

# DeXtreme: Transfer of Agile In-hand Manipulation from Simulation to Reality



Ankur Handa\*

Arthur Allshire\*

Viktor Makoviychuk\*

Aleksei Petrenko\*

Ritvik Singh\*

Jingzhou Liu\*

Denys Makoviichuk

Karl Van Wyk

Alexander Zhurkevich

Balakumar Sundaralingam

Yashraj Narang

Jean-Francois Lafleche

Dieter Fox

Gavriel State

NVIDIA



Figure 1: The DeXtreme system using an Allegro Hand in action in the real world.



## Abstract

Recent work has demonstrated the ability of deep reinforcement learning (RL) algorithms to learn complex robotic behaviours in simulation, including in the domain of multi-fingered manipulation. However, such models can be challenging to transfer to the real world due to the gap between simulation and reality. In this paper, we present our techniques to train a) a policy that can perform robust dexterous manipulation on an anthropomorphic robot hand and b) a robust pose estimator suitable for providing reliable real-time information on the state of the object being manipulated. Our policies are trained to adapt to a wide range of conditions in simulation. Consequently, our vision-based policies significantly outperform the best vision policies in the literature on the same reorientation task and are competitive with policies that are given privileged state information via motion capture systems. Our work reaffirms the possibilities of sim-to-real transfer for dexterous manipulation in diverse kinds of hardware and simulator setups, and in our case, with the Allegro Hand and Isaac Gym GPU-based simulation. Furthermore, it opens up possibilities for researchers to achieve such results with commonly-available, affordable robot hands and cameras. Videos of the resulting policy and supplementary information, including experiments and demos, can be found at <https://dextreme.org/>.



\*Equal contribution

All authors are with NVIDIA (except for Denys Makoviichuk, who is with Snap). In addition to NVIDIA, Arthur Allshire, Jingzhou Liu, and Ritvik Singh are also with the University of Toronto, and Aleksei Petrenko is with the University of Southern California.

Author contributions listed at the end of the paper.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Method</b>	<b>3</b>
2.1	Task . . . . .	3
2.2	Hardware . . . . .	4
2.3	Policy Learning with RL . . . . .	4
2.4	Reward Formulation . . . . .	6
2.5	Simulation . . . . .	6
2.6	Domain Randomisation . . . . .	7
2.6.1	Physics Randomisations . . . . .	8
2.6.2	Non-physics Randomisations . . . . .	10
2.6.3	Measuring ADR Performance in Training and in the Real World . . . . .	11
2.7	Pose Estimation . . . . .	11
<b>3</b>	<b>Results</b>	<b>13</b>
3.1	Training in Simulation . . . . .	13
3.2	Real-World Policy Performance . . . . .	14
3.3	Quirks, Problems, and Surprises . . . . .	16
<b>4</b>	<b>Related work</b>	<b>17</b>
<b>5</b>	<b>Limitations</b>	<b>18</b>
<b>6</b>	<b>Acknowledgements</b>	<b>19</b>
<b>7</b>	<b>Contributions</b>	<b>20</b>
<b>A</b>	<b>Appendix</b>	<b>25</b>
A.1	Compute Budget Comparisons . . . . .	25
A.2	Hardware Comparisons . . . . .	25
A.3	PPO Hyperparameters . . . . .	25
A.4	Isaac Gym Simulation Parameters . . . . .	26
A.5	Progressive Improvements in Consecutive Successes in the Real World . . . . .	26
A.6	Default KUKA configuration . . . . .	26
A.7	Software Tools Used in the Work . . . . .	26

## 1 Introduction

Multi-fingered robotic hands offer an exciting platform to develop and enable human-level dexterity. Not only do they provide kinematic redundancy for stable grasps, but they also enable the repertoire of skills needed to interact with a wide range of day-to-day objects. However, controlling such high-DoF end-effectors has remained challenging. Even in research, most robotic systems today use parallel-jaw grippers.

In 2018, OpenAI et al. [1] showed for the first time that multi-fingered hands with a purely end-to-end deep-RL based approach could endow robots with unprecedented capabilities for challenging contact-rich in-hand manipulation. However, due to the complexity of their training architecture, and the *sui generis* nature of their work on sim-to-real transfer, reproducing and building upon their success has proven to be a challenge for the community. Recent advancements in in-hand manipulation with RL have made progress with multiple objects and an anthropomorphic hand [2], but those results have only been in simulation.

While the NLP and computer vision communities have reproduced and extended the successes of large-scale models like GPT-3 [3] and DALL-E [4, 5] respectively, similar efforts have remained elusive in robotics due to hardware and infrastructure challenges. Using large-scale data from simulations may provide avenues to unlock a similar step function in robotics capabilities.

This paper builds on top of the prior work in [1]. We use a comparatively affordable Allegro Hand with a locked wrist and four fingers, using only position encoders on servo motors; the Shadow Hand used in OpenAI’s experiments costs an order of magnitude more than the Allegro Hand. We also develop a simple vision system that requires no specialised tracking or infrastructure on the hand; the system works on three off-the-shelf RGB cameras compared to OpenAI’s expensive marker-based setup, making our system easily accessible for everyone. Furthermore, we use the GPU-based Isaac Gym physics simulator [6] as opposed to the CPU-based MuJoCo [7], which allows us to reduce the amount of computational resources used and the complexity of the training infrastructure. Our best models required only 8 NVIDIA A40 GPUs to train, as opposed to OpenAI’s use of a CPU cluster composed of 400 servers with 32 CPU-cores each, as well as 32 NVIDIA V100 GPUs [8] (compute requirements for block reorientation). Our more affordable hand, in combination with the simple vision system architecture and accessible compute, dramatically simplifies the process of developing and deploying agile and dexterous manipulation. We summarise our contributions below:



- We demonstrate a system for learning-based dexterous in-hand manipulation that uses low-cost hardware (one order of magnitude less expensive than [1]), uses a purely vision-based pipeline, sets more diverse pose targets, uses orders-of-magnitude cheaper compute, and offers further insights into this problem with detailed ablations.
- We develop a highly robust pose estimator trained entirely in simulation which works through heavy occlusions and in a variety of robotic settings e.g. [https://www.youtube.com/watch?v=-MTsm0Uh\\_5o](https://www.youtube.com/watch?v=-MTsm0Uh_5o).
- While not directly comparable to [1] due to different hardware, our purely vision-based state estimation results not only outperform their best vision-based results, but also fare comparably to their marker-based results.
- We will also release both our vision and RL pipelines for reproducibility. We seek to provide a much broader segment of the research community with access to a novel state-of-the-art in-hand manipulation system in hopes of catalyzing further studies and advances.

## 2 Method

### 2.1 Task

We propose a method for performing object reorientation on an anthropomorphic hand. Initially the object to be manipulated is placed on the palm of the hand and a random target orientation is sampled in  $SO(3)$ <sup>2</sup>. The policy then orchestrates the motion of the fingers so as to bring the object to its desired target orientation. Similar to OpenAI et al. [1], if the object orientation is within a specified threshold of 0.4 radians of the target orientation, we sample a new target orientation. The fingers continue from the current configuration and aim to move the object to its new target orientation. The success criterion is the number of consecutive target orientations achieved without dropping the object or having the object stuck in the same configuration for more than 80 seconds. Importantly, each consecutive success becomes increasingly harder to achieve as the fingers have to keep the object in the hand without dropping, hence testing the policy’s ability to model the dynamics on the go.

<sup>2</sup>In contrast to previous works [1, 8], which limited it to configurations with flat faces pointing upwards.

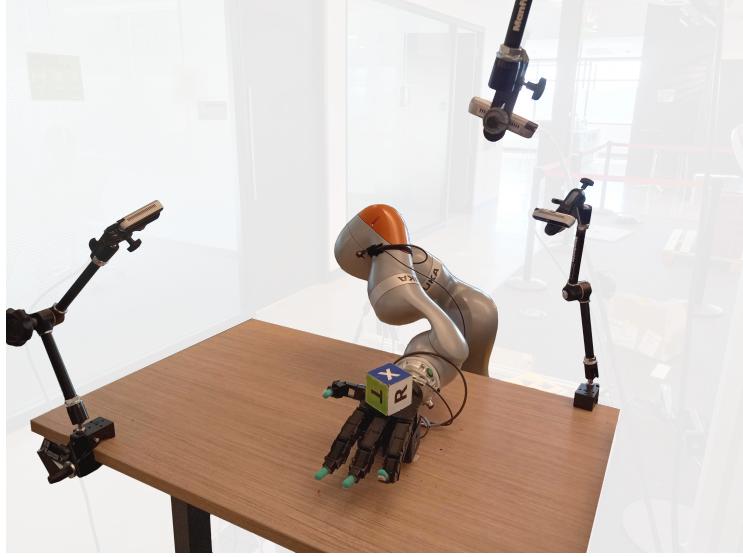


Figure 2: The hardware setup used in this work, unlike [1], is not housed in a cage, and our system is robust enough to perform the task in an open laboratory environment. The background in the image is alpha-blended for visibility.

For a quick and high level understanding of this work, we encourage readers to watch the video <https://www.youtube.com/watch?v=TAUiAYAVkfI>. Figure 3 provides a quick overview of different components involved in the system.

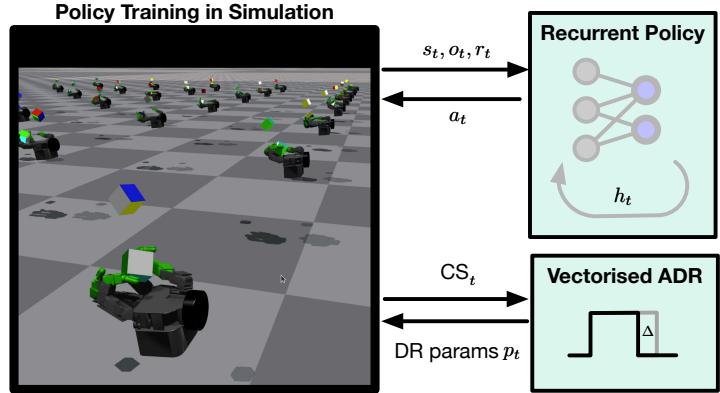
## 2.2 Hardware

Our hardware setup (see Fig 2) consists of an Allegro Hand rigidly mounted at the wrist. We use 3 Intel D415 cameras for object tracking with RGB frames *i.e.* no depth images were used. The cameras are extrinsically calibrated relative to the palm link of the hand. Our object tracking is done entirely using a vision-based system, and in contrast to [1], we do not use any marker-based system to track the cube or fingertip states.

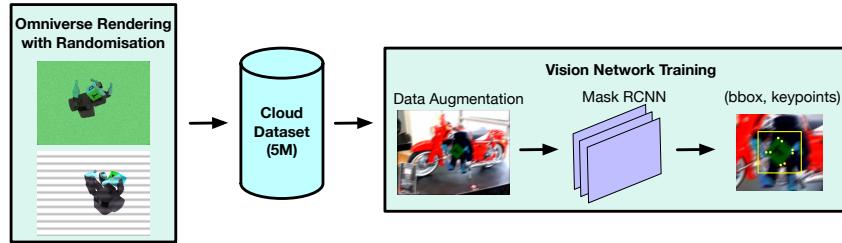
The object we learn to manipulate is a 6.5 cm cuboid with coloured and lettered stickers on the sides. These stickers allow the vision system to distinguish between different faces (see Sec. 2.7). The pose of the cube is represented with respect to the palm of the robot hand. The camera-camera extrinsics and camera-robot calibration allow us to transform the cube pose from the canonical reference frame of a camera to the palm. Since the cube is represented locally in the palm reference frame, the policy performance is not dependent on the physical location of the setup, enabling us to move the setup freely whenever desired.

## 2.3 Policy Learning with RL

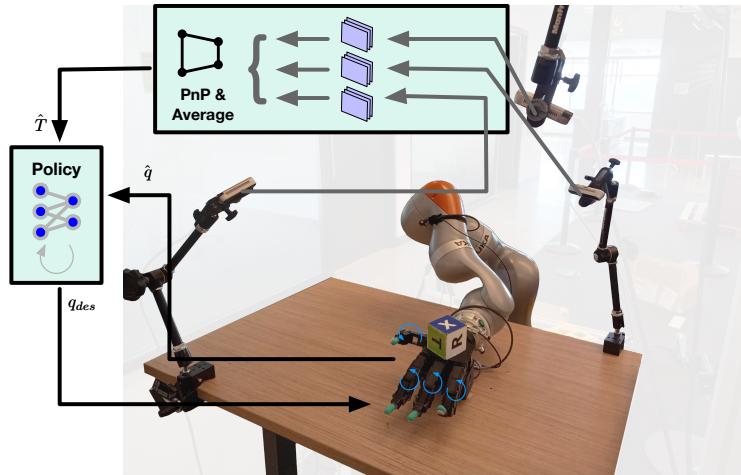
**RL Formulation:** The task of manipulating the cube to the desired orientation is modelled as a sequential decision making problem where the agent interacts with the environment in order to maximise the sum of discounted rewards. In our case, we formulate it as a discrete-time, partially observable Markov Decision Process (POMDP). We use Proximal Policy Optimisation (PPO) [9] to learn a parametric stochastic policy  $\pi_\theta$  (actor), mapping from observations  $o \in \mathcal{O}$  to actions  $a \in \mathcal{A}$ . PPO additionally learns a function  $V_\phi^\pi(s, o)$  (critic) to approximate the on-policy value function. Following Pinto et al. [10], the critic does not take in the same observations as the actor, but receives additional observations including states  $s \in \mathcal{S}$  in the POMDP. The actor and critic observations are detailed in Table 1.



(a) Policy training.



(b) Vision data generation and training pipeline.



(c) Functioning in the real world.

Figure 3: High level overview of the training and inference systems.

Input	Dimensionality	Actor	Critic
Object position with noise	3D	✓	✓
Object orientation with noise	4D (quaternion)	✓	✓
Target position	3D	✓	✓
Target orientation	4D (quaternion)	✓	✓
Relative target orientation	4D (quaternion)	✓	✓
Last actions	16D	✓	✓
Hand joints angles	16D	✓	✓
Stochastic delays	4D	✗	✓
Fingertip positions	12D	✗	✓
Fingertip rotations	16D (quaternions)	✗	✓
Fingertip velocities	24D	✗	✓
Fingertip forces and torques	24D	✗	✓
Hand joints velocities	16D	✗	✓
Hand joints generalised forces	16D	✗	✓
object scale, mass, friction	3D	✗	✓
Object linear velocity	3D	✗	✓
Object angular velocity	3D	✗	✓
Object position	3D	✗	✓
Object rotation	4D	✗	✓
Random forces on object	3D	✗	✓
Domain randomisation params	78D	✗	✓
Gravity vector	3D	✗	✓
Rotation distances	2D	✗	✓
Hand scale	1D	✗	✓

Table 1: Observations of the policy and value networks. The input vector is 50D in size for policy and 265D for the value function.

We use a high-performance PPO implementation from rl-games [11] with the following hyperparameters: discount factor  $\gamma=0.998$ <sup>3</sup>, clipping  $\epsilon=0.2$ . While in some experiments the learning rate was updated adaptively based on a fixed KL threshold 0.016, our best result was obtained using linear scheduling of the learning rate for the policy (start value  $lr = 1e-4$ ) and a fixed learning rate for the value function ( $lr = 5e-5$ ). Our best policy  $\pi_\theta : \mathcal{O} \times \mathcal{H} \rightarrow \mathcal{A}$  was a Long Short-Term Memory (LSTM) network [12] taking in environment observations  $o$  and previous hidden state  $h \in \mathcal{H}$ . We use an LSTM backpropagation through time (BPTT) truncation length of 16. The LSTM has 1024 hidden units with layer normalization and is followed by 2 multilayer perceptron (MLP) layers with sizes 512 and ELU activation [13]. The action space  $\mathcal{A}$  of our policy is the PD controller target for each of the 16 joints on the robot hand. The value function LSTM layer has 2048 hidden units with layer normalization, followed by 2 MLP layers with 1024 and 512 units respectively with ELU activation. The output of the policy is low-pass filtered with an exponential moving average (EMA) smoothing factor. During training this factor is annealed from 0.2 to 0.15. Our best results in the real world were obtained with an EMA of 0.1, which provided a balance between agility and stability of the motion, preventing the hardware from breaking or motor cables from burning.

## 2.4 Reward Formulation

The reward formulation is inspired by the Shadow hand environment in Isaac Gym[6], and described and justified in Table 2.

## 2.5 Simulation

Our aim in this paper is to learn dexterous manipulation behaviours. Current on-policy learning algorithms can struggle to accomplish this on real robots due to the number of samples required. Hence, we learn our behaviours entirely in simulation. We use the GPU-based Isaac Gym physics

<sup>3</sup>We found that following [1] and setting the higher discount of 0.998 (as opposed to 0.99 as used with MLPs [6]) was essential to allowing us to train LSTMs.

Reward	Formula	Weight	Justification
Rotation Close to Goal	$1/(d + 0.1)$	1.0	Shaped reward to bring cube close to goal
Position Close to Fixed Target	$\ \mathbf{p}_{\text{object}} - \mathbf{p}_{\text{goal}}\ $	-10.0	Encourage the cube to stay in the hand
Action Penalty	$\ \mathbf{a}\ ^2$	-0.001	Prevent actions that are too large
Action Delta Penalty	$\ \mathbf{targ}_{\text{curr}} - \mathbf{targ}_{\text{prev}}\ ^2$	-0.25	Prevent rapid changes in joint target
Joint Velocity Penalty	$\ \mathbf{v}_{\text{joints}}\ ^2$	-0.003	Stop fingers from moving too quickly
Reset Reward	Condition	Value	
Reach Goal Bonus	$d < 0.1$	250.0	Large reward for getting the cube to the target

Table 2: Reward terms are computed, multiplied by their weight, and summed to produce the reward at each timestep.  $d$  represents the rotational distance from the object's current to the target orientation.  $\mathbf{p}_{\text{object}}$  and  $\mathbf{p}_{\text{goal}}$  are the position of the object and goal respectively.  $a$  is the current action.  $\mathbf{targ}_{\text{curr}}$  and  $\mathbf{targ}_{\text{prev}}$  are the current and previous joint position targets.  $\mathbf{v}_{\text{joints}}$  is the current joint velocity vector.

Parameter	Type	Distribution	Initial Range	ADR-Discovered Range
<b>Hand</b>				
Mass	Scaling	uniform	[0.4, 1.5]	[0.4, 1.5]
Scale	Scaling	uniform	[0.95, 1.05]	[0.95, 1.05]
Friction	Scaling	uniform	[0.8, 1.2]	[0.54, 1.58]
Armature	Scaling	uniform	[0.8, 1.02]	[0.31, 1.24]
Effort	Scaling	uniform	[0.9, 1.1]	[0.9, 2.49]
Joint Stiffness	Scaling	loguniform	[0.3, 3.0]	[0.3, 3.52]
Joint Damping	Scaling	loguniform	[0.75, 1.5]	[0.43, 1.6]
Restitution	Additive	uniform	[0.0, 0.4]	[0.0, 0.4]
<b>Object</b>				
Mass	Scaling	uniform	[0.4, 1.6]	[0.4, 1.6]
Friction	Scaling	uniform	[0.3, 0.9]	[0.01, 1.60]
Scale	Scaling	uniform	[0.95, 1.05]	[0.95, 1.05]
External Forces	Additive	Refer to [1]	—	—
Restitution	Additive	uniform	[0.0, 0.4]	[0.0, 0.4]
<b>Observation</b>				
Obj. Pose Delay Prob.	Set Value	uniform	[0.0, 0.05]	[0.0, 0.47]
Obj. Pose Freq.	Set Value	uniform	[1.0, 1.0]	[1.0, 6.0]
Obs Corr. Noise	Additive	gaussian	[0.0, 0.04]	[0.0, 0.12]
Obs Uncorr. Noise	Additive	gaussian	[0.0, 0.04]	[0.0, 0.14]
Random Pose Injection	Set Value	uniform	[0.3, 0.3]	[0.3, 0.3]
<b>Action</b>				
Action Delay Prob.	Set Value	uniform	[0.0, 0.05]	[0.0, 0.31]
Action Latency	Set Value	uniform	[0.0, 0.0]	[0.0, 1.5]
Action Corr. Noise	Additive	gaussian	[0.0, 0.04]	[0.0, 0.32]
Action Uncorr. Noise	Additive	gaussian	[0.0, 0.04]	[0.0, 0.48]
RNA $\alpha$	Set Value	uniform	[0.0, 0.0]	[0.0, 0.16]
<b>Environment</b>				
Gravity (each coord.)	Additive	normal	[0, 0.5]	[0, 0.5]

Table 3: Domain randomisation parameter ranges for policy learning

simulator [6], which models contacts differently than MuJoCo’s soft-contact model [7] used in [1]. Isaac Gym gives the advantage of being able to simulate thousands of robots in parallel on a single GPU, mitigating the need for large amounts of CPU resources.

## 2.6 Domain Randomisation

It is widely known that there is a "sim-to-real" gap between physics simulators and real life [14]. Compounding this is the fact that the robot as a system can change from day to day (*e.g.*, due to wear-and-tear) and even from timestep to timestep (*e.g.*, stochastic noise). To help overcome this, we introduce various kinds of randomisations [15] into the simulated environment as listed in Table 3.

**Vectorised Automatic Domain Randomisation:** In our best policies, we set the parameters of the domain randomisations via a vectorised implementation of Automatic Domain Randomisation (ADR, introduced in [8]). ADR automatically adjusts the range of domain randomisations to keep them as wide as possible while keeping the policy performance above a certain threshold. This allows us to train policies with less randomisation earlier in training (enabling behaviour exploration) while producing final policies that are robust to the largest range of environment conditions possible at

the end of training, with the aim of improving sim-to-real transfer by learning policies which are robust and adaptive to a range of environment randomisation parameters. Using the parallelisation provided by Isaac Gym, we implement a vectorised variant of the algorithm, which we call Vectorised Automatic Domain Randomisation (VADR, see Algorithm 1).

The range of randomisations for each environment parameter in ADR is modelled as a uniform distribution  $d^n \sim U(p^{2n}, p^{2n+1})$ , where  $p^{2n}$  and  $p^{2n+1}$  are the current lower and upper randomisation boundaries, and  $n \in 0, \dots, D - 1$  is the parameter index for each of the  $D$  ADR dimensions. Each dimension starts with initial values from system calibration or best-guesses for the randomisation bounds,  $p_{init}^{2n}$  and  $p_{init}^{2n+1}$  for the lower and upper bounds of parameter  $n$ , respectively. Optional minimum and maximum bounds on the randomisations may also be specified,  $p_{min}^n$  and  $p_{max}^n$ . Unlike in [8], we choose the size of step  $\Delta^n$  separately for each parameter. This trades off more tuning work for more stable training and the mitigation of the need for custom, secondary distributions on top of certain randomisation dimensions, as were used in that work.

Evaluation proceeds as follows: environments sample a value for each randomisation dimension uniformly between the upper and lower bounds. A fraction (40%) of the vectorised environments are dedicated to evaluation. In these environments, one of the ADR randomisation dimensions is fixed to the current lower or upper boundary (the rest of the dimensions are sampled from the aforementioned uniform distribution set by ADR). The episode proceeds to roll out; the number of consecutive successes is recorded at the end of the episode. This figure is added to a queue for the boundary of maximum length  $N = 256$ . Then, if mean consecutive successes on the boundary is above a certain threshold,  $t_H = 20$ , the range is widened, and if the performance is below a lower threshold  $t_L = 5$ , then the range is tightened on that bound. This is depicted in Figure 4. If on a particular step the value of a bound changes, the queue is cleared (as the previous performance data then becomes invalid). In this way, ADR will discover the maximum ranges over which a policy can perform, including for example discovering the limits of parameters impacting physics stability. Our vectorised ADR (VADR) full algorithm is described below in Algorithm 1 and Algorithm 2.

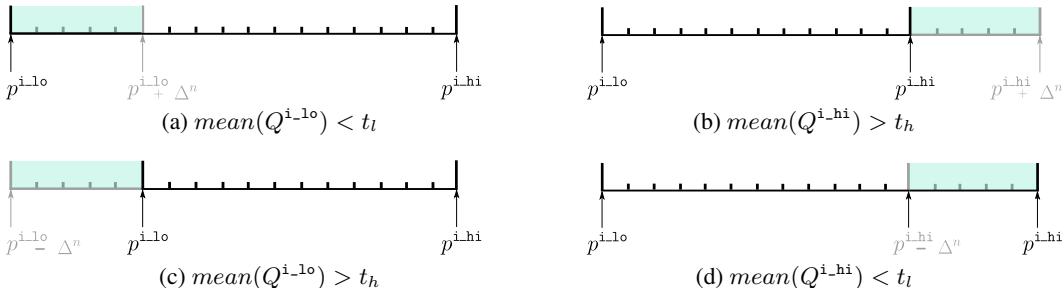


Figure 4: Parameter range adjustments,  $p^{i\_lo}$  and  $p^{i\_hi}$ , with ADR based on the performance of policy at the boundaries  $Q^{i\_lo}$  and  $Q^{i\_hi}$  with respect to the thresholds  $t_l$  and  $t_h$ . If the mean performance at the boundary  $Q^{i\_lo}$  is less than threshold  $t_l$ , the range is tightened (a) and if it is above a threshold  $t_h$ , the range is expanded (c). Similarly, if the mean performance at the boundary  $Q^{i\_hi}$  is above a threshold  $t_h$ , the range is expanded (b) and if it is lower than  $t_l$ , the range is tightened (d).

When training on multiple GPUs (8 for our best policies), we ran VADR separately on each one. This was done for two reasons: firstly, to avoid additional synchronisation overhead of buffers. Secondly, to partially mitigate the disadvantage of ADR noted below in Section 3.3 caused by the failure to model the joint distribution; having multiple independent parameter sets to some extent will allow multiple sets of extreme parameters. All randomisations are set by ADR.<sup>4</sup> In the following, we describe the physics and non-physics randomisations in more detail.

### 2.6.1 Physics Randomisations

We apply physics randomisations to account for both changing real-world dynamics and the inevitable gaps between physics in simulation and reality. These include basic properties such as mass, friction

<sup>4</sup>Except for mass and scale, which are randomised within a fixed range, as collision morphologies cannot be changed at run-time.

---

**Algorithm 1:** Vectorised Automatic Domain Randomisation (VADR)

---

```

Function ADRUpdate (p, success_counts, ADR_mode)
    /* Function to decide whether to update each ADR boundary's value
       based on the success level of just-completed episodes. */
    for n = 0, ..., D - 1 do
        i_lo = 2n ;
        i_hi = 2n + 1 ;
         $Q^{i\_lo} \leftarrow \text{success\_counts}[\text{ADR\_modes} == i\_lo];$ 
         $Q^{i\_hi} \leftarrow \text{success\_counts}[\text{ADR\_modes} == i\_hi];$ 
        if length( $Q^{i\_lo}$ ) == N then
            if mean( $Q^{i\_lo}$ ) > t_h then
                 $p^{i\_lo} \leftarrow p^{i\_lo} - \Delta^n$ 
            else if mean( $Q^{i\_lo}$ ) < t_l then
                 $p^{i\_lo} \leftarrow p^{i\_lo} + \Delta^n$ 
            if  $p^{i\_lo}$  changed then
                 $p^{i\_hi} \leftarrow \text{clip}(p^{i\_lo}, p_{min}^n, p_{max}^n);$ 
                Clear( $Q^{i\_lo}$ )
            if length( $Q^{i\_hi}$ ) == N then
                if mean( $Q^{i\_hi}$ ) > t_h then
                     $p^{i\_hi} \leftarrow p^{i\_hi} + \Delta^n$ 
                else if mean( $Q^{i\_hi}$ ) < t_l then
                     $p^{i\_hi} \leftarrow p^{i\_hi} - \Delta^n$ 
                if  $p^{i\_hi}$  changed then
                     $p^{i\_hi} \leftarrow \text{clip}(p^{i\_hi}, p_{min}^n, p_{max}^n);$ 
                    Clear( $Q^{i\_hi}$ )
            end
        return p
Function ResetADRVals (p, ADR_modes, ADR_vals)
    /* Function to resample ADR values for the next episode just after an
       episode is done. */
    /* NB. all operations here are vectorised, so eg. ADR_modes is a
       tensor with 60% of values -1. */
     $\text{ADR\_modes} = \begin{cases} -1, & \text{with } p = 60\%; \\ \sim \text{Categorical}(0, 2D - 1), & \text{with } p = 40\%; \end{cases};$ 
    /* Resample ADR values for each dimension for the next episode. */
    for n = 0, ..., D - 1 do
        is_eval = ADR_modes == 2n | ADR_modes == 2n+1
         $\text{ADR\_vals}\{\text{n}\} = \begin{cases} p^{\text{ADR\_modes}}, & \text{where } \text{i\_eval} \\ \sim U(p^{2n}, p^{2n+1}), & \text{otherwise} \end{cases};$ 
    end
    return ADR_modes, ADR_vals

```

---

**Algorithm 2:** Training with VADR.

---

Initialise tensor  $\text{ADR\_modes} \in \mathbb{Z}^{N_{envs}}$  storing whether each environment is a normal environment (entry= -1) or performing sampling at one of the boundaries (entry $\in 0, \dots, 2D - 1$ );

Initialise tensor  $\text{ADR\_vals} \in R^{N_{envs} \times D}$  storing ADR randomisation values for each parameter in each environment (matrix storing values of  $\phi$  from text).;

Initialise ADR boundaries  $p \in \mathbb{R}^{2D}$

**while** training **do**

- dones, successes = *train\_step*(*ADR\_vals*);
- p* = *ADRUpdate*(*p*, successes[dones], *ADR\_mode*[dones]);
- ADR\_modes*[dones], *ADR\_vals*[dones] = *ResetADRVals*(*p*, *ADR\_modes*[dones], *ADR\_vals*[dones]);

**end**

---

and restitution of the hand and object. We also randomly scale the hand and object to avoid over-reliance on exact morphology. On the hand, joint stiffness, damping, and limits are randomised. Furthermore, we add random forces to the cube in a similar fashion to [1].

**Joint Stiffness, Joint Damping, and Effort** are scaled using the value sampled directly from the ADR-given uniform distribution.

**Mass and Scale** are randomised within a fixed range due to API limitations currently. However, we did not observe this as a significant limitation for our experiments, and our policies nevertheless achieved rollouts with high consecutive successes in the real world.

**Gravity** cannot be randomised per-environment in Isaac Gym currently, but a new gravity value is sampled every 720 concurrent simulation steps for all environments.

### 2.6.2 Non-physics Randomisations

In addition to normal physics randomisations, Table 3 lists action and observation randomisations, which we found to be critical to achieving good real-world performance. To make our policies more robust to the changing inference frequency and jitter resulting from our ROS-based inference system, we add stochastic delays to cube pose and action delivery time as well as fixed-for-an-episode action latency. To the actions and observations, we add correlated and uncorrelated additive Gaussian noise. To account for unmodelled dynamics, we use a Random Network Adversary (RNA, see below).

#### Observation & Action Noise

We apply Gaussian noise to the observations and actions with the noise function

$$f_{\delta, \epsilon}(x) = x + \delta + \epsilon$$

Where  $\delta$  and  $\epsilon$  are sampled from Gaussian distributions parameterised by the ADR values  $p^i, p^j$ ,  $\delta \sim \mathcal{N}(\cdot; 0, \text{var}(p^i))$ ,  $\epsilon \sim \mathcal{N}(\cdot; 0, \text{var}(p^j))$  where  $\text{var}(a) = \exp[a^2] - 1$

For  $\delta$ , this sampling happens once per episode at the beginning of the episode, corresponding to correlated noise. For  $\epsilon$ , sampling happens at every timestep. Note that the formula for var has a cutoff at 0 noise. This allows ADR to set a certain fraction of environments to have 0 noise, which we found an important case that is not covered in previous works when setting fixed or above-zero cutoff variance (since during inference, zero white noise is added).

#### Latency & Delay

We apply three forms of delay. The first is an exponential delay, where the chance of applying a delay each step is  $p^i$  and is given by  $f(x; x_{last}) = x_{last} \cdot d + x \cdot (1 - d)$  and  $d \sim \text{Bern}(\cdot; p^i)$  is the Bernoulli distribution parametrised by the  $i$ -th ADR variable,  $p^i \in [0, 1]$ . This delay case, applied to both observations of cube pose and actions, mimics random jitter in latency times.

The second form of delay is action latency, where the action from  $n$  timesteps ago is executed. For this parameter, we slightly modify the vanilla ADR formulation to allow smooth increase in delay with ADR value despite the discretisation of timesteps. The bounds are still continuously modified, but the sampling from the range is done from a categorical distribution. Specifically, let  $\epsilon \sim U(0, b) + U(-0.5, 0.5)$  be the sampled ADR value (plus random noise used to allow probabilistic blending of delay steps when sampling on the ADR boundary). Then the delay  $k$  is  $k = \text{round}(\epsilon)$ .

A third form of delay, this time on observation, is that caused by the refresh rate of the cameras in the real world. To compensate for this, we have randomisation on the refresh rate. Similarly to the aforementioned action latency, we use ADR to sample a categorical action delay  $d \in \{1, \dots, \text{delay}_{max}\}$ . We then only update the cube pose observation if  $(t + r) \bmod d = 0$ , effectively mimicing a pose estimation frequency of  $d \cdot \Delta t$  (where  $r$  is a randomly sampled alignment variable to offset updates from the beginning of the episode randomly).

#### Random Pose Injection

We noticed that due to heavy occlusion and caging from the fingers, our cube pose estimator exhibited occasional jumps. To ensure that the policy performance did not deteriorate and LSTM hidden state become corrupted by this, we occasionally inject completely random cube poses into the network. At the start of each episode for each environment, we sample a probability  $p \in U(0, 0.3)$ .

## Random Network Adversary

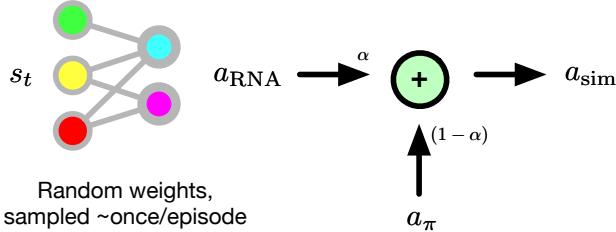


Figure 5: The functioning of the Random Network Adversary

Then each step we sample a variable  $m \sim \text{Bern}(\cdot; p)$ , and the cube pose becomes:  $\text{pose\_obs} = \text{pose} \cdot (1 - m) + \text{random\_pose} \cdot m$ .

### Random Network Adversary

Random Network Adversary, introduced in [8], uses a randomly-generated neural network each episode to introduce much more structured, state-varying noise patterns into the environment, in contrast to normal Gaussian noise. As we are doing simulation on GPU rather than CPU, instead of using a new network per environment-episode and wasting memory on thousands of individual MLPs, we generate a single network across all environments and use a unique and periodically refreshed dropout pattern per environment. Actions from the RNA network are blended with those from the policy by  $a = \alpha \cdot a_{RNA} + (1 - \alpha) \cdot a_{\text{policy}}$ , where  $\alpha$  is controlled by ADR.

### 2.6.3 Measuring ADR Performance in Training and in the Real World

Nats per Dimension (npd) was a metric developed by OpenAI in [8] to measure the amount of randomisation through the average entropy across the ADR distributions. While it does not directly capture the difficulty of the environment (since each dimension is not normalised for difficulty), it provides a rough proxy for how much randomisation there is in the environment. The formula for nats per dimension is given by:

$$\text{npd} = \frac{1}{D} \sum_{n=0}^{D-1} \log(p^{2n+1} - p^{2n}) \quad (1)$$

Currently, we measure real-world performance based on the number of consecutive successes. An avenue for future work is directly exploring how real-world policy performance corresponds to ADR randomisation levels in total and across different dimensions.

## 2.7 Pose Estimation

**Data Generation and Processing:** We use NVIDIA Omniverse Isaac Sim with Replicator<sup>5</sup> to generate 5M images of the cube in hand in just under a day. Each image is  $320 \times 240$  in resolution and contains visual domain randomisations as summarised in Table 4 to add variety to the training set. Such visual domain randomisations allow the network to be robust to different camera parameters and visual effects that may be present in the scene. In addition, we apply data augmentations during training on a batch in order to add even more variety to the training set. As such, a single rendered image from the dataset can provide multiple training examples with different data augmentation settings, thereby saving both rendering time and storage space. For instance, motion blur (important in our case where we have a fast-moving object to track) can be especially time-consuming at rendering time. Instead we generate it on the fly via motion-blur data augmentation by smearing the image with a Gaussian kernel with the blur direction and extent chosen randomly. The data augmentations used on the images are listed in Table 5. Each augmentation is applied with a fixed probability to ensure that the batch consists of a combination of the original as well as augmented images.

<sup>5</sup>See <https://developer.nvidia.com/isaac-sim>

Parameter	Probability Distribution
Albedo desaturation	uniform(0.0, 0.3)
Albedo add	uniform(-0.2, 0.2)
Albedo brightness	uniform(0.5, 1.0)
Diffuse tint	normal(1.0, 0.3)
Reflection roughness constant	uniform(0.0, 1.0)
Metallic constant	uniform(0.0, 0.5)
Specular level	uniform(0.0, 1.0)
<b>Enable camera noise</b>	<b>bernoulli(0.3)</b>
Enable scan lines	bernoulli(0.3)
Scan line spread	uniform(0.1, 0.5)
Enable vertical lines	bernoulli(0.1)
Enable film grain	bernoulli(0.1)
Grain amount	uniform(0.0, 0.15)
Grain size	uniform(0.7, 1.2)
Enable random splotches	bernoulli(0.1)
Colour amount	uniform(0.0, 0.3)
Hand visibility	bernoulli(0.75)
Camera pose	Hemispherical shell <sup>6</sup>
Camera f/l multiplier	uniform(0.99, 1.01)

Table 4: Ranges of vision parameter randomisations.

Data Augmentation Type	Probability
CutMix (see [16])	0.5
Random Blurring	0.5
Random Background	0.6
Random Rotation	0.5
Random Brightness and Contrast	0.5
Random Cropping and Resizing	0.5

Table 5: Various data augmentations applied to the images on the fly during training. We also set a probability for each one of them.

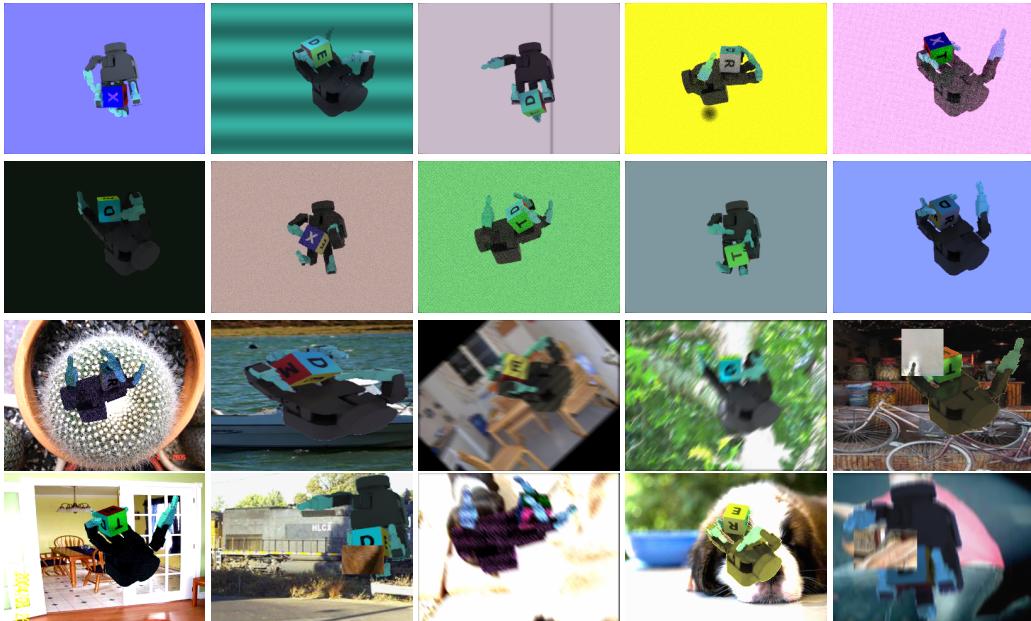


Figure 6: Illustration of generated data. *Top row*: Images generated from NVIDIA Isaac Sim, demonstrating different camera locations, lighting conditions, and camera poses. *Bottom row*: Images after data augmentations such as CutMix, random cropping and rotation, and motion blur are applied during training.

We also collect configurations of the cube in-hand generated by our policies running in the real world and play them back in Isaac Sim to render data where the pose estimates are not fully reliable *i.e.* the pose estimator is not accurate all the time. This happens due to the sim-to-real gap as a result of either sparse sampling or insufficient lighting randomisations for those configurations. Playback in simulation enables dense sampling of pose and lighting around these configurations with more randomisations. This allows us to generate larger datasets that improve the reliability of the pose estimator in the real world *i.e.* closing the perception loop between real and sim by mining configurations from the current best policy in the real world and using them in sim to render more images. We use the same intrinsics of the cameras in the real world, but randomise extrinsics when rendering data.

<sup>6</sup>Camera pose is randomly sampled from a hemispherical shell with thickness 0.3m around the hand, with a random focus selected with a 0.2m radius around the hand origin.

Experiment	Avg. Rotation Error	Avg. Translation Error		
		X	Y	Z
Sim	$5.3 \pm 0.11^\circ$	$1.9 \pm 0.1$ mm	$4.1 \pm 0.2$ mm	$6.9 \pm 0.4$ mm

Table 6: Rotation and translation error on test dataset with 90% confidence intervals.

**Training setup and inference:** We use a torchvision Mask-RCNN [17]-inspired network <sup>7</sup> that regresses to the bounding box, segmentation, and keypoints located at the 8 corners of the cube. The bounding box localises the cube in the image, and the keypoint head regresses to the positions of the 8 corners of the cube within the bounding box. The networks are trained with cross-entropy loss for segmentation and keypoint location regression, and smooth L1 loss for bounding box regression. We use the ADAM optimiser with a learning rate of 1e-4. The network runs on three cameras at an inference rate of 20Hz on an NVIDIA RTX 3090 GPU and a 32-core AMD Ryzen Threadripper CPU. However, because the policy was trained with a control frequency of 30Hz in simulation, the pose estimator was locked to run at 15Hz to ensure that the policy receives pose observations at a constant integer interval of once every two control steps. To make the pose estimate reliable for the downstream policy, we first perform classic PnP [18] on each of the three cameras independently and then filter out the ones where the projected keypoints from the PnP pose do not match the inferred keypoints from the network. We triangulate the keypoints from the filtered cameras and register them against the model of the cube to obtain the pose in a canonical reference frame. We use the OpenCV implementation of PnP and romax [19] for registering keypoints against the model of the cube. We benchmark the pose on a test set consisting of 50K images and provide results in Table 6. Since we do not use any marker-based system in the real world, we can only precisely evaluate the performance of the pose estimator in simulation. Our ablation studies in Section 3.2 do test the strength of the pose estimator for manipulation in the real world. One important difference between our approach and OpenAI et al. [1] is that our pose estimation is not done end-to-end. Since we detect keypoints in the image and use geometric computer vision to obtain the pose, our pose estimator is not tied to a fixed camera setup, unlike [1].

### 3 Results

In the following section, we present the results we achieved in object reorientation in the simulations and then real world using the methods described in Section 2. We then follow it up with tests of policy robustness in reality and simulation.

#### 3.1 Training in Simulation

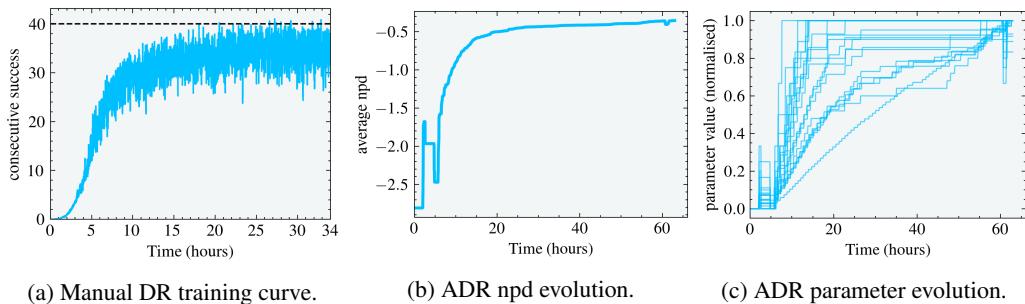


Figure 7: Training curve evolution for (a) manual domain randomisation, which takes  $\sim 24$  hours to reach an average of 35 consecutive successes, (b) ADR experiments showing how the average nats per distribution (npd), an indicator of the extent of randomisation, increases as a function of training (c) shows how parameters (normalised) optimised with ADR evolve over time. All curves are from experiments in simulations.

For all of our experiments, we use a simulation  $dt$  of  $\frac{1}{60}$  s and a control  $dt$  of  $\frac{1}{30}$  s. We train with 16384 agents per GPU and use a goal-reaching orientation threshold of 0.1 rad but test with 0.4 rad

<sup>7</sup>We also tried U-Net based direct regression of the keypoints from the image but found it to be somewhat unreliable. Sometimes the keypoints were detected at entirely different locations in the image where the cube was not even present. Having a network that first localises the cube and then detects the keypoints within the bounding box prevents such mis detections.

as in [1, 6] for all experiments both in simulation and the real world. All policies are trained with randomisations described in Table 3. Most importantly, our ADR policies using the same compute resources — 8 NVIDIA A40s — achieved the best performance in the real world after training for only 2.5 days in contrast to [8] that trained for 2 weeks to months for the task of block reorientation<sup>8</sup>. Various training curves for our task a) with manual DR, b) with automatic domain randomisation (ADR), and c) ADR parameter evolution are presented in Figure 7. We note that due to differences in physics engines and hand morphology, our simulation average consecutive successes are not directly comparable, but we achieve performance on par with [1, 8].

Training with manual DR takes roughly 32 hours to converge on 8 NVIDIA A40s generating a combined (across all GPUs) frame rate of 700K frames/sec. With a  $dt = \frac{1}{60}$ , this amounts to  $\frac{32 \times 700000}{60 \times 24 \times 365}$  which is  $\sim 42$  years of real-world experience.

### 3.2 Real-World Policy Performance

Experiment	Cons. Success Trials (sorted)	Average	Median
Best Model	1, 6, 6, 10, 10, 18, 18, 36, 61, 112	$27.8 \pm 19.0$	14.0
	3, 4, 7, 16, 19, 22, 29, 31, 58, 77	$26.6 \pm 13.2$	20.5
	1, 5, 5, 11, 12, 12, 33, 36, 42, 51	$20.8 \pm 9.8$	12.0
Best Model (Goal frame count=10)	6, 8, 10, 16, 16, 17, 20, 33, 39, 45	$21.0 \pm 7.4$	16.5
	9, 11, 13, 13, 15, 16, 27, 29, 32, 36	$20.1 \pm 5.4$	15.5
	2, 3, 3, 9, 11, 12, 14, 15, 43, 44	$16.6 \pm 8.4$	11.5
Non-ADR Model	2, 3, 7, 7, 13, 16, 22, 23, 26, 29	$14.8 \pm 5.4$	14.5
	1, 1, 3, 7, 8, 11, 14, 17, 22, 35	$11.9 \pm 5.8$	9.5
	0, 7, 8, 8, 9, 10, 10, 11, 17, 20	$10.0 \pm 3.0$	9.5

Table 7: The results of running different models on the real robot. We run 10 trials per policy [1] to benchmark the average consecutive successes. Individual rows within each experiment indicate running the experiment on different days [20] and  $\pm$  indicates 90% confidence interval. Our best model was trained with ADR while non-ADR experiments had DR ranges manually tuned. The second experiment shows results when the cube is held at a goal for additional consecutive frames once the target cube pose is reached.

It is worth noting that, while in simulations, state information is derived directly from physics buffers, in all real-world experiments we use the pose estimator described in Section 2.7 to obtain the pose of the cube and provide it as input to the policy. The qualitative results of this are best illustrated by the accompanying videos at <https://dextreme.org>.

The deployment pipeline of the system is shown in Figure 8. We use three separate machines to run various components. The main machine has an NVIDIA RTX 3090, which runs both the policy as well as the pose estimator. We also do live visualisation of the real-world manipulation in Omniverse but disable the physics.

Similar to [1], we observe a large range of different behaviours in the policies deployed on the real robot. Our real-world quantitative results measuring average consecutive successes are illustrated in Table 7. We collect 10 trials for each policy to obtain the average consecutive successes and also collect different sets of trials across different days to understand the inter-day variability that may arise due to different dynamics, temperature, and lighting conditions. We believe such inter-day variations are important to benchmark in robotics [20] and have endeavoured to highlight this specifically in this challenging task. We find that our policies do not show a dramatic drop in average performance, indicating that they are mostly robust to inter-day variations.

We benchmark both ADR and non-ADR (manually-tuned DR ranges) policies in Table 7 and like [1] find that the policies trained with ADR perform the best, suggesting that the sheer diversity of data gleaned from the simulator endows the policies with the extreme robustness needed in the real world. Importantly, we observed that policies trained with non-ADR exhibited ‘stuck’ behaviours (as shown in Figure 9), which ADR-based policies were able to overcome due to increased diversity in training data. We also find that on an average, the trials with ADR achieve more consecutive successes than non-ADR policies. Table 9 puts our results in perspective alongside the previous works of [1] and [8]. **We demonstrate performance which significantly improves upon the best vision policies**

<sup>8</sup>Although [8] focused on the Rubik’s cube, they also trained for block reorientation (pp. 20, Table 3) with the same infrastructure to reproduce their results from [1].

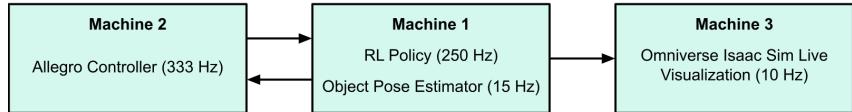


Figure 8: The system runs on a combination of three machines. The pose estimation and policy are run on the same machine, whereas the Allegro control and live Omniverse visualisation are done on two separate machines that communicate with the main machine at different rates via ROS messages.

**from [1] and ADR (XL)<sup>9</sup> policies given high-quality state information from a motion capture system in [8].** Our policies do not achieve the average successes seen in [8] with ADR (XXL) with *state information*. We hypothesise that this maybe due to (a) better accuracy of the state information from the PhaseSpace markers (b) higher frame-rate of state observations with PhaseSpace markers (c) increased diversity of data with ADR (XXL). However, our best vision-based policy generated  $\sim 2.5 \times$  higher peak consecutive successes and  $\sim 1.5 \times$  higher mean consecutive successes than the vision-based policies in [1] as shown in Table 9.



Figure 9: Policies trained with manual DR exhibited ‘stuck’ behaviours, where the cube remained stuck in certain configurations and was unable to recover. An example of such behaviour can be observed here: <https://www.youtube.com/watch?v=tJgq18VbL3k>.

Additionally, we also benchmark our results beyond the basic experiment of goal reaching by making the policy hold the cube at a target orientation  $N$  frames in a row (we use  $N = 10$ ), *i.e.*, we count the number of times the cube orientation is within the threshold of 0.4 rad in a row (as described in 3.1), refresh the goal only if the frame counter reaches  $N$ , and reset the counter to zero every time it goes outside the threshold. In the basic experiment of goal reaching without the hold, the cube may shoot past the target, making it difficult to tell if the target was achieved merely due to noise in the pose estimation or if the cube orientation was indeed accurately estimated. Therefore, holding the cube for  $N$  frames in a row ensures that the goal was not achieved by chance, highlighting the robustness of the pose estimator. We conduct this experiment only with our ADR policies as shown in (Table 7, 2<sup>nd</sup> row). It is worth noting that the policy was not trained explicitly to hold the cube and that the experiment is meant to be a test of accuracy of pose. We chose  $N = 10$  based on our simulation experiments and found that setting  $N$  too high led to a dramatic drop in the performance because the LSTM was not trained for such scenarios (see Table 8). On the other hand, setting it too low did not change the simulation performance;  $N = 10$  was a good balance between performance and LSTM stability. This also lets us separate the drop in performance due to LSTM instability from pose estimation errors in the real world. From the trials, we find that while there is a noticeable drop in performance *i.e.* the maximum consecutive successes are only 45 as opposed to 112 in the other case, the average consecutive successes have not dropped as dramatically suggesting both the robustness of the policy and the pose estimator.

Frame Hold (N)	Cons. Successes
0	38.4
5	35.3
10	33.3
20	27.3

Table 8: Performance in simulation with ADR policies with respect to  $N$ . Therefore, holding the cube for  $N$  frames

Method	DR type	Pose estimation type	Training time	Cons. Successes	Median	Best rollout
OpenAI <i>et al.</i> [1] (state)	Manual	PhaseSpace	2.08 days (50 hours)	$18.8 \pm 5.4$	13.0	50
OpenAI <i>et al.</i> [1] (state, locked wrist)	Manual	PhaseSpace	2.08 days (50 hours)	$26.4 \pm 7.3$	28.5	50
OpenAI <i>et al.</i> [1] (vision)	Manual	Neural Network	2.08 days (50 hours)	$15.2 \pm 7.8$	11.5	46
OpenAI <i>et al.</i> [8] (state)	ADR (L)	PhaseSpace	13.76 days	$13.3 \pm 3.6$	11.5	—
OpenAI <i>et al.</i> [8] (state)	ADR (XL)	PhaseSpace	Multiple Months	$16.0 \pm 4.0$	12.5	—
OpenAI <i>et al.</i> [8] (state)	ADR (XXL)	PhaseSpace	Multiple Months	$32 \pm 6.4$	42.0	—
Ours (vision)	Manual	Neural Network	1.41 days (34 hours)	$14.8 \pm 5.4$	14.5	29
Ours (vision, best avg)	ADR	Neural Network	2.5 days (60 hours)	$27.8 \pm 19.0$	14.0	112
Ours (vision, best median)	ADR	Neural Network	2.5 days (60 hours)	$26.6 \pm 13.2$	20.5	77
Ours (vision, max successes capped at 50)	ADR	Neural Network	2.5 days (60 hours)	$23.1 \pm 9.4$	20.5	50

Table 9: We compare our block reorientation results against the previous work of OpenAI *et al.* [1] and OpenAI *et al.* [8]. It is important to note that we use the less capable but more affordable, 4-fingered **Allegro Hand** with a different morphology and a locked wrist, whereas they use the high-end, tendon-based **Shadow Hand** with five fingers and a movable wrist. Notwithstanding this disparity between the two robot platforms, we have tried to provide the best possible comparison in this table. Our best vision based policy on average performs better than the vision-based policy in [1] and the state-based policy trained with ADR(XL) in [8] with PhaseSpace markers while taking only 2.5 days to train. OpenAI *et al.* [8] do not mention the precise time to train with ADR (XL) and ADR (XXL) but do say “multiple months” in the paper. Missing entries in the last column mean that the information was not available in the paper. The table only considers the time taken by the control policy to train, and training time for vision models is not included. Our best rollout with vision achieves 112 consecutive successes, which is roughly  $2.5\times$  more than theirs with vision[1]. It is worth reminding that both [1] and [8] capped the maximum consecutive successes to 50 (Section 6.2, OpenAI *et al.*[1]). Therefore, we also provide the statistics with successes capped at 50 in the last row of this table. The compute budget comparisons are in Section 1 and Appendix A.1.

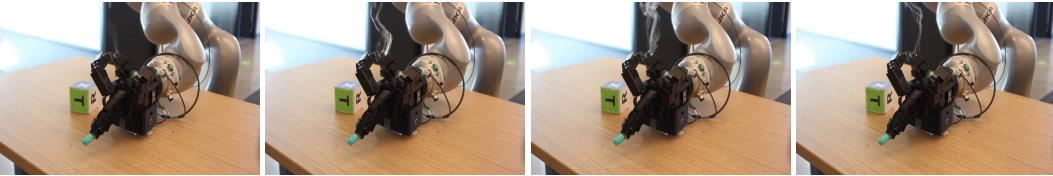


Figure 10: Ribbon cables can burn with aggressive maneuvers or excessive usage. Full video is available at <https://www.youtube.com/watch?v=H5x0er9jxJc>. The images show one of the instances of running the policy leading to burning cables and smoke coming out of the hand. Regular health checkups are necessary, involving replacing cables, tightening screws and resting the hand for a few hours after running about 10-15 trials.

### 3.3 Quirks, Problems, and Surprises

Throughout our work, we experienced a variety of recurring problems and observations, which we disclose here in the interest of transparency and providing avenues for future research. Due to complexities involved in training and slow turnaround time in results because of real hardware involved in the loop, we do not have rigorous experiments to prove these. However, they constitute “tricks of the trade” that will hopefully be of use to other researchers.

- Even though we are not training on the real robot hand and our policies are very gentle due to the low-pass filters we apply, the hand is quite fragile and requires regular health checks and cable replacements *e.g.* the ribbon cables connecting joints on the fingers can break or burn, disabling the corresponding joints (see Figure 10). We also apply hot glue at the either ends of the ribbon cables going into connectors to tighten the connection and prevent them from coming out.
- We had applied 300lse tape on fingers and palm of the hand for another unrelated task of grasping, but it added significant friction, preventing the cube from sliding and moving smoothly on the palm. Higher friction also meant that the fingers tried aggressively to manipulate the cube when it was stuck in the hand. Therefore, we removed the tape from the palm. However, we kept the tape on the sides of the fingers to allow better grips on the object.

<sup>9</sup>XL and XXL denote the degree of randomisations.

<sup>9</sup>To fully hold the cube stationary in hand for a target orientation means zero velocities at the target, which requires changing the reward function.

- We found that deployment in the real world was quite variable between policies trained with different seeds, even with the same domain randomisation parameters. This is an issue that was also noted in [21]. We suspect that this is because, despite the extreme levels of randomisation we do, there is a "null space" of possible policies which perform similarly in simulation but differently in the real world.
- Automatic Domain Randomisation [8], which was used to achieve our best results, does not model the joint distribution and relationships between performance with different parameters. This means it can provide different results depending on the order in which randomisation ranges increase, and furthermore it may not explore the Pareto-limits of performance on any particular dimension. This can lead to cases where ADR disproportionately increases the performance of the policy on dimension A at the expense of dimension B.
- Our best policy trained in a non-ADR setting exhibited slightly more natural and smooth behaviour than the ADR one. We also observed that the ADR policies tended to cage the object, resulting in letters occluded by the fingers from all cameras. This made the pose estimation unreliable in those configurations (although the policy was able to absorb the errors and performed better overall). It is possible temporal smoothing of the pose estimator can help as long as it does not introduce any latency.
- We also tried pose estimation without the letters on the faces and found that to be quite challenging — the network was confused by colour changes due to lighting in the real scene and the quality of the cameras used *e.g.* yellow turning into white in the image when the normal of a cube face is perpendicular to the camera. Having a robot cage as used in [1, 8] and [22] with an constrained artificial light can prevent this from happening to an extent, but it also makes the system confined to the cage.

On the other hand, we were surprised positively about some aspects of our system:

- Our pose estimator proved to be surprisingly robust even in real-world scenarios outside of Allegro Hand manipulation. This hints at the power of extreme randomisation with synthetic data in simulation for developing such general systems.
- The ability to, at test time, adjust the speed of the policy by tuning the EMA value (see Section 2.3 - we trained with 0.15 but tested with 0.1) was very useful to avoid damaging the hardware while at the same time having agile policies.
- The agility of our policies in the real world on the Allegro Hand, which we initially expected to be a limitation as it is relatively large and is motor- rather than tendon-driven, was a positive surprise to us.
- Most of our experiments and trials conducted on the robot were done with a worn-out thumb on the Allegro Hand — one of the wires connecting a joint to the circuit board had a loose connection — and we were quite surprised that the policies produced high consecutive successes despite a malfunctioning actuator, suggesting the robustness of learning-based approaches. The slow turnaround time involved in repairing the hardware motivated us to do it ourselves regularly during the experiments, but it was only a temporary solution.
- Since we do not regress to the pose via an end-to-end network, we found that our pose estimator was not tied to a particular camera configuration as keypoint detection worked reliably from different configurations<sup>10</sup> *i.e.* our earlier results as described in A.5 were obtained with a camera configuration that was different to the one we used afterwards with ADR.
- Our best policies were trained with continuous actions unlike OpenAI et al. [1, 8], where discrete actions were used.

## 4 Related work

In-hand manipulation is a longstanding challenge in robotics. Classical methods [23–25] have focussed on directly leveraging the robot’s kinematic and dynamic model to analytically derive controllers for the object in hand. These approaches work well while an object maintains no-slip

---

<sup>10</sup>While extrinsics change with different camera configurations, the intrinsics remain the same.

contacts with the hand, but struggle to achieve good results in dynamic tasks where complex sequences of making-and-breaking of contacts is needed.

Reinforcement learning has proven to be a powerful method for learning complex behaviours in robots with many degrees of freedom. Hwangbo et al. [26] and Lee et al. [27] showed how robust legged locomotion can be learned even over challenging terrain using a combination of deep RL, fast simulation, and domain randomisation. A crucial inspiration for our work is OpenAI et al. [1] on learning in-hand reorientation of cubes via reinforcement learning in simulation and subsequent OpenAI et al. [8] extending this task to solving Rubik’s cubes. A variety of recent works [6, 28–30] have leveraged new RL techniques and simulators in order to reproduce or extend in simulation the anthropomorphic in-hand manipulation capabilities shown in [1]. However, these works have not shown sim-to-real transfer, demonstrating that this remains a significant challenge for learned in-hand manipulation. There have also been a variety of recent approaches attempting in-hand manipulation from the perspective of finger gaiting [31, 32]. However, these often fail to reproduce the agile dexterity present in human hands, as the limitations of such a sequential approach to control place corresponding limits on speed. Similarly, Allshire et al. [22] and Shi et al. [33] achieve real-world in-hand manipulation using reinforcement learning, but on a platform that cannot mimic the dexterity of a human hand. Sievers et al. [34] learned in-hand manipulation using tactile sensing, but the lack of vision meant that the kinds of re-posing behaviour and speed of the manipulation were limited. There has also been research investigating using reinforcement learning directly on hardware platforms as opposed to in simulation [35, 36]. However, this limits the complexity of learning that can be done due to the small number of trials available on real hardware platforms, as well as the wear-and-tear imposed.

Pose estimation for robotic manipulation is a widely studied area [37, 38]. However, relatively few works outside of OpenAI et al. [1] have applied it to the problem of contact-rich, dexterous, in-hand manipulation, which introduces challenges that exclude many off-the-shelf pose estimators (due to the large degree of occlusions, motion blur, *etc.*).

## 5 Limitations

Despite our best efforts, the gap between simulations and the real world is still noticeable. Our non-ADR (manual DR) based policy achieves an average of 35 consecutive successes (see Figure 7(a)) in simulation, but only obtains an average of about 15 consecutive successes in the real world. While our ADR policies do perform better, they still fail to reach the average successes seen in simulations. Another limitation of our work is that our ADR policies do not consistently obtain high consecutive successes as seen in ADR (XXL) in OpenAI et al. [8]. It is unclear whether we need a more reliable pose estimation *e.g.* marker-based systems, the policy needs more diversity in the data, or if it was due to the malfunctioning thumb on our Allegro Hand. Our best npd with ADR policies is around -0.2, and it is possible that further improvements can be made. This is something we would like to investigate in future work.

We also observed that there is still some sim-to-real gap in pose estimation. This is manifested when we played back the real states in sim (real-to-sim) with physics enabled, which sometimes resulted in interpenetrations. Therefore, we were not able to easily calibrate physics parameters of the cube, which could have improved our sim-to-real transfer.

Lastly, putting this work in context, it is worth remembering that some of the key reasons for successful transfer of this task involve:

- Being able to simulate the task, and having a clear and well-defined reward function so that the policies can be trained in simulation.
- Randomisation of interpretable parameters exposed by the simulator and providing a curriculum for training in simulation. For instance, we randomised *friction*, *damping*, *stiffness*, *etc.*, which were crucial for the transfer. ADR provided a curriculum.
- Ability to evaluate successful execution of the task so that we can track the improvements. In this work, we compared the current orientation with the desired, and if the difference was within a user-specified threshold (0.4 rad in our case), the goal-reaching was considered successful.

Many real-world tasks are hard to simulate and sometimes defining reward functions is not possible as it is in this task. Even if we could simulate and have a well-defined reward function, evaluating a successful execution of a task, *e.g.* cooking a meal, is not straightforward.

## 6 Acknowledgements

The authors would like to thank Nathan Ratliff, Lucas Manuelli, Erwin Coumans, and Ryan Hickman for helpful discussions. Maciej Bala assisted with multi-GPU training; Michael Lin helped with hardware and Zhutian Yang with video editing. Nick Walker provided valuable suggestions on the teaser video and proofreading. Thanks also to Ankit Goyal for the help with high frame-rate video capture and Jie Xu for proofreading.

## 7 Contributions

### Reinforcement Learning & Simulation

- **Viktor Makoviychuk** implemented the first version of the Allegro Hand environment in Isaac Gym.
- **Ankur Handa, Arthur Allshire, and Viktor Makoviychuk** developed the domain randomisations in Isaac Gym that assisted in sim2real transfer.
- **Arthur Allshire** developed the vectorised Automatic Domain Randomisation (ADR) in Isaac Gym.
- **Ankur Handa, Arthur Allshire, Viktor Makoviychuk, and Aleksei Petrenko** trained RL policies and added various features to the environment to improve sim2real transfer.
- **Denys Makoviichuk** implemented the RL Library used in this project and helped implement specific extensions required for this work.
- **Yashraj Narang and Gavriel State** advised on tuning simulations.
- **Dieter Fox and Gavriel State** advised on experiments.

### Vision

- **Ankur Handa** developed the code to train pose estimation and data augmentation for the vision models.
- **Arthur Allshire** wrote the vision data rendering system and developed the domain randomisations in Omniverse. **Ritvik Singh and Jingzhou Liu** generalised and extended the rendering pipeline.
- **Ankur Handa, Ritvik Singh, and Jingzhou Liu** trained pose estimation models and did real-to-sim with vision to improve the pose estimator.
- **Alexander Zhurkevich** helped speed up the inference.

### Real-World Experiments

- **Ankur Handa** conducted the real-world experiments. **Arthur Allshire** helped with experiments in the early stages, and **Viktor Makoviychuk** helped with experiments in the later stages of the project.
- **Arthur Allshire** wrote the code to perform policy & vision inference on the real robot. **Ankur Handa** maintained and extended it.
- **Arthur Allshire** developed the live visualisation pipeline in Omniverse.
- **Karl Van Wyk** managed the Allegro Hand infrastructure to run experiments on the real hand.
- **Balakumar Sundaralingam and Ankur Handa** provided assistance repairing the Allegro Hand when it broke.
- **Karl Van Wyk** developed an automatic calibration system for camera-camera and camera-robot calibration.

### Organisational

- **Arthur Allshire and Ankur Handa** drafted the paper.
- **Yashraj Narang, Ritvik Singh, Jingzhou Liu, and Gavriel State** helped to edit the paper.
- **Gavriel State, Dieter Fox, and Jean-Francois Lafleche** provided resources and support for the project.
- **Ankur Handa** edited the videos. **Gavriel State, Arthur Allshire, Viktor Makoviychuk, Jingzhou Liu, Yashraj Narang and Ritvik Singh** examined the videos and provided feedback.
- **Ankur Handa** led the project. **Ankur Handa, Arthur Allshire and Viktor Makoviychuk** designed the roadmap of the project.

## References

- [1] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub W. Pachocki, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *CoRR*, abs/1808.00177, 2018. URL <http://arxiv.org/abs/1808.00177>. 3, 4, 6, 7, 10, 13, 14, 15, 16, 17, 18, 25
- [2] Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. *Conference on Robot Learning*, 2021. 3
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>. 3
- [4] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL <https://arxiv.org/abs/2102.12092>. 3
- [5] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL <https://arxiv.org/abs/2204.06125>. 3
- [6] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning. *CoRR*, abs/2108.10470, 2021. URL <https://arxiv.org/abs/2108.10470>. 3, 6, 7, 14, 18, 26
- [7] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012. 3, 7
- [8] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik’s cube with a robot hand. *CoRR*, abs/1910.07113, 2019. URL <http://arxiv.org/abs/1910.07113>. 3, 7, 8, 11, 14, 15, 16, 17, 18, 25
- [9] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms, 2017. 4
- [10] Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. *CoRR*, 2017. URL <http://arxiv.org/abs/1710.06542>. 4
- [11] Denys Makoviichuk and Viktor Makoviychuk. rl-games: A high-performance framework for reinforcement learning. [https://github.com/Denys88/rl\\_games](https://github.com/Denys88/rl_games), May 2022. 6, 26
- [12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. 6
- [13] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL <http://arxiv.org/abs/1511.07289>. 6
- [14] Nick Jakobi, Phil Husbands, and Inman Harvey. Noise and the reality gap: The use of simulation in evolutionary robotics. In Federico Morán, Alvaro Moreno, Juan Julián Merelo, and Pablo Chacón, editors, *Advances in Artificial Life*, 1995. 7

- [15] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 3803–3810. IEEE, 2018. 7
- [16] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. *CoRR*, abs/1905.04899, 2019. URL <http://arxiv.org/abs/1905.04899>. 12
- [17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. URL <http://arxiv.org/abs/1703.06870>. 13
- [18] Wikipedia. Perspective n point. URL <https://en.wikipedia.org/wiki/Perspective-n-Point>. 13
- [19] Romain Brégier. Deep regression on manifolds: a 3d rotation case study. *CoRR*, abs/2103.16317, 2021. URL <https://arxiv.org/abs/2103.16317>. 13, 26
- [20] Google. The importance of a/b testing in robotics, 2021. URL <https://ai.googleblog.com/2021/06/the-importance-of-ab-testing-in-robotics.html>. 14
- [21] Lei M. Zhang, Matthias Plappert, and Wojciech Zaremba. Predicting sim-to-real transfer with probabilistic dynamics models. *CoRR*, abs/2009.12864, 2020. URL <https://arxiv.org/abs/2009.12864>. 17
- [22] Arthur Allshire, Mayank Mittal, Varun Lodaya, Viktor Makoviychuk, Denys Makoviichuk, Felix Widmaier, Manuel Wuthrich, Stefan Bauer, Ankur Handa, and Animesh Garg. Transferring Dexterous Manipulation from GPU Simulation to a Remote Real-World TriFinger. *CoRR*, 2021. 17, 18
- [23] J. Kenneth Salisbury and John J. Craig. Articulated hands: Force control and kinematic issues. *The International Journal of Robotics Research*, 1(1):4–17, 1982. doi: 10.1177/027836498200100102. URL <https://doi.org/10.1177/027836498200100102>. 17
- [24] Matthew T. Mason and J. Kenneth Salisbury. *Robot Hands and the Mechanics of Manipulation*. MIT Press, Cambridge, MA, USA, 1985. ISBN 0262132052.
- [25] Zexiang Li, Ping Hsu, and Shankar Sastry. Grasping and coordinated manipulation by a multi-fingered robot hand. *The International Journal of Robotics Research*, 8(4):33–50, 1989. doi: 10.1177/027836498900800402. URL <https://doi.org/10.1177/027836498900800402>. 17
- [26] Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning Agile and Dynamic Motor Skills for Legged Robots. *Science Robotics*, Jan 2019. 18
- [27] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47):eabc5986, 2020. doi: 10.1126/scirobotics.abc5986. URL <https://www.science.org/doi/abs/10.1126/scirobotics.abc5986>. 18
- [28] Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. 2021. URL <https://arxiv.org/abs/2111.03043>. 18
- [29] Wenlong Huang, Igor Mordatch, Pieter Abbeel, and Deepak Pathak. Generalization in dexterous manipulation via geometry-aware multi-task learning. 2021. URL <https://arxiv.org/abs/2111.03062>.
- [30] Vittorio Caggiano, Huawei Wang, Guillaume Durandau, Massimo Sartori, and Vikash Kumar. Myosuite – a contact-rich simulation suite for musculoskeletal motor control, 2022. URL <https://arxiv.org/abs/2205.13600>. 18
- [31] Ryosuke Higo, Yuji Yamakawa, Taku Senoo, and Masatoshi Ishikawa. Rubik’s cube handling using a high-speed multi-fingered hand and a high-speed vision system. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6609–6614, 2018. doi: 10.1109/IROS.2018.8593538. 18

- [32] Andrew S. Morgan, Kaiyu Hang, Bowen Wen, Kostas E. Bekris, and Aaron M. Dollar. Complex in-hand manipulation via compliance-enabled finger gaiting and multi-modal planning. *RAL*, abs/2201.07928, 2022. URL <https://arxiv.org/abs/2201.07928>. 18
- [33] Fan Shi, Timon Homberger, Joonho Lee, Takahiro Miki, Moju Zhao, Farbod Farshidian, Kei Okada, Masayuki Inaba, and Marco Hutter. Circus anymal: A quadruped learning dexterous manipulation with its limbs, 2020. 18
- [34] Leon Sievers, Johannes Pitz, and Berthold Bäuml. Learning purely tactile in-hand manipulation with a torque-controlled hand. 2022. doi: 10.48550/ARXIV.2204.03698. URL <https://arxiv.org/abs/2204.03698>. 18
- [35] Vikash Kumar, Emanuel Todorov, and Sergey Levine. Optimal control with learned local models: Application to dexterous manipulation. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 378–383, 2016. doi: 10.1109/ICRA.2016.7487156. 18
- [36] Vikash Kumar, Abhishek Gupta, Emanuel Todorov, and Sergey Levine. Learning dexterous manipulation policies from experience and imitation. *CoRR*, abs/1611.05095, 2016. URL <http://arxiv.org/abs/1611.05095>. 18
- [37] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects. In *CoRL*, 2018. 18
- [38] Yann Labb  , Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosopose: Consistent multi-view multi-object 6d pose estimation. 2020. URL <https://arxiv.org/abs/2008.08465>. 18
- [39] Charles R. Harris, K. Jarrod Millman, St  fan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fern  ndez del R  o, Mark Wiebe, Pearu Peterson, Pierre G  rard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL <https://doi.org/10.1038/s41586-020-2649-2>. 26
- [40] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, St  fan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2. 26
- [41] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, pages 90–95, 2007. 26
- [42] Morgan Quigley, Brian Gerkey, Ken Conley, Josh Faust, Tully Foote, Jeremy Leibs, Eric Berger, Rob Wheeler, and Andrew Ng. Ros: an open-source robot operating system. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*, Kobe, Japan, may 2009. 26
- [43] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alch  -Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf>. 26

[44] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. [26](#)

[45] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Jan 2019. [27](#), [28](#)

## A Appendix

### A.1 Compute Budget Comparisons

Method	DR type	Compute Infrastructure	Training time	Training Cost
OpenAI et al. [1]	Manual DR	384 CPU servers with 16-cores each and 8 NVIDIA V100s	2.08 days	\$14,280.0
OpenAI et al. [8]	ADR	400 CPU servers with 32-cores each and 32 NVIDIA V100s	13.76 days	\$215,685.1
Ours	Manual DR	8 NVIDIA A40s	1.41 days	\$553.8
Ours	ADR	8 NVIDIA A40s	2.50 days	\$977.2

Table 10: Compute budget comparisons for block reorientation task of our work against the previous work.

The costs are estimated from the AWS EC2 instance pricing as of *October 19, 2022* at <https://ec2pricing.net/>. The estimated costs for OpenAI et al. [1, 8] may have been higher in 2018 and 2019 respectively.

For the OpenAI [1] equivalent, a p3.16xlarge  $8 \times$ V100 machine at \$24.48/hr and a c6i.4xlarge with 16 CPU cores at \$0.68/hr add up to a total cost of  $(24.48 + 0.68 \times 384) \times 50 = \$14280$ . Similarly, for an equivalent of OpenAI [8] with ADR, a 32-core c6i.8xlarge machine at \$1.36/hr, the overall cost can be estimated to be  $(24.48 \times 4 + 1.36 \times 400) \times 24 \times 14 = \$215,685$ .

While AWS does not offer NVIDIA A40s, it does provide NVIDIA A10Gs which are similar in performance. A g5.48xlarge instance with 8 NVIDIA A10G GPUs costs at \$16.288/hr. For our manual DR experiments the overall compute adds up to  $16.288 \times 34 = \$553.8$  and for ADR it is  $16.288 \times 60 = \$977.28$  in total. The cost for such experiments on p4d.24xlarge instances offering 8-GPU A100 machines at \$32.77/hr would be \$1114.18 and \$1966.2 respectively.

### A.2 Hardware Comparisons

Method	Robot Hand	RGB	Markers on Cube	Markers on Fingertips	Tactile Sensors	Robot Cage
OpenAI et al. [1]	Shadow Hand	✓	✓	✓	✗	✓
OpenAI et al. [8]	Shadow Hand	✗	✓	✓	✗	✓
Ours	Allegro Hand	✓	✗	✗	✗	✗

Table 11: Our hardware setup compared against the one used in OpenAI et al. [1] and OpenAI et al. [8]. Note that the experiment pertaining to the block reorientation in [8] was done with marker-based systems and not RGB cameras.

### A.3 PPO Hyperparameters

Hyperparameter	Value
Hardware configuration	8 NVIDIA A40
Action distribution	16D Continuous actions
Discount factor	0.998
Generalised advantage estimation	0.95
Entropy regularisation coefficient	0.002
PPO clipping parameter	0.2
KL-divergence threshold	0.16
Optimiser	Adam
Learning rate (Actor)	1e-4
Learning rate (Critic)	5e-4
Minibatch size	16384
Learning epochs	4
Horizon length	16
LSTM input sequence length	16
Action bounds loss coefficient	0.005

Table 12: Various hyperparameters used to train the policy with PPO.

#### A.4 Isaac Gym Simulation Parameters

Hyperparameter	Value
Sim $dt$	1/60s
Control $dt$	1/30s
Num Envs	8192/16384
Num Substeps	2
Num Pos Iterations	8
Num Vel Iterations	0

Table 13: Simulation settings for the Allegro Hand environment.

#### A.5 Progressive Improvements in Consecutive Successes in the Real World

Date	Max Consecutive Success	Notes
29-Mar-2022	10	Manual DR
14-Jun-2022	15	Pose wrt wrist
21-Jun-2022	20	Increased DR ranges
26-Jul-2022	31	Increased network sizes
27-Jul-2022	43	Random pose injection and RNA
20-Aug-2022	70	ADR
27-Aug-2022	77	ADR
30-Aug-2022	112	ADR and deeper networks

#### A.6 Default KUKA configuration

Joint Name	A1	A2	A3	A4	A5	A6	A7
	-42.38	-53.38	66.01	116.45	7.20	57.55	81.53

Table 14: Default rest configuration of the KUKA arm joints (in degrees) used in this work.

#### A.7 Software Tools Used in the Work

Software	Source
Numpy [39]	<a href="https://numpy.org/">https://numpy.org/</a>
Scipy [40]	<a href="https://scipy.org/">https://scipy.org/</a>
Matplotlib [41]	<a href="https://matplotlib.org/">https://matplotlib.org/</a>
ROS [42]	<a href="http://wiki.ros.org/rospy">http://wiki.ros.org/rospy</a>
Isaac Gym [6]	<a href="https://developer.nvidia.com/isaac-gym">https://developer.nvidia.com/isaac-gym</a>
IsaacGymEnvs [6]	<a href="https://github.com/NVIDIA-Omniverse/IsaacGymEnvs">https://github.com/NVIDIA-Omniverse/IsaacGymEnvs</a>
RL Games [11]	<a href="https://github.com/Denys88/rl_games">https://github.com/Denys88/rl_games</a>
PyTorch [43]	<a href="https://pytorch.org/">https://pytorch.org/</a>
PyTorch3D [44]	<a href="https://pytorch3d.org/">https://pytorch3d.org/</a>
Roma [19]	<a href="https://github.com/naver/roma">https://github.com/naver/roma</a>
Pangolin	<a href="https://github.com/stevenlovegrove/Pangolin">https://github.com/stevenlovegrove/Pangolin</a>
Omniverse Isaac Sim	<a href="https://developer.nvidia.com/isaac-sim">https://developer.nvidia.com/isaac-sim</a>
iMovie	<a href="https://www.apple.com/imovie/">https://www.apple.com/imovie/</a>

We present a model card for DeXtreme in Table 15, following Mitchell et al. [45].

Model Details		
Organization Developing the Model	NVIDIA	
Model Date	October 2022	
Model Type	torchvision mask-rcnn inspired model for pose estimation (details in Section 2.7) and LSTM model for policy training (details in Section 2.3) .	
Feedback on the Model	ahanda@nvidia.com, aallshire@nvidia.com, jason-liu@nvidia.com, ritviks@nvidia.com, gstate@nvidia.com, vmakoviychuk@nvidia.com	
Intended Uses		
Primary Intended Uses	The primary use is research on pose estimation and training control policies conditioned on the pose.	
Primary Intended Users	Researchers working in sim-to-real. We will make some of the code available publicly.	
Out-of-Scope Uses	None. The models are meant to work on a particular combination of hardware used in this work.	
Metrics		
Model Performance Measures	We focus on the performance of model in very challenging conditions <i>e.g.</i> occlusions, blur and lighting changes. However, since there is no real-world dataset with ground truth for the use case we have. Therefore, most of our quantitative evaluations are done only in simulated settings Section 2.7.	
Decision thresholds	N/A	
Approaches to Uncertainty and Variability	N/A	
Training Data		
Datasets	We use our dataset rendered with NVIDIA Omniverse. We do not intend to make that dataset public.	
Motivation	We test the pose estimator in various challenging conditions. The simulation parameters are tuned to optimise the performance in the real world. The policy is trained with RL with domain randomisation to enable sim-to-real transfer.	
Preprocessing	Images are processed so that their mean and variance are 0 and 1 respectively. Similarly, for policy we subtract mean and divide by standard deviation.	
Evaluation Data		
Datasets	We test the model in laboratory settings on real-world data collected with our robot. We do not intend to make that dataset public.	
Quantitative Analyses		

Unitary Results	We refer to Section 3.2 for the full details of our quantitative study.
<b>Performance and Limitations</b>	
Section 3.2 is about performance of the model in the real world and Section 5 describes the current limitations of our work.	

Table 15: **DeExtreme Model Card.** We follow the framework presented in Mitchell et al. [45].