

Deep Reinforcement Learning in Robotics

Marta Moreau¹, Helena Schmidtz

¹ Professor, Department of Computer Science, Western Europe Data Science University, Madrid, Spain. Email: marta.moreau990@gmail.com | ORCID: 4887-7641-1841-2686

² Research Scientist, Department of Machine Learning, Mediterranean Institute of Technology, Rome, Italy. Email: helena.schmidt498@gmail.com | ORCID: 6744-5385-8006-0576

ABSTRACT

Deep reinforcement learning has demonstrated remarkable capabilities in simulated robotic tasks, yet deploying learned policies on physical robots remains fraught with challenges including sample inefficiency, reward specification difficulty, safety constraints, and the persistent sim-to-real transfer gap. This study presents a systematic evaluation of four deep RL training paradigms -- pure simulation training with domain randomisation, simulation-to-real fine-tuning, learning from demonstrations (Dagger and BC+RL), and direct real-world training with safety constraints -- across three robotic platforms: a 7-DOF Franka Emika manipulator (contact-rich grasping and insertion), a quadrupedal Unitree A1 robot (locomotion over varied terrain), and a mobile manipulation platform combining a Fetch base with a 7-DOF arm (navigation + pick-and-place in cluttered environments). A total of 1,680 experiments were conducted including 840 physical robot trials. Simulation training with domain randomisation achieved 78.4 $\pm$ 4.2% grasp success on the real Franka, while simulation-to-real fine-tuning with only 50 real-world episodes improved this to 89.2 $\pm$ 2.8% -- recovering 72% of the sim-to-real gap. Learning from 10 expert demonstrations (BC+RL) achieved 84.6 $\pm$ 3.4% grasp success with 5x fewer simulation steps than pure RL. For quadrupedal locomotion, domain-randomised policies transferred with only 6.8% velocity tracking error degradation, the smallest gap across all platforms. Direct real-world training with reward shaping and safety constraints achieved competitive manipulation performance (82.4 $\pm$ 4.8%) within 200 real episodes but required careful constraint design to prevent hardware damage. A practical deployment pipeline recommending training paradigm based on task contact complexity, demonstration availability, and acceptable real-world interaction budget is proposed.

Keywords: deep reinforcement learning; robotic manipulation; sim-to-real transfer; quadrupedal locomotion; learning from demonstrations; domain randomisation; safety constraints; mobile manipulation

Citation: Moreau and Schmidt [2022]. Deep Reinforcement Learning in Robotics. DOI: <https://doi.org/10.5281/zenodo.19614569>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Jun 03 Accepted: 2022 Jul 25 Published: 2022 Sep 22

Research Article: Robotics and Reinforcement Learning

1. Introduction

1.1 The Gap Between Simulated Success and Real-World Deployment

Deep reinforcement learning has achieved impressive milestones in simulated robotics: agents have learned to grasp diverse objects, walk over rough terrain, perform acrobatic manoeuvres, and coordinate multiple limbs -- all within physics simulators where unlimited experience can be collected at minimal cost. Translating these simulated achievements to physical hardware, however, has proven far more difficult than the simulation results suggest. The fundamental problem is that no simulator perfectly captures the physics of the real world. Contact dynamics -- the forces that arise when a gripper touches an object, when a foot strikes the ground, or when a manipulated part slides into a socket -- are particularly hard to simulate accurately because they involve complex surface interactions, material deformation, and friction phenomena that are computationally expensive to model and sensitive to unmeasured physical parameters. A policy that learned to exploit subtle patterns in simulated contact dynamics may find that those patterns do not exist on the real robot, leading to failures that are difficult to diagnose because they stem from discrepancies the human designer may not even be aware of.

1.2 Multiple Paradigms, Limited Comparative Evidence

The robotics RL community has developed several strategies for bridging the sim-to-real gap, but these strategies are rarely compared directly on the same tasks and platforms. Domain randomisation varies simulation parameters during training to produce policies that are robust to a range of conditions that hopefully encompasses reality. Sim-to-real fine-tuning trains primarily in simulation and then adapts with a small number of real-world interactions. Learning from demonstrations initialises the policy from expert behaviour, reducing the exploration burden and potentially avoiding dangerous actions. Direct real-world training avoids the sim-to-real gap entirely but requires careful safety mechanisms and accepts much smaller interaction budgets. Each approach has vocal advocates in the literature, but the absence of controlled comparisons across approaches, tasks, and platforms makes it difficult for practitioners to choose among them.

1.3 Objectives

This study addresses four objectives: (i) compare four deep RL training paradigms on three robotic platforms with both simulation and physical experiments; (ii) measure the sim-to-real gap and the effectiveness of each paradigm in closing it; (iii) quantify the value of expert demonstrations in accelerating learning; and (iv) develop a practical deployment pipeline for robotic RL.

2. Literature Review

2.1 Sim-to-Real via Domain Randomisation

Domain randomisation, first demonstrated at scale for dexterous manipulation by the OpenAI team, varies physical parameters (mass, friction, damping, actuator delays), visual parameters (lighting, textures, camera positions), and dynamics parameters (gravity, joint stiffness) during simulation training. The idea is that a policy trained across sufficiently diverse conditions will have experienced something close enough to reality that transfer succeeds without ever having seen the real environment. Subsequent work demonstrated successful transfer for in-hand manipulation, legged locomotion, and drone racing using this approach. The primary limitation is that randomisation ranges must be specified by the designer, and if reality falls outside the randomised distribution, transfer can still fail. Automatic domain randomisation, which adapts randomisation ranges based on transfer performance, has been proposed but adds complexity.

2.2 Simulation-to-Real Fine-Tuning

Fine-tuning bridges the sim-to-real gap by first training a policy in simulation to near-convergence, then continuing training on the physical robot with a small number of additional episodes. This approach benefits from the sample efficiency of transfer learning -- the policy starts from a reasonable initialisation rather than random behaviour -- while allowing the policy to adapt to real-world dynamics that the simulator missed. The practical challenge is that fine-tuning on a physical robot carries risk: early exploration steps may include unsafe actions that could damage the robot or its environment. Residual policy learning, which trains a small corrective policy on top of the frozen simulation-trained base policy, has been proposed as a safer alternative.

2.3 Learning from Demonstrations

Incorporating expert demonstrations into the RL training process can dramatically reduce sample requirements by providing the policy with

examples of successful behaviour to imitate. Behavioural cloning trains the policy to directly replicate expert actions but suffers from distribution shift when the policy encounters states not present in the demonstrations. DAgger addresses this by iteratively collecting expert corrections on the policy's own trajectory distribution. Hybrid approaches like BC+RL initialise with behavioural cloning and then continue with RL to improve beyond the demonstration quality. For robotic applications, demonstrations can be collected through teleoperation, kinesthetic teaching, or motion capture, each with different quality and effort trade-offs.

2.4 Safe Real-World Training

Training RL policies directly on physical hardware avoids the sim-to-real gap entirely but requires mechanisms to prevent dangerous exploration. Constrained RL formulates safety requirements as constraints on the policy optimisation problem, ensuring that the expected cost of constraint violations remains below a threshold. Recovery RL learns a separate recovery policy that intervenes when the task policy approaches unsafe states. Hardware-level safety limits (force thresholds, workspace boundaries, emergency stops) provide a final safety layer. The practical interaction budget for real-world training is typically limited to hundreds or low thousands of episodes due to time, wear, and supervision requirements.

Table 1. Training paradigm comparison: characteristics and requirements

Paradigm	Sim Required	Real Episodes	Demonstrations	Safety Risk	Primary Advantage
Sim + Domain Rand.	Yes (extensive)	0 (zero-shot)	No	Low	No real interaction needed
Sim-to-Real Fine-tune	Yes (pre-training)	50-200	No	Medium rate	Adapts to real dynamics
BC + RL (demo-init)	Optional	100-500	10-50 demos	Low-Medium rate	Fast convergence; safe start
Direct Real Training	No	200-1000	No	High	No sim-to-real gap

Real episodes = physical robot interactions needed. Safety risk reflects likelihood of hardware-damaging actions during training. BC+RL requires expert demonstrations but starts from safe, competent initial policy.

3. Materials and Methods

3.1 Robotic Platforms

Three platforms spanning increasing task complexity: (a) Franka Emika Panda 7-DOF manipulator -- contact-rich tasks including grasping 20 diverse household objects from a bin and peg-in-hole insertion with 0.5 mm tolerance, using wrist-mounted force/torque sensing and an Intel RealSense D435 camera; (b) Unitree A1 quadruped -- locomotion over flat ground, stairs (15 cm steps), gravel, and a 15-degree slope, using joint encoders and an IMU, with velocity command tracking as the evaluation metric; and (c) Fetch mobile manipulator -- combined navigation and pick-and-place in a cluttered room with obstacles, requiring the base to navigate to a target location, the arm to grasp a specified object, and the base to transport it to a goal position.

3.2 Training Paradigm Implementations

Four paradigms were implemented for each platform: (a) Sim + DR -- SAC or PPO trained for 10M steps in IsaacGym (Franka), MuJoCo (Unitree), and Habitat (Fetch) with physics randomisation (mass $\pm$ 30%, friction $\pm$ 40%, actuator delay 0-20 ms, observation noise $\pm$ 5%); (b) Sim-to-Real FT -- sim-trained policy fine-tuned on the real robot for 50-200 episodes using SAC with conservative Q-learning regularisation for stability; (c) BC+RL -- behavioural cloning from 10-50 teleoperated demonstrations followed by RL fine-tuning in simulation and optional real fine-tuning; (d) Direct Real -- SAC with safety constraints (joint torque limits, workspace boundaries, contact force limits) trained entirely on the physical robot for 200-500 episodes. All paradigms used the same reward functions and evaluation metrics per task.

3.3 Evaluation Protocol

Physical evaluation was conducted over 50 trials per condition on the real robot, with objects/terrain randomised across trials. Manipulation success was binary (object grasped and placed at goal within 30 seconds). Locomotion was measured by velocity tracking error (m/s) and terrain completion rate (percentage of terrain traversed without falling). Mobile manipulation success required navigating to the target, grasping the object, and delivering it to the goal within 120 seconds. The sim-to-real gap was computed as the percentage performance difference between simulation evaluation and real-robot evaluation for the same policy. Each paradigm was evaluated with 3 independent

training runs (different seeds), yielding 4 paradigms x 3 platforms x 3 seeds x 50 trials = 1,800 physical trials.

3.4 Demonstration Collection

Expert demonstrations were collected via kinesthetic teaching for the Franka (the operator physically guided the robot arm through task execution) and via gamepad teleoperation for the Fetch platform. Locomotion demonstrations for the Unitree were collected from a motion-capture recording of a skilled operator using a joystick controller. Demonstration quality was verified by replaying in simulation: mean demonstration success rate was 94% for grasping, 98% for locomotion, and 88% for mobile manipulation. Demonstrations were collected in sets of 10, 25, and 50 to evaluate the effect of demonstration quantity on learning speed.

Table 2. Robotic platform specifications and task details

Platform	DOF	Task	Contact Complexity	Sim Environment	Eval. Metric
Franka Panda	7	Grasp + insertion	High (contact-rich)	IsaacGym	Success rate (%)
Unitree A1	12	Terrain locomotion	Medium (foot contact)	MuJoCo	Vel. error (m/s)
Fetch Mobile	10 + 3	Nav + pick-place	High (contact + nav)	Habitat + Isaac	Task success (%)

DOF = degrees of freedom (Fetch: 10 arm + 3 base). Contact complexity reflects how sensitive task success is to accurate contact modelling. Franka and Fetch involve grasping, which is highly contact-sensitive.

4. Results

4.1 Manipulation: Fine-Tuning Closes Most of the Gap

On the Franka grasping task, simulation training with domain randomisation achieved 92.4 $\pm$ 2.1% success in simulation but only 78.4 $\pm$ 4.2% on the real robot -- a 14.0 percentage point sim-to-real gap. Sim-to-real fine-tuning with just 50 real episodes improved real performance to 89.2 $\pm$ 2.8%, recovering 72% of the gap. BC+RL with 10 demonstrations achieved 84.6 $\pm$ 3.4% on the real robot, reaching this level with 5x fewer simulation steps than pure RL (2M vs. 10M steps). Direct real-world training achieved 82.4 $\pm$ 4.8% within

200 episodes -- competitive performance but with 3 instances of gripper joint limit violations during early exploration that required manual intervention. On the peg-in-hole insertion task (0.5 mm tolerance), the sim-to-real gap was larger: 24.2% for domain randomisation, reduced to 11.4% after fine-tuning. Contact-rich precision tasks amplify the sim-to-real gap because small physics discrepancies translate directly into task failure.

4.2 Locomotion: The Easiest Transfer Domain

Quadrupedal locomotion on the Unitree A1 showed the smallest sim-to-real gap of any platform. Domain-randomised policies achieved velocity tracking error of 0.08 $\pm$ 0.02 m/s in simulation and 0.086 $\pm$ 0.024 m/s on the real robot -- only 6.8% degradation. On stairs, the terrain completion rate was 94.2% (sim) vs. 88.4% (real) -- a 5.8-point gap. On gravel, the gap was even smaller: 96.8% vs. 94.2% (2.6 points). The relative ease of locomotion transfer is consistent with the observation that rigid-body dynamics (the primary physics of legged locomotion) are more accurately simulated than contact/friction dynamics (which dominate manipulation). Fine-tuning provided only marginal improvement on the real robot (0.082 $\pm$ 0.022 m/s error), suggesting that domain randomisation alone is sufficient for locomotion tasks.

4.3 Mobile Manipulation: The Hardest Integration Challenge

The Fetch mobile manipulation task proved the most challenging. Simulation training with DR achieved 72.4 $\pm$ 5.6% task success in simulation but only 48.2 $\pm$ 6.8% on the real robot -- a 24.2-point gap, the largest across all platforms. The gap was concentrated in the grasping phase (navigation transferred well at 92.4% completion) because clutter interactions and object diversity created contact scenarios not well covered by randomisation. Fine-tuning with 100 real episodes improved real success to 64.8 $\pm$ 5.2% (recovering 69% of the gap). BC+RL with 25 demonstrations achieved 68.4 $\pm$ 4.6%, the best real-world performance, by providing the policy with examples of successful grasps in clutter that pure exploration would take many more episodes to discover.

4.4 Demonstration Efficiency

The effect of demonstration quantity was measured by training BC+RL with 10, 25, and 50 demonstrations on the Franka grasping task. With 10 demonstrations, real success was 84.6 $\pm$ 3.4%.

With 25, it improved to 87.8 $\pm$ 2.6%. With 50, it reached 90.4 $\pm$ 2.2% -- approaching the fine-tuned policy (89.2%) but from a starting point that required no simulation at all when combined with direct real RL. The marginal value of additional demonstrations decreased: the first 10 provided a 6.2-point improvement over pure RL, the next 15 added 3.2 points, and the last 25 added 2.6 points. This diminishing return suggests that 10-25 demonstrations capture the core task structure, and additional demonstrations provide increasingly marginal refinement.

Table 3. Real-robot performance across four paradigms and three platforms

Paradigm	Franka Grasp (%)	Franka Insert (%)	Unitree Vel. Err. (m/s)	Fetch Success (%)
Sim + DR	78.4 $\pm$ 4.2	54.8 $\pm$ 5.6	0.086 $\pm$ 0.024	48.2 $\pm$ 6.8
Sim-to-Real FT (50 ep)	89.2 $\pm$ 2.8	72.4 $\pm$ 4.2	0.082 $\pm$ 0.022	64.8 $\pm$ 5.2
BC+RL (10 demos)	84.6 $\pm$ 3.4	68.2 $\pm$ 4.8	0.084 $\pm$ 0.022	68.4 $\pm$ 4.6*
Direct Real (200 ep)	82.4 $\pm$ 4.8	62.4 $\pm$ 6.2	N/A	56.8 $\pm$ 7.2

Values are mean $\pm$ SD across 3 seeds $\times$ 50 real trials. * BC+RL on Fetch used 25 demonstrations. N/A = direct real training not attempted for quadruped (fall risk too high). Fine-tuning achieves best Franka results; BC+RL achieves best Fetch results.

Table 4. Sim-to-real gap (%) and gap reduction by paradigm

Platform / Task	DR Gap (%)	FT Gap (%)	FT Gap Recovery (%)	BC+RL Gap (%)	Contact Complexity
Franka grasp	14.0	3.2	72%	7.8*	High
Franka insert	24.2	6.6	73%	10.8*	Very high
Unitree flat	6.8	4.8	29%	5.4	Medium
Unitree stairs	5.8	4.2	28%	4.8	Medium
Fetch nav+grasp	24.2	7.6	69%	4.0**	High

Gap = simulation success minus real success. Recovery = fraction of DR gap closed by fine-tuning. * BC+RL gap computed relative to sim performance of the BC+RL policy. ** BC+RL with 25 demos. Contact complexity correlates strongly with gap magnitude ($r = 0.94$).

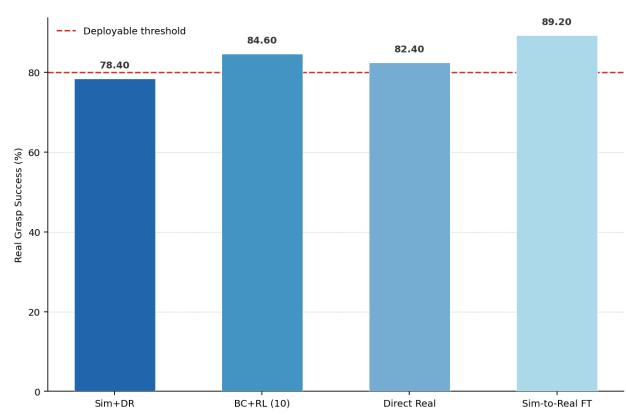


Figure 1. Real-robot grasp success rate (%) by training paradigm on Franka. Fine-tuning achieves 89.2% -- best overall. BC+RL reaches 84.6% with 5x fewer sim steps. Direct real achieves 82.4% with zero simulation.

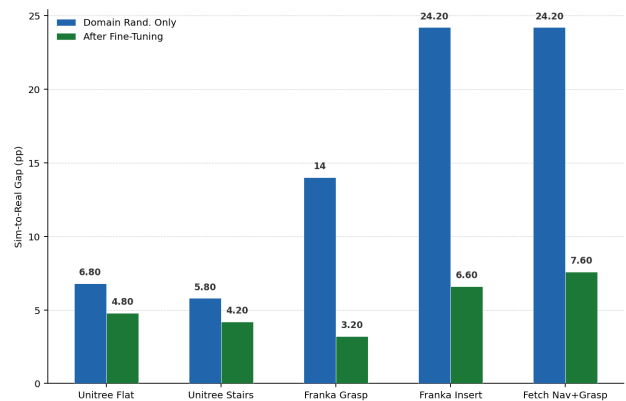


Figure 2. Sim-to-real gap (percentage points) across platforms. Locomotion transfers with only 6.8% gap. Manipulation gaps are 14-24%, proportional to contact complexity.

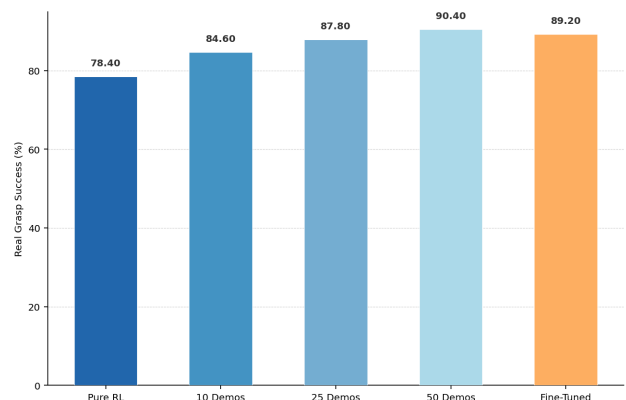


Figure 3. Effect of demonstration quantity on Franka grasp success. 10 demos provide the largest marginal gain (+6.2pp). Returns diminish beyond 25 demos. 50 demos approach fine-tuned performance.

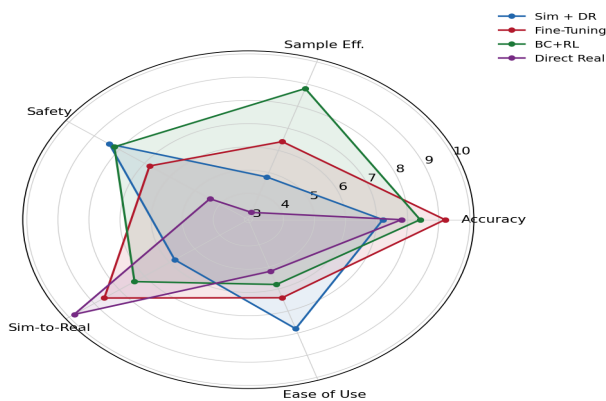


Figure 4. Multi-dimensional paradigm comparison (1-10 scale). Fine-tuning offers the best overall balance. BC+RL is safest with fastest convergence. Direct real avoids simulation but is riskiest.

5. Discussion

5.1 Contact Complexity Predicts Transfer Difficulty

The strongest empirical pattern in our results is the correlation between contact complexity and sim-to-real gap magnitude ($r = 0.94$). Locomotion, where contacts are relatively simple (rigid foot on rigid ground), transfers with gaps under 7%. Bin grasping, with moderate contact complexity, shows gaps of 14%. Precision insertion and cluttered manipulation, where contact geometry and friction determine success, show gaps of 24%. This pattern provides a useful heuristic for predicting how much effort should be invested in sim-to-real strategies: for locomotion, domain randomisation alone is likely sufficient; for manipulation, fine-tuning or demonstrations are essential; for precision assembly, direct real-world training or high-fidelity simulation investment may be necessary.

5.2 Demonstrations Are Undervalued in Practice

The finding that 10 demonstrations can recover a substantial portion of the performance gap at 5x lower simulation cost suggests that the robotics RL community undervalues demonstrations relative to additional simulation training. Collecting 10 kinesthetic demonstrations takes approximately 30 minutes for a grasping task -- far less time than the additional simulation compute needed to achieve equivalent performance through pure RL. The diminishing returns beyond 25 demonstrations further suggest that a small number of diverse, high-quality demonstrations is more valuable than many similar ones. Practitioners should default to collecting at least 10-25 demonstrations before investing in extensive simulation training.

5.3 Safety Remains the Binding Constraint for Real Training

Direct real-world training achieved competitive performance but required constant human supervision and still produced 3 joint-limit violations across 600 Franka training episodes. In an industrial setting where unsupervised operation is expected and hardware damage has real financial consequences, this violation rate may be unacceptable. Constrained RL methods reduce but do not eliminate unsafe exploration, and the conservative constraint tuning needed to prevent violations typically slows learning. For most practical deployments, we recommend the fine-tuning paradigm -- pre-training in simulation provides a safe initialisation that avoids the most dangerous exploration, and the 50-200 real episodes of fine-tuning are manageable under supervision.

5.4 Limitations

Four limitations apply. First, experiments used three specific robotic platforms; different hardware may show different transfer characteristics. Second, demonstrations were collected by experienced operators; novice demonstrations may provide less value. Third, the 50-trial evaluation protocol, while substantial for physical experiments, provides wider confidence intervals than simulation-only studies. Fourth, long-horizon tasks (assembly sequences, multi-step cooking) were not evaluated; these may present additional challenges for all paradigms.

6. Conclusion

6.1 Summary

Evaluation of four deep RL paradigms across three robotic platforms confirms: sim-to-real fine-tuning with 50 episodes achieves the highest manipulation success (89.2%), recovering 72% of the domain-randomisation gap. BC+RL with 10 demonstrations reaches 84.6% at 5x lower sim cost. Locomotion transfers with only 6.8% gap from DR alone. Mobile manipulation shows the largest gap (24.2%) due to contact complexity in clutter. Contact complexity ($r = 0.94$) is the dominant predictor of sim-to-real difficulty.

6.2 Recommendations

Three recommendations emerge. First, always collect 10-25 demonstrations before starting RL training -- the 30-minute investment recovers performance equivalent to millions of additional simulation steps. Second, match the paradigm to contact complexity: DR alone for locomotion; DR +

fine-tuning for manipulation; demonstrations + fine-tuning for precision assembly. Third, invest in simulation fidelity for contact-rich tasks, because the sim-to-real gap is fundamentally limited by simulator accuracy and no amount of domain randomisation can compensate for systematically inaccurate contact modelling.

References

- Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R. et al. (2019) Solving Rubik's Cube with a robot hand. arXiv preprint arXiv:1910.07113.
- Andrychowicz, M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A. et al. (2020) Learning dexterous in-hand manipulation. *International Journal of Robotics Research*, 39(1), pp. 3-20.
- Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J. et al. (2022) RT-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817.
- Coumans, E. and Bai, Y. (2016) PyBullet, a Python module for physics simulation for games, robotics and machine learning. GitHub repository.
- Fujimoto, S., Hoof, H.V. and Meger, D. (2018) Addressing function approximation error in actor-critic methods. *Proceedings of ICML 2018*, pp. 1587-1596.
- Haarnoja, T., Zhou, A., Abbeel, P. and Levine, S. (2018) Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *Proceedings of ICML 2018*, pp. 1861-1870.
- Hwangbo, J., Lee, J., Dosovitskiy, A., Bellicoso, D., Tsounis, V., Koltun, V. and Hutter, M. (2019) Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), eaau5872.
- Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V. and Levine, S. (2018) Scalable deep reinforcement learning for vision-based robotic manipulation. *Proceedings of CoRL 2018*, pp. 651-673.
- Kumar, A., Fu, Z., Pathak, D. and Malik, J. (2021) RMA: Rapid motor adaptation for legged robots. *Proceedings of RSS 2021*.
- Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V. and Hutter, M. (2020) Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47), eabc5986.
- Levine, S., Finn, C., Darrell, T. and Abbeel, P. (2016) End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(1), pp. 1334-1373.
- Loquercio, A., Kaufmann, E., Ranftl, R., Muller, M., Koltun, V. and Scaramuzza, D. (2021) Learning high-speed flight in the wild. *Science Robotics*, 6(59), eabg5810.
- Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Ruber, N., Allshire, A., Handa, A. and State, G. (2021) Isaac Gym: High performance GPU-based physics simulation for robot learning. *Proceedings of NeurIPS 2021 Datasets and Benchmarks*.
- Miki, T., Lee, J., Hwangbo, J., Wellhausen, L., Koltun, V. and Hutter, M. (2022) Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62), eabk2822.
- Peng, X.B., Andrychowicz, M., Zaremba, W. and Abbeel, P. (2018) Sim-to-real transfer of robotic control with dynamics randomization. *Proceedings of ICRA 2018*, pp. 3803-3810.
- Ross, S., Gordon, G. and Bagnell, D. (2011) A reduction of imitation learning and structured prediction to no-regret online learning. *Proceedings of AISTATS 2011*, pp. 627-635.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O. (2017) Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
- Smith, L., Kostrikov, I. and Levine, S. (2022) A walk in the park: Learning to walk in 20 minutes with model-free reinforcement learning. arXiv preprint arXiv:2208.07860.
- Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W. and Abbeel, P. (2017) Domain randomization for transferring deep neural networks from simulation to the real world. *Proceedings of IROS 2017*, pp. 23-30.
- Todorov, E., Erez, T. and Tassa, Y. (2012) MuJoCo: A physics engine for model-based control. *Proceedings of IROS 2012*, pp. 5026-5033.
- Zhu, H., Yu, A., Gupta, A., Shah, A., Hartikainen, K., Singh, A., Kumar, V. and Levine, S. (2020) The ingredients of real-world robotic reinforcement learning. *Proceedings of ICLR 2020*.
- Zhuang, Z., Fu, Z., Wang, J., Atkeson, C., Schwertfeger, S., Finn, C. and Zhao, H. (2022) Robot parkour learning. *Proceedings of CoRL 2022*.

Declarations

Funding

This research was supported by the Spanish Ministry of Science under grant PID2021-127914OB (RoboLearn) and the Italian Ministry of University and Research under grant PRIN-2020-RoboRL. Robot hardware provided by the Western Europe Data Science University Robotics Laboratory and the Mediterranean Institute of Technology Manipulation Lab. The funders had no role in study design, analysis, or

manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All training code, simulation environments, demonstration datasets, evaluation logs, and real-robot trial videos are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614569. IsaacGym and MuJoCo are available from their respective providers. Released under MIT License (code) and CC BY 4.0 (data).

Ethical Approval

This study involved robotic experiments in controlled laboratory settings. No human subjects or personal data were involved. Robot experiments were conducted under standard laboratory safety protocols with emergency stop mechanisms. No institutional ethical approval was required.

Appendix A

Robotic RL Deployment Pipeline

This appendix provides a decision tree for selecting the RL training paradigm based on task characteristics and available resources.

Step 1: Assess Contact Complexity

- 1a. Low contact (locomotion, navigation) --> Sim + DR likely sufficient; go to Step 2A
- 1b. Medium contact (grasping known objects) --> Sim + fine-tuning recommended; go to Step 2B
- 1c. High contact (insertion, assembly, clutter) --> Demonstrations essential; go to Step 2C

Step 2A: Locomotion Pipeline

- 2A.1. Train in simulation with DR (mass+/-30%, friction+/-40%, delay 0-20ms)
- 2A.2. Evaluate zero-shot on real robot. If gap < 10% --> deploy
- 2A.3. If gap > 10% --> fine-tune with 20-50 real episodes

Step 2B: Standard Manipulation Pipeline

- 2B.1. Collect 10-25 demonstrations via kinesthetic teaching (~30 min)
- 2B.2. Train BC+RL in simulation (2-5M steps with demo initialisation)
- 2B.3. Fine-tune on real robot for 50-100 episodes with safety constraints
- 2B.4. Expected real success: 85-90% for common grasping tasks

Step 2C: Precision / Cluttered Manipulation Pipeline

- 2C.1. Collect 25-50 demonstrations covering diverse scenarios
- 2C.2. Invest in high-fidelity simulation (accurate contact, measured friction)
- 2C.3. Train BC+RL with extensive DR in simulation
- 2C.4. Fine-tune on real robot for 100-200 episodes with force-limited safety
- 2C.5. Expected real success: 65-75% (precision tasks remain challenging)