Enhancing Sim-to-Real Transfer Learning with PPO and Domain Randomization

Carl Zhang, Ruichen Zhao
Department of Computer Science,
Duke University,
Durham, NC
Email: {carl.zhang, ruichen.zhao}@duke.edu

Abstract— **This study explores the challenge of sim-to-real transfer, focusing on how discrepancies between simulated environments and real-world conditions affect agent performance. Using the CartPole environment as a test base, we examine the effects of various simulation modifications, including friction dynamics, observation noise, and curriculum learning. We employ Proximal Policy Optimization (PPO) to train policies under different conditions, comparing performance between agents trained with standard environments, domain randomization, and progressive difficulty adjustments (curriculum learning).**

Our experimental results show that while domain randomization improves generalization in environments with unseen variations, curriculum learning provides a smoother progression but does not always outperform direct training in harder conditions. We further evaluate the robustness of trained models by introducing unseen friction values and dynamic environmental perturbations. This exploratory work highlights the strengths and limitations of different sim-to-real strategies, providing insights into the adaptability of reinforcement learning agents under varying simulation complexities.

Keywords: Sim-to-Real transfer, reinforcement learning, PPO, domain randomization, curriculum learning, simulation dynamics.

I. INTRODUCTION

Robotic systems trained in simulated environments often fail to achieve adequate performance when deployed in real-world settings due to discrepancies in physical dynamics, sensor inaccuracies, and environmental variability. Bridging this sim-to-real gap is critical for applications such as autonomous driving, industrial automation, and service robotics, where operational precision and adaptability are non-negotiable.

Proximal Policy Optimization (PPO), recognized for its balance between computational simplicity and sample efficiency, serves as a strong foundation for training robust policies in reinforcement learning (RL). Domain randomization, a technique that systematically varies simulation parameters (e.g., friction, noise, and dynamics) during training, further enhances the generalizability of learned policies by exposing agents to a diverse set of simulated conditions.

In this work, we extend and integrate these concepts by:

1. Implementing a PPO-based framework with an enhanced domain randomization protocol.

2. Performing a comparative performance analysis of policies trained under standard, randomized, and progressively difficult environments.

3. Developing a scalable and adaptive simulation framework to evaluate the robustness of RL agents against real-world uncertainties.

Our contributions aim to explore how dynamic adjustments in simulation parameters impact policy performance, providing insights into the strengths and limitations of domain randomization and curriculum learning for sim-to-real transfer.

II. RELATED WORK

Sim-to-real transfer is a pivotal concern in robotics, where the goal is to leverage the efficiency of simulations for robust real-world applications. The PPO algorithm has been instrumental in advancing reinforcement learning, particularly due to its effectiveness in environments with high-dimensional action spaces [1]. On the other hand, domain randomization has been proposed as a technique to improve the transferability of simulation-trained models by introducing variability in simulation parameters such as lighting conditions, textures, and physical properties [2]. While these approaches have independently shown promise, there is a gap in research that combines these methods with adaptive algorithms to address the dynamic nature of real-world environments. Our work builds on these foundations and addresses existing gaps by:

1. Integrating PPO with a dynamic domain randomization protocol that adjusts simulation parameters over time.

2. Exploring the impact of progressively challenging environments (curriculum learning) on policy robustness.

3. Conducting an extensive comparative analysis to evaluate generalization across different environmental variations, such as varying levels of friction and observation noise.

By systematically combining domain randomization, PPO, and curriculum learning, our study provides new insights into creating more resilient and adaptable policies for sim-to-real transfer tasks.

III. METHODS

A. PPO Implementation

Our implementation of Proximal Policy Optimization (PPO) optimizes a clipped surrogate objective function to ensure stable and controlled updates during training [1]. The clipping mechanism constrains the policy updates, preventing large deviations from previous policies, which is particularly advantageous when combined with domain randomization. By systematically exposing the policy to a variety of conditions, the clipped objective enables stable learning across diverse scenarios without overfitting to any single environment.

To integrate domain randomization effectively, we dynamically adjust the environmental parameters at the start of each training epoch. This process ensures that the policy continually adapts to new conditions, mimicking incremental learning observed in natural and engineered systems. Such an approach encourages the development of generalized policies capable of handling a wide range of dynamic and uncertain real-world environments.

B. *Domain Randomization Implementation*

Our domain randomization strategy dynamically varies key simulation parameters at the start of each episode. By introducing controlled stochasticity into the environment, the agent encounters a diverse range of training conditions, improving its generalizability and robustness. The following parameters were randomized within empirically chosen, realistic bounds to reflect potential real-world variations:

Gravity: Simulate gravitational inconsistencies.

Pole Mass: Reflecting changes in mechanical properties.

Cart Friction: Introducing resistance to cart movement.

Sensor Noise: Gaussian noise with a standard deviation of 0.5 was added to observations to simulate sensor inaccuracies.

C. *Environment Setup*

We used the CartPole-v1 environment as a baseline for initial testing due to its simplicity and widespread use as a benchmark in reinforcement learning studies. To enhance the realism of the simulation, we extended the standard CartPole environment by introducing frictional dynamics and parameter variability, as described above.

To test the effectiveness of our training methods, we created a customized version of CartPole, named CartPoleWithFriction, where velocities of the cart and pole are dampened to simulate energy loss. Randomized environments were configured by dynamically adjusting the simulation parameters at runtime. This extended environment setup serves as an intermediate step toward bridging the sim-to-real gap, allowing us to evaluate the adaptability and generalizability of PPO-based policies under increasing environmental complexity.

D. *Training Procedure*

We trained and evaluated two sets of models for comparative analysis:

Standard Environment: Policies were trained in the unmodified CartPole-v1 environment with fixed parameters.

Randomized Environment: Policies were trained in our domain-randomized environment, where key parameters were varied across episodes.

Each model was trained for 10,000 timesteps, with periodic evaluations conducted to monitor training progress and performance. We adjusted critical hyperparameters such as the learning rate and clipping range based on training feedback to ensure optimal convergence. Figure 1 illustrates the training progress of the model in the standard environment.

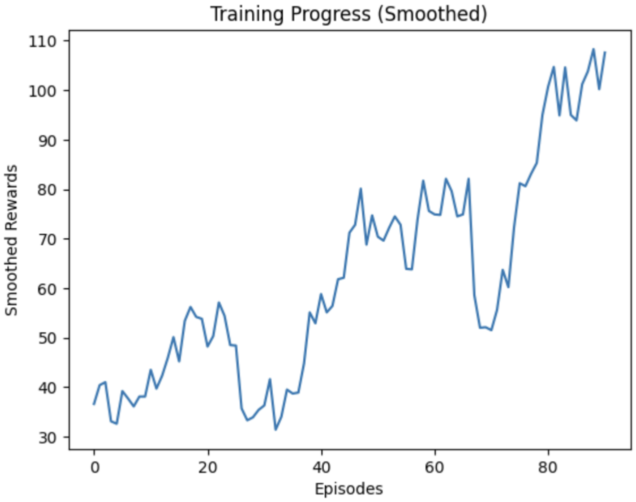


Fig. 1: Training Progress

E. *Evaluation Protocol*

Our evaluation focuses on both the average reward and the variability of rewards across episodes, providing a dual perspective on the effectiveness and consistency of the trained policies. This comprehensive evaluation helps in assessing not only the peak performance but also the reliability of the policies under varying conditions.

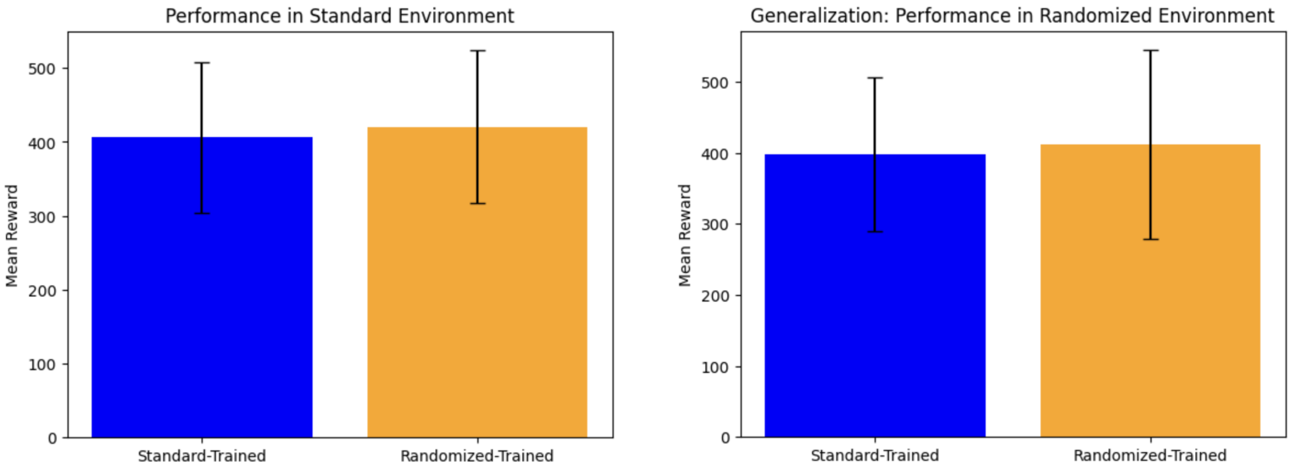
IV. Detailed Findings

A. *Domain Randomization Results*

In the domain randomization tests, the models trained in randomized environments outperformed those trained in standard environments when tested in both settings. Figure 2. Key observations include:

- In the randomized environment, the standard-trained model achieved a mean reward of 397.36 with a standard deviation of 107.91. In contrast, the randomized-trained model showed improved performance with a mean reward of 411.74 and a standard deviation of 132.42.
- In the standard environment, the standard-trained model recorded a mean reward of 406.14 with a standard deviation of 102.15, while the randomized-trained model exhibited a superior mean reward of 420.32 and a standard deviation of 103.42.

These results underscore the improved generalization capabilities of policies trained with domain randomization, achieving higher rewards and adaptability across diverse settings.



(a) Standard Training

(b) Randomized Training

Fig. 2: Domain Randomization Results

B. *Sim-to-Sim Training with a More Complex Environment*

When subjected to a more complex environment featuring additional simulation parameters, the findings were presented in Figure 3, and specifically:

- The standard-trained model achieved a mean reward of 349.6 and a standard deviation of 60.77.
- The randomized-trained model showed an enhanced performance with a mean reward of 363.1 and a standard deviation of 51.89.

Further exploration involved fine-tuning both models in this complex environment, yielding significant improvements:

- The fine-tuned standard model dramatically increased its performance to a mean reward of 497.8 with a minimal standard deviation of 7.46.
- The fine-tuned randomized model also improved, achieving a mean reward of 483.3 with a standard deviation of 24.64.

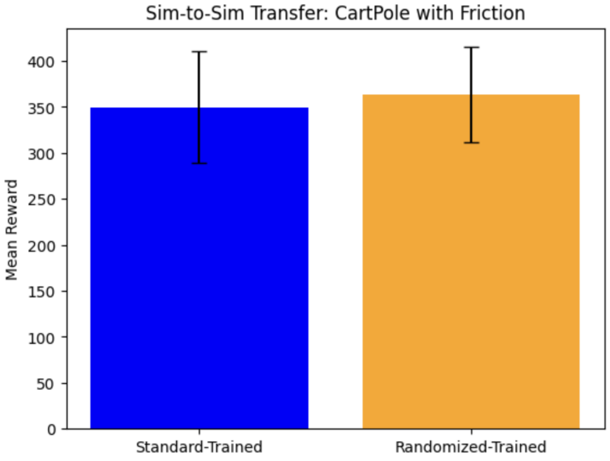


Fig. 3: Sim-to-Sim Evaluation Result.

These results highlight the importance of fine-tuning policies in increasingly complex settings, where initial training provides a strong foundation, and subsequent adaptation maximizes performance.

C. *Exploration in Noisy Environments*

The robustness of the models was evaluated in environments with varying levels of sensor noise (Figure 4):

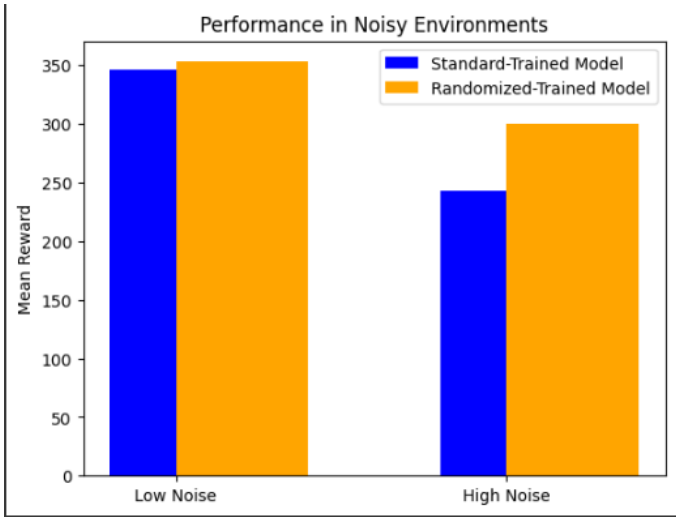


Fig. 4: Performance in Noisy Environments.

1) High-Noise Environment: In high-noise settings, the results indicated that:

- The **Standard-Trained Model** achieved a mean reward of 243.05 with a standard deviation of 100.74, reflecting a considerable decrease in performance under challenging conditions.
- The **Randomized-Trained Model** performed better, with a mean reward of 299.75 and a standard deviation of 107.04. This suggests that the randomized training

approach enhances the model's ability to handle unexpected environmental noise.

2) Low-Noise Environment: In environments with lower levels of noise, both models exhibited improved performance, yet the randomized-trained model still showed a slight advantage:

- The **Standard-Trained Model** recorded a mean reward of 345.8 and a standard deviation of 81.22, demonstrating respectable robustness in more controlled settings.
- Conversely, the **Randomized-Trained Model** achieved a mean reward of 352.95 with a standard deviation of 117.60, indicating not only better performance but also a higher variability in reward distribution, which could suggest a greater adaptability to subtle environmental variations.

3) Analysis and Implications: The results indicate that domain randomization enhances the model's ability to handle noisy and uncertain observations, achieving higher rewards under both high and low noise conditions. However, the increased variability (higher standard deviation) observed in the randomized-trained model suggests a trade-off between robust adaptability and performance consistency. This trade-off warrants further exploration to optimize stability while maintaining generalization capabilities.

V. Summary of Findings

Our findings demonstrate the effectiveness of domain randomization and its impact on model performance across a variety of test conditions:

- Domain Randomization: Policies trained with domain randomization consistently outperformed standard-trained models in both standard environments and randomized environments.
- Complex Environments: Fine-tuning models in increasingly complex simulations significantly improved performance, showcasing the potential of incremental adaptation for sim-to-real transfer.
- Noisy Environments: Randomized-trained models exhibited greater resilience to sensor noise, achieving higher rewards compared to standard-trained models. However, variability in performance (higher standard deviation) suggests a need for balancing adaptability and consistency.

Overall, the results validate the hypothesis that training with domain randomization enhances model generalization and robustness, preparing policies to handle real-world variability and uncertainties more effectively.

VI. Limitations and Future Work

A. Real-World Testing Constraints

A significant limitation of this study is the inability to test the trained models in real-world environments due to time and resource constraints. While simulated environments incorporating domain randomization and curriculum learning offer valuable insights, they remain approximations of real-world physics and sensor characteristics. The absence of real-world validation highlights a need for further testing to confirm the practical applicability of the learned policies.

B. Exploration of Curriculum Training

Curriculum training was explored as a method to gradually introduce challenges to the model, starting with easier tasks and progressively moving towards more complex scenarios. The hypothesis was that this approach would lead to a more robust learning process, enabling the model to adapt better to different levels of difficulty. However, the results were unexpected, as both the curriculum-trained model and the baseline model achieved the maximum reward of 500.0, with a standard deviation of 0.0 in both low and high friction environments Figure 5.

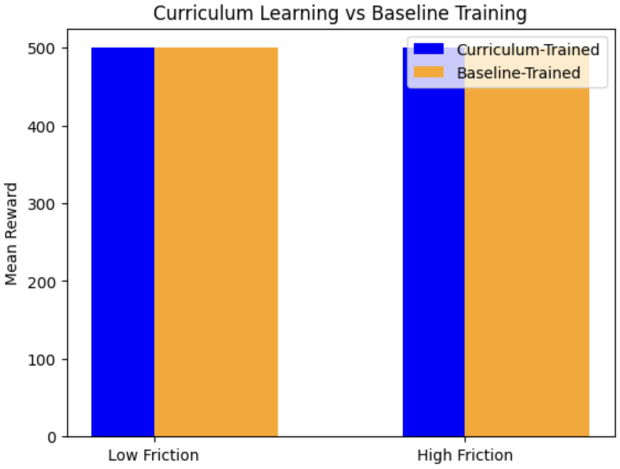


Fig. 5: Curriculum Training Result

1) Analysis of Maximum Reward Achievement: The uniform success of both models suggests the following explanations:

- **Saturation of Learning Capacity:** The CartPole task may lack sufficient complexity to challenge the models beyond a basic skill threshold. Both models may have quickly reached their optimal performance due to the simplicity of the environment.
- **Insufficient Differentiation in Curriculum Stages:** The progression of curriculum stages might not have introduced sufficient incremental challenges. If the difficulty increases too gradually or fails to test the agent's adaptation abilities, the curriculum provides limited benefit over direct training.

2) Implications and Further Research: These findings highlight the importance of carefully designing curriculum stages to ensure meaningful learning progression. Future work will involve:

- Developing tasks with greater environmental complexity to prevent premature convergence to optimal performance.
- Introducing more distinct and challenging stages to better assess adaptability.

VII. APPENDIX

A. Variability in Model Performance

While models trained with domain randomization generally demonstrated superior performance, variability in results was observed across different conditions. Notably, in high-noise environments, the standard-trained model occasionally outperformed the randomized-trained model, suggesting limitations in the robustness and consistency of the randomized policies.

1) Potential Reasons for Performance Variability: Several factors may explain the observed fluctuations:

- **Overfitting to Randomized Conditions:** While domain randomization exposes the model to a range of dynamics, there is a risk of overfitting to specific types of noise or randomness encountered during training. This can limit the model's ability to generalize to underrepresented conditions.
- **Algorithmic Stability:** The inherent sensitivity of PPO to initial conditions and hyperparameter settings can impact training stability. Suboptimal configurations may lead to inconsistencies in performance, particularly under high variability.

2) Implications and Further Research: Addressing these limitations requires a focus on improving model stability and generalization. Future research could include:

- **Enhanced Generalization:** Implementing techniques such as cross-validation within the training process to better generalize across different noise profiles and environmental conditions.
- **Hyperparameters Tuning:** Systematic hyperparameter tuning and sensitivity analysis could help identify optimal settings that minimize performance variability across different training scenarios.
- **Algorithmic Enhancements:** Exploring adaptive learning rates or alternative RL algorithms (e.g., SAC or TRPO) that may exhibit greater stability under challenging conditions.

REFERENCES

[1] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[2] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 23–30.