The Implementation of Test Time Augmentation in Deep Reinforcement Learning

Chenye Dai

South China University of Technology, Guangzhou University City, Guangzhou, China

Abstract. Deep Reinforcement Learning (DRL) models often struggle with generalization and robustness, requiring costly retraining to adapt to environmental changes. To address this, the study proposes Test Time Augmentation (TTA) as a post-training method to enhance the policy stability of DRL agents. This work introduces a novel approach that applies TTA to DRL by leveraging controlled state perturbation, majority voting, and dynamic scaling of augmentations. This method allows agents to adapt to varying conditions without modifying the original model parameters, offering a lightweight yet effective solution to improving robustness. Experimental results on the LunarLander-v2 environment using Deep Q-Networks (DQN) demonstrate a 4.78% performance improvement and a 9% success rate improvement under stable conditions and increased resilience against moderate noise. However, performance declines in highly chaotic environments, highlighting TTA's limitations under extreme randomness. Overall, this study bridges the gap between TTA in computer vision and DRL, offering insights into practical and computationally efficient methods for improving policy robustness without retraining.

1 Introduction

Test Time Augmentation (TTA) has been widely employed in computer vision to improve model robustness by applying controlled transformations to input data during inference. These augmentations help models handle distribution shifts and unforeseen variations without necessitating additional retraining. While computer vision models benefit significantly from TTA, deep reinforcement learning (DRL) algorithms exhibit notable generalization issues, particularly when deployed in dynamic, uncertain, or noisy environments. DRL agents are often trained within a well-defined distribution of environmental states, leading to overfitting and a lack of adaptability when encountering unfamiliar scenarios. This weakness limits their practical applicability, as real-world decision-making tasks frequently involve unpredictable environmental conditions and non-stationary data distributions.

Traditional techniques for improving DRL robustness, such as domain randomization, modify environmental parameters during training to expose agents to diverse conditions. While effective in certain scenarios, this approach requires significant computational

202230382226@mail.scut.edu.cn

resources and hyperparameter tuning, making it difficult to generalize across different environments. Similarly, fine-tuning trained models on new data can be computationally expensive and does not guarantee improved performance in out-of-distribution settings. Given these challenges, alternative methods for improving DRL robustness without retraining are highly desirable [1, 2, 3].

Motivated by the effectiveness of TTA in computer vision, this study investigates its application in DRL to address generalization challenges. A novel TTA framework is introduced, integrating state-space perturbations and adaptive decision-making mechanisms to enhance policy robustness during testing. This framework operates without modifying the original DRL model, making it computationally efficient and easily deployable. To validate its feasibility, the proposed method is tested on the LunarLander-v2 environment, a widely used DRL benchmark. Experimental evaluation provides insights into the effectiveness of TTA in DRL settings, highlighting its potential as a lightweight approach for enhancing policy stability in the presence of environmental uncertainty [4, 5, 6].

This study offers a detailed exploration of the ways in which Test Time Augmentation can be applied in DRL to improve the robustness of agents, particularly in noisy or dynamically changing environments. The research is hoped to stimulate more studies on the application of TTA in reinforcement learning systems, opening up a new direction for improving the generalization capability of DRL algorithms on real-world tasks.

The contributions of this study are threefold as follows.

(1) This research introduces a new augmentation method called dynamic augmentation scaling mechanism that scales state-space perturbations based on the agent's performance, allowing for more adaptive and context-dependent noise injection.

(2) By conducting extensive experiments to investigate the impact of TTA on agent performance as a function of varying noise intensities, defining the ranges over which TTA effectively improves robustness without compromising the ability of the agent to act optimally.

(3) This work also explains how the perturbations induced by noise can act as a form of regularization, preventing overfitting to single states and encouraging decision-making policies that are more generalizable.

2 Related Work

Test Time Augmentation originated in computer vision as a strategy for improving classification robustness by introducing random transformations into input data at inference time. The goal is to diminish the effect of noise or slight changes in data that might degrade model performance. More recent work by Lyzhov et al. showed that learnable augmentation policies that vary based on the input or model state can outperform static transformations that produce the same fixed augmentations regardless of the input [7]. In this manner, more dynamic and flexible model responses are achievable, which has been shown to have potential for robustness improvement.

In the DRL paradigm, most data augmentation techniques have been explored at training time as a regularization technique, aiming to improve generalization and prevent overfitting [5, 6, 8]. However, data augmentation at test time, i.e., applying Test Time Augmentation, has been less explored. Wang et al. emphasized the central importance of robustness for DRL policies, especially in environments where noise or dynamic change is involved [3]. Raileanu et al. also demonstrated that increasing the diversity of augmentations during training can have a profound impact on the ability of an agent to generalize across environments [6].

3 Methodology

The suggested TTA framework for DRL comprises three synergistic elements: physics-aware state perturbation, ensemble-based decision consolidation, and performance-adaptive augmentation scaling. The technical realization and theoretical underpinning of these are discussed in this section.

3.1 Manuscript Formatting Requirements

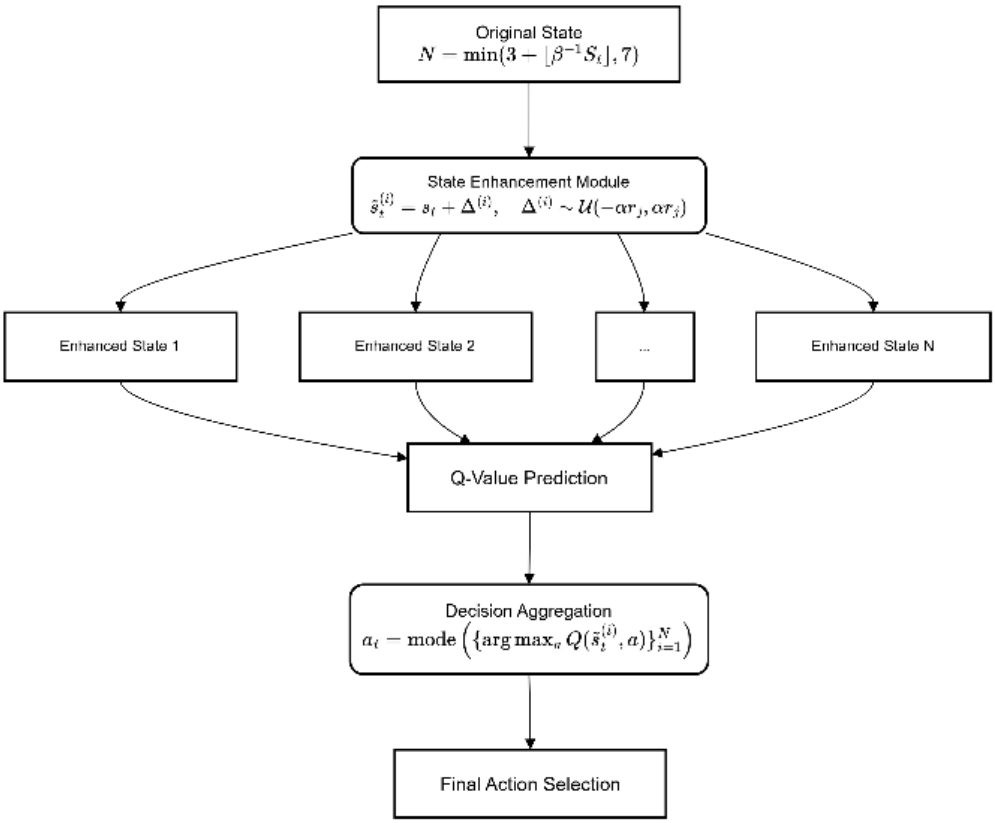

Fig. 1 Workflow for Test Time Augmentation. (Picture credit: Original)

The workflow of the process is outlined in Figure 1. Initially, the agent receives a raw state, denoted as (S_t), from the environment. This state includes all the relevant information needed for the agent to make an informed decision about its next action, such as positions, velocities, and other environmental factors. Based on feedback from the agent's performance, the number of augmentations N, is dynamically adjusted. This adjustment tailors the level of augmentation applied to the state according to the agent's ability to handle noise or variability, allowing for more targeted regularization.

With the updated value of (N), a set of physically valid perturbed states is generated, introducing noise or variations that align with the physical dynamics of the environment. These perturbed states provide useful diversity while ensuring they remain valid. A parallel set of forward passes through the DQN is then performed for each augmented state,

estimating the Q-values for each state-action pair. This enables the agent to evaluate the potential outcomes of each possible action based on the perturbed states. Finally, the agent selects the most frequent optimal action by considering the predicted Q-values across the augmented states. This voting mechanism aggregates information from the multiple forward passes and selects the action with the highest frequency of being identified as optimal across the perturbed states.

3.2 Physics-Constrained State Perturbation

Traditional vision-based TTA employs pixel-level transformations (e.g., rotations, flips) that lack semantic meaning for RL states. This paper's method introduces physical range-constrained noise to preserve environmental semantics. For each state dimension s_j, this method calculates its maximum historical variation range, $r_j = \max_t(s_j) - \min_t(s_j)$ during training. The augmentation noise follows formula (1).

$$\Delta_j^{(i)} \sim \mathcal{U}(-\alpha r_j, \alpha r_j) \tag{1}$$

Where $\alpha \in [0,0.5]$ controls perturbation intensity, ensuring a reasonable range of s described in formula (2):

$$s_j^{min} \leq \tilde{s}_j^{(i)} \leq s_j^{max} \tag{2}$$

This prevents generating physically implausible states (e.g., lunar lander below ground level). Compared to Gaussian noise used in [9], uniform distribution provides explicit bound guarantees while maintaining diversity [10].

3.3 Consensus-Driven Action Selection

Traditional reinforcement learning (RL) agents rely on greedy Q-maximization to select actions, but this approach becomes brittle under perturbed observations. To enhance robustness, the proposed Majority Voting with Graceful Degradation mechanism first generates N augmented states $\{\tilde{s}_t^{(i)}\}_{i=1}^N$ and computes the candidate actions as formula (3).

$$a^{(i)} = \arg\max_a Q(\tilde{s}_t(I), a) \tag{3}$$

If a mode exists among the candidate actions, the agent selects it as the final action: $a_t = \text{mode}(\{a^{(i)}\})$. However, in the case of a tie, the agent instead chooses the action that maximizes the sum of Q-values across all augmented states shown in formula (4).

$$a_t = \arg\max_a \sum_{i=1}^{N} Q(\tilde{s}_t^{(I)}, a) \tag{4}$$

This two-stage approach effectively combines action frequency statistics with Q-value magnitudes, ensuring a more stable decision-making process and addressing edge cases where multiple actions receive similar vote counts.

3.4 Performance-Adaptive Augmentation Scaling

Fixed N implementations ([2] in CV) prove suboptimal for RL's sequential decision-making. This paper introduce Dynamic Augmentation Budgeting:

$$N_t = \min(3 + \lfloor \beta^{-1} S_t \rfloor, 7) \quad (5)$$

Where S_t represents the current trajectory's cumulative reward, and $\beta = 50$ controls scaling sensitivity. This allows Baseline operation (N=3) when $S_t \approx S_{baseline}$ then gradually increase as performance improves until reaching the upper bound preventing the computational explosion.

4 Experimental Results

The LunarLander-v2 environment [4] with controllable noise and observation variability is a standard testbed well-suited to the assessment of DRL policies under changing conditions. It is therefore an excellent platform for investigating the potential implications of TTA, with a clearly defined benchmark for how inference-time augmentations can enhance policy robustness to noise and uncertainty.

4.1 Clean Environment Performance

In the clean environment, TTA enhanced DQN's final average score by 4.78%. This is also reflected in the distribution (Figure 2), where TTA consistently achieved more stable results with fewer erratic actions. As shown in Figure 3, the score comparison between Baseline and TTA also illustrates the efficacy of TTA in reaching higher and more stable scores.

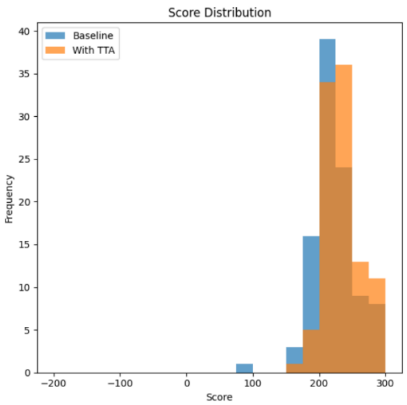

Figure 2: Score distribution of Baseline and TTA in clean environment. (Picture credit: Original)

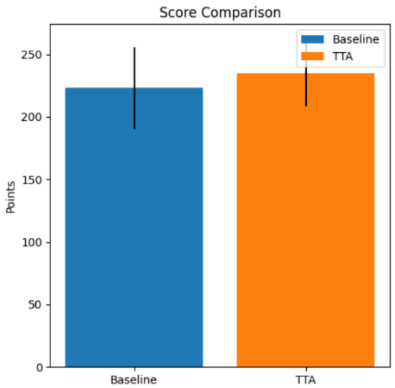

Figure 3: Score comparison of Baseline and TTA in clean. (Picture credit: Original)

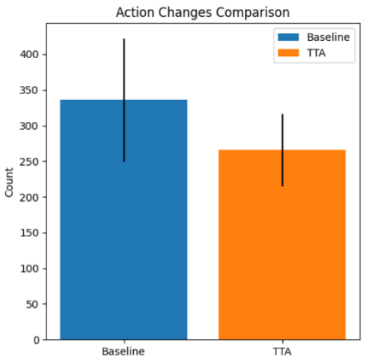

Figure 4: Action changes comparison of Baseline and TTA in clean environment. (Picture credit: Original)

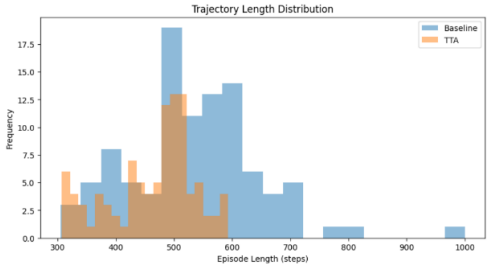

Figure 5: Trajectory length distribution of Baseline and TTA in clean environment. (Picture credit: Original)

Additionally, Figure 4 illustrates the reduced action changes with TTA, revealing a more stable decision-making process, and Figure 5 illustrates the trajectory length distribution, demonstrating that TTA led to shorter and more effective trajectories. These findings in total demonstrate that TTA led to increased stability and efficiency in the agent's decision-making process.

4.2 Noise Robustness Analysis

Next, this research tested the performance of DQN in clean, weak noise, and strong noise environments. As shown in Table 1, this paper compared the average score, success rate, and failure rate between the baseline and TTA under each environment. The results indicate that TTA consistently outperformed the baseline in noisy environments with **85%** success rate and **227.4** avg score. Also, in the clean environment, TTA leads to a increasement in both success rate and average score.

Table 1: A comparison of Avg Score, Success Rate and Fail Rate with Baseline and TTA applied to clean, weak noise and strong noise environment.

Test Case	Avg Score	Success Rate	Fail Rate
Clean Baseline	230.9 ± 33.2	79.0%	0.0%
Clean TTA	233.6 ± 28.0	88.0%	0.0%
Noise Baseline	220.6 ± 55.1	82.0%	0.0%
Noise TTA	227.4 ± 50.1	85.0%	1.0%
High Noise Baseline	170.2 ± 58.4	34.0%	1.0%
High Noise TTA	153.6 ± 79.6	26.0%	4.0%

Table 2: A comparison of Avg Score, Success Rate and Fail Rate after using TTA in clean, weak noise and strong noise environment

TTA effect	Clean	Weak Noise	Strong Noise
Success Rate	9%↑	3%↑	8%↓
Avg Score	2.7↑	6.8↑	16.6↓
Error Range	5.2↓	5.0↓	21.2↑

Further analysis in Table 2 provides a more detailed comparison of the performance metrics after applying TTA. The table shows that under clean conditions, TTA leads to a 9% increase in success rate, a 2.7-point improvement in the average score, and a 5.2-point reduction in error range, indicating overall performance enhancement. In weak noise conditions, TTA still provides benefits, with a 3% increase in success rate and a 6.8-point rise in the average score, while also slightly reducing the error range by 5.0 points, suggesting improved stability. However, in strong noise conditions, the effectiveness of TTA declines, as shown by an 8% drop in success rate, a 16.6-point decrease in the average score, and a 21.2-point increase in error range, highlighting the challenge of adapting to severe perturbations. These findings reinforce that while TTA enhances stability and decision-making in low to moderate noise environments, its advantages diminish as noise levels increase.

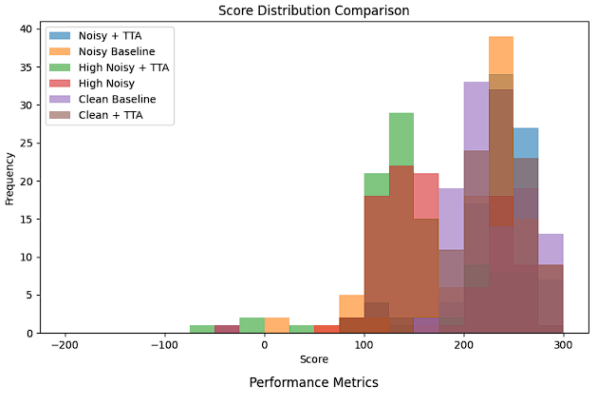

Figure 6: Score distribution comparison of Baseline and TTA applied to clean, weak noise and strong noise environments. (Picture credit: Original)

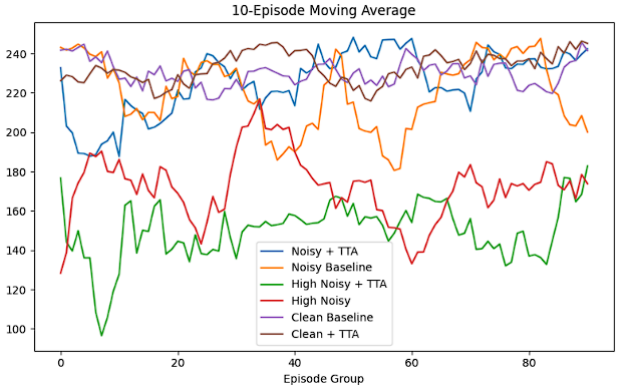

Figure 7: 10-Episode Average comparison of Baseline and TTA applied to clean, weak noise and strong noise environment. (Picture credit: Original)

In Figure 6, the score distribution comparison between Baseline and TTA across clean, weak noise, and strong noise environments shows that TTA generally resulted in a more consistent and stable performance, particularly in noisy conditions. Similarly, Figure 7 presents the 10-episode average scores, which reveal that TTA, even under noise, consistently outperformed the baseline in terms of overall performance.

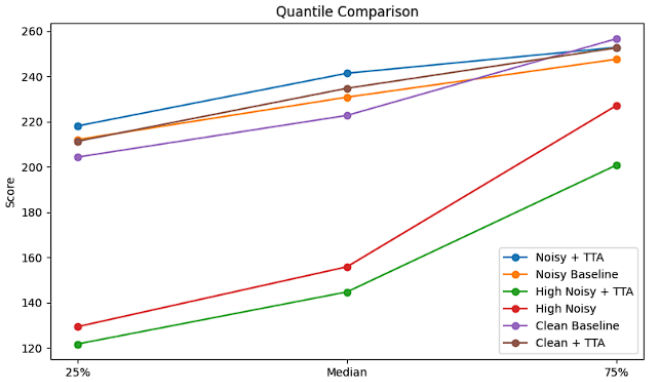

Figure 8: Quantile comparison of Baseline and TTA applied to clean, weak noise and strong noise environments. (Picture credit: Original)

However, as is apparent from Figure 8, which presents the quantile comparison, in very noisy conditions, TTA did not lead to performance improvement. On the contrary, it led to a drop in scores, which suggests that the benefits of TTA decrease as noise levels increase. This observation is further supported by the data presented in Table 1, which shows that TTA's effectiveness is limited under highly noisy conditions.

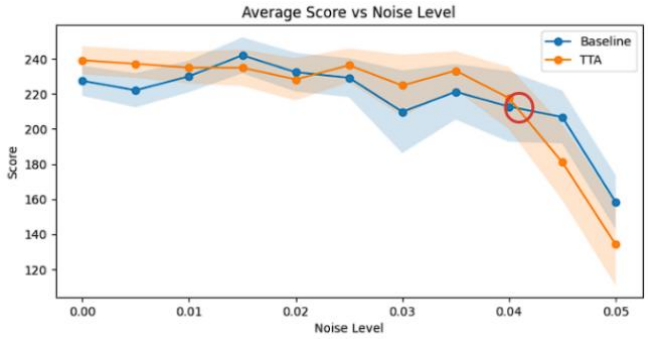

Figure 9: Average Score of Baseline and TTA in different noise levels. (Picture credit: Original)

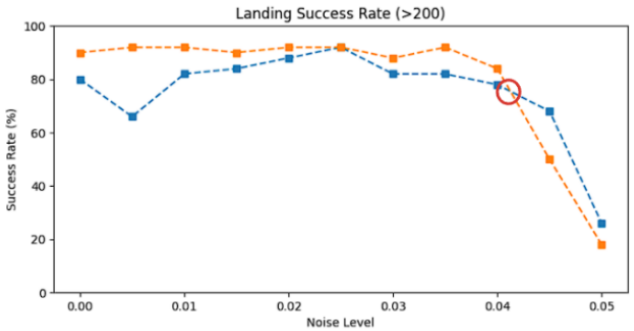

Figure 10: Success Rate comparison of Baseline and TTA in different noise levels. (Picture credit: Original)

Further testing with varying amounts of noise, as graphed in Figure 9 and Figure 10, showed that in mildly noisy conditions, TTA improved the success rate and provided a modest improvement in the average score for the DQN. Beyond a certain noise level (σ = 0.04), though, TTA caused a decrease in average score and success rate. This means that while TTA assists under low-noise conditions, its usefulness is limited under high-noise conditions, leading to poorer performance.

These results highlight the robustness of TTA in handling weak noise, but also point to the diminishing returns when faced with stronger disturbances. The findings suggest that while TTA is a useful technique for enhancing decision-making in noisy environments, its application may need to be adapted or optimized for extreme noise conditions.

4.3 Experimental Findings

When mild noise is introduced, specifically when $\alpha \leq 0.04$, TTA acts as a strong regularization technique. The results suggest that by incorporating controlled noise during inference, the agent's generalization ability across unseen states is significantly improved. This finding aligns with previous research, which has shown that noise can be a powerful tool for preventing overfitting and enhancing the agent's performance in more varied situations [9].

In addition, the dynamic TTA approach proves to be more effective than static ones. By intelligently adjusting the number of augmentations based on the agent's current performance, the dynamic strategy ensures that the augmentation level is always tuned to what the agent requires. This adaptability allows for maximum benefits from the noise perturbations without unnecessary computational overhead, ultimately leading to more

efficient computations and better overall performance, particularly in noisy environments where the noise severity can vary.

However, a performance degradation was observed when the noise level reached a certain threshold ($\sigma > 0.04$). At this point, TTA no longer led to performance improvements; instead, it caused a significant decline in the agent's ability to make optimal decisions. This suggests that there are inherent limits to the utility of input-space augmentation, beyond which excessive noise actually harms the agent's decision-making ability, rather than enhancing it.

5 Conclusion

The results demonstrate that Test Time Augmentation enhances the robustness of Deep Reinforcement Learning agents, particularly in environments with predictable dynamics. TTA improves the agent's generalization ability and performance, showing its potential to prevent overfitting and adapt to varied situations. However, while TTA proves effective in mild noise conditions, its benefits diminish in high-noise environments, where it may even lead to performance degradation.

A key limitation of TTA is its decreasing effectiveness under extreme noise levels. This calls for future research to explore ways to mitigate this weakness. One promising direction is the combination of TTA with latent-space augmentation techniques. By introducing perturbations at a more abstract level, this approach could offer a more robust form of regularization that avoids the detrimental effects of severe noise, potentially leading to more stable performance in noisy and unpredictable environments.

Looking ahead, there is an opportunity to further optimize the TTA framework by dynamically adjusting the level of augmentation based on real-time feedback from the environment. This adaptability could enhance the agent's performance while minimizing unnecessary computational overhead.

The significance of this research lies in its contribution to improving the robustness and efficiency of DRL agents, particularly in real-world applications where noise and unpredictable dynamics are prevalent. By understanding the interplay between TTA and noise, this paper pave the way for more resilient agents capable of handling complex environments, ultimately advancing the field of reinforcement learning in practical settings.

References

1. Mnih, V., Kavukcuoglu, K., Silver, D., et al.: 'Human-level control through deep reinforcement learning', Nature, 2015, 518, (7540), pp. 529–533
2. Wang, J., Perez, L.: 'The effectiveness of data augmentation in image classification using deep learning', Convolutional Neural Networks Vis. Recognit., 2017, 11, pp. 1–8
3. Wang, X., Wang, S., Liang, X., et al.: 'Deep reinforcement learning: A survey', IEEE Trans. Neural Netw. Learn. Syst., 2022, 35, (4), pp. 5064–5078
4. Awasthi, A.: 'Evaluating reinforcement learning algorithms for LunarLander-v2: A comparative analysis of DQN, DDQN, DDPG, and PPO', Research Square Preprint, 2025, Version 1
5. Laskin, M., Lee, K., Stooke, A., et al.: 'Reinforcement learning with augmented data', Adv. Neural Inf. Process. Syst., 2020, 33, pp. 19884–19895

6. Raileanu, R., Goldstein, M., Yarats, D., et al.: 'Automatic data augmentation for generalization in deep reinforcement learning', arXiv preprint, 2020, arXiv:2006.12862
7. Lyzhov, A., Molchanova, Y., Ashukha, A., et al.: 'Greedy policy search: A simple baseline for learnable test-time augmentation'. Proc. Conf. Uncertainty in Artificial Intelligence (UAI), August 2020, pp. 1308–1317
8. Kirk, R., Zhang, A., Grefenstette, E., et al.: 'A survey of generalisation in deep reinforcement learning', arXiv preprint, 2021, arXiv:2111.09794
9. Igl, M., Ciosek, K., Li, Y., et al.: 'Generalization in reinforcement learning with selective noise injection and information bottleneck', Adv. Neural Inf. Process. Syst., 2019, 32
10. Wang, J., Wang, Z., Li, H., et al.: 'Maximum entropy evolutionary reinforcement learning method based on adaptive noise', Acta Automatica Sinica, 2023, 49, (1), pp. 54–66