Influence of Rudder Action Resolution on Autonomous Ship Path Following

Md Shadab Alam* and Pavlo Bazilinskyy
Eindhoven University of Technology, The Netherlands
Emails: m.s.alam@tue.nl, p.bazilisnkyy@tue.nl
*Corresponding author

***Abstract*—Rudder action resolution directly influences how a deep reinforcement learning controller balances path tracking and steering demand. This study compares three autonomous ship controllers: a Deep Q Network (DQN) with three discrete rudder commands, a DQN with five discrete commands, and a Proximal Policy Optimization (PPO) controller with continuous rudder control. All controllers are evaluated on the same three degree of freedom KRISO Container Ship model using common observations, reward formulation, actuator limits, and performance metrics. In calm water, DQN 3 achieves the lowest mean cross track RMSE of 1.16L, while PPO gives the lowest mean rudder demand of 0.16. Under wind and wave disturbances, PPO achieves the lowest mean RMSE of 0.73L and the lowest mean rudder demand of 0.17. In model scale experiments, PPO records the smallest tracking error at 0.79L, compared to 0.83L for DQN 3 and 0.86L for DQN 5. In a tightening spiral, DQN 3 reaches 32 of 40 waypoints, compared to 28 for PPO and 27 for DQN 5.**

***Index Terms*—autonomous surface vessel, deep reinforcement learning, DQN, PPO, path following, rudder control, action space**

I. Introduction

Autonomous surface vessels require reliable steering control to maintain accurate navigation under nonlinear manoeuvring dynamics, changing path curvature, environmental disturbances, and uncertainties between simulated and physical operation. Deep reinforcement learning (DRL) has consequently attracted increasing attention in autonomous marine navigation, including heading control, path following, collision avoidance, and decision making [1]–[7]. Unlike conventional controllers that rely strongly on predefined models and control laws, DRL can learn control behaviour through interaction with the operating environment. This flexibility makes it attractive for marine systems in which nonlinear dynamics and changing operating conditions can complicate controller design.

An important but comparatively less studied aspect of DRL control is the representation of the action space. In rudder based ship control, the action space determines the steering commands that a learned policy can generate. Coarse discrete actions provide a limited number of strong steering choices, whereas finer discrete actions permit intermediate commands and continuous action spaces remove command quantisation within the allowable range. Action resolution can therefore influence how a controller responds to path deviations, maintains sustained turning, and regulates steering activity. These characteristics are particularly relevant for autonomous vessels because tracking accuracy and control demand need not improve simultaneously.

Previous studies have demonstrated the potential of both discrete and continuous DRL approaches for marine control. Sivaraj et al. investigated DQN based heading control and path following in calm water and waves [2], while Deraj et al. examined DQN based KCS path following in simulation and model scale experiments [1]. Woo et al. demonstrated DRL based path following through free running vessel experiments [5], and Wang et al. investigated autonomous ship path following across different reference trajectories [6]. Continuous policy approaches have also been explored, including PPO for path following and obstacle avoidance [4], while broader comparisons of DRL algorithms have been reported for realistic maritime navigation scenarios [3]. More recent research has extended DRL based marine control towards robustness, physical validation, and operation under environmental disturbances [8]. DRL has also been increasingly investigated for autonomous collision avoidance and integrated navigation problems [7].

Despite this progress, the action space itself has received considerably less attention than algorithm selection, reward design, state representation, and network architecture. In particular, it remains unclear how changing rudder action resolution affects the balance between tracking accuracy and steering demand, whether the resulting behaviour remains consistent under environmental disturbances, and whether performance trends observed in simulation persist during physical operation. Recent reviews of autonomous ship navigation similarly identify robustness, generalisation, real world validation, and consistent evaluation as important challenges for learning based maritime systems [9]. Understanding the role of action resolution is therefore important not only for DRL formulation but also for selecting control behaviour appropriate to the operational requirements of an autonomous vessel.

A. Aim of the Study

The aim of this study is to investigate how rudder action resolution influences the performance and behaviour of DRL based autonomous ship controllers. The study focuses on the relationship between action resolution, path tracking accuracy, steering demand, disturbance response, sustained turning, and transfer from simulation to physical operation. Particular

attention is given to the tradeoffs associated with coarse discrete, finer discrete, and continuous rudder representations rather than seeking a single controller that is optimal under every criterion. The objective is to clarify when different levels of action resolution are advantageous and thereby provide a more informed basis for action space selection in autonomous ship control.

II. Ship Model and Control Formulation

A. KCS maneuvering model

The vessel is represented by the three degree of freedom Maneuvering Modeling Group formulation for surge, sway, and yaw [10]. The equations of motion are

$$(m + m_x)\dot{u} - mvr - mx_G r^2 = X_H + X_R + X_P + X_E, \tag{1}$$

$$(m + m_y)\dot{v} + mx_G\dot{r} + mur = Y_H + Y_R + Y_E, \tag{2}$$

$$(I_{zz} + J_{zz})\dot{r} + mx_G\dot{v} + mx_G ur = N_H + N_R + N_E, \tag{3}$$

where m is the mass of the vessel, I_{zz} is the yaw moment of inertia and m_x, m_y, and J_{zz} denote the added terms of inertia. The subscripts H, R, P, and E denote hull, rudder, propeller, and environmental contributions, respectively. In calm water, the environmental terms are omitted. Wind forces and moments are computed from the relative wind velocity in the fixed frame of the body, while second order mean wave drift forces are obtained from hydrodynamic coefficients on wave frequency, speed of the vessel, and relative wave direction [11].

The policy outputs the commanded rudder angle δ_c. The realised rudder angle δ follows first order actuator dynamics,

$$T_R\dot{\delta} + \delta = \delta_c, \tag{4}$$

with $T_R = 0.1$ and a maximum rudder rate of $5°$/s at full scale. Thus, discrete policies still produce a continuous physical rudder response, although their target angles belong to a finite set.

This actuator model is central to interpreting the action spaces. A three action controller obtains intermediate realised rudder angles only transiently while moving between discrete targets. Moderate steering can therefore be approximated through command switching, potentially increasing command activity and rudder oscillation. The five action controller reduces this quantisation by providing $\pm 20°$ targets, while a continuous policy removes target angle quantisation altogether. The same rate limit and first order lag constrain all controllers. The comparison therefore concerns the commanded rudder resolution rather than different physical actuator limits.

B. Environmental disturbances

The wind is represented in the body fixed frame using the heading of the vessel ψ, the speed of the wind V_w and the global direction of the wind β_w,

$$u_w = V_w \cos(\beta_w - \psi), \tag{5}$$

$$v_w = V_w \sin(\beta_w - \psi). \tag{6}$$

The corresponding surge force, sway force, and yaw moment are

$$W_x = C_{wx}(\gamma_w)\frac{\rho_a}{\rho}A_x U_{wr}^2, \tag{7}$$

$$W_y = C_{wy}(\gamma_w)\frac{\rho_a}{\rho}A_y U_{wr}^2, \tag{8}$$

$$W_\psi = C_{w\psi}(\gamma_w)\frac{\rho_a}{\rho}A_y L_{OA} U_{wr}^2, \tag{9}$$

where γ_w is the relative wind direction, U_{wr} is the relative speed of the wind, A_x and A_y are projected areas, and ρ_a/ρ is the ratio of air density to water. The coefficients C_{wx}, C_{wy}, and $C_{w\psi}$ vary with relative direction of the wind. Wave effects are represented by second order mean drift forces obtained using a panel method over wave frequency, vessel speed, and relative wave direction and interpolated in the current vessel state [11]. The policies do not receive direct wind or wave measurements, so disturbance rejection occurs through the navigation states in the observation vector.

C. State, reward, and action spaces

The observation vector is

$$\mathbf{s} = [d_c, \chi_e, d_{wp}, r], \tag{10}$$

where d_c is the cross track error, χ_e is the course angle error, d_{wp} is the distance to the next waypoint, and r is the yaw rate. The reward follows the path following formulation of Deraj et al. [1]:

$$R = 2e^{-0.08d_c^2} - 1 + 1.3e^{-10|\chi_e|} - 0.3 - 0.25d_{wp}. \tag{11}$$

A terminal reward of 100 is assigned when the destination is reached. Episodes terminate when the destination is reached, the vessel passes the destination while moving away, or the maximum episode length is reached.

The controllers differ in policy class and rudder command space. The two DQN controllers use discrete actions, while PPO uses a continuous command:

$$\mathcal{A}_{\mathrm{DQN3}} = \{-35°, 0°, 35°\}, \tag{12}$$

$$\mathcal{A}_{\mathrm{DQN5}} = \{-35°, -20°, 0°, 20°, 35°\}, \tag{13}$$

$$\mathcal{A}_{\mathrm{PPO}} = [-35°, 35°]. \tag{14}$$

DQN approximates the state action value function over discrete rudder commands [12], whereas PPO learns a continuous stochastic policy using clipped policy updates [13]. This distinction enables command resolution to be examined while keeping the dynamics, observations, reward terms, and rudder limits of the vessel unchanged.

D. Training configuration

All three policies were trained in calm water. The DQN controllers use two hidden layers with 64 units and $\tanh$ activations, experience replay with a capacity of 10^5 transitions, and a batch size of 128. DQN 3 uses a discount factor of 0.95 and an exploration schedule decaying to $\epsilon = 0.2$, whereas DQN 5 uses 0.96 and decays exploration to zero. PPO uses separate policy and value networks with two 128 unit $\tanh$ layers,

TABLE I
MAIN TRAINING SETTINGS OF THE EVALUATED POLICIES.

Setting	DQN 3	DQN 5	PPO
Hidden layers	$64, 64$	$64, 64$	$128, 128$
Activation	$\tanh$	$\tanh$	$\tanh$
Discount γ	0.95	0.96	0.96
Initial learning rate	10^{-3}	10^{-3}	10^{-3}
Training episodes	10001	10001	5000
Evaluation checkpoint	ep. 7000	ep. 10000	iter. 60

generalised advantage estimation with $\lambda = 0.95$, a clipping ratio of 0.2, and an entropy regularisation coefficient of 0.01. Ten optimisation epochs are applied to each PPO batch.

The evaluated checkpoints correspond to episode 7000 for DQN 3, episode 10000 for DQN 5, and iteration 60 for PPO. These checkpoints are used throughout the simulation and physical comparisons. Table I summarises the main settings. Action resolution is not the only difference because each algorithm retains its own optimisation configuration. The results therefore compare three trained controllers under a common plant and evaluation protocol rather than a controlled ablation with identical optimiser hyperparameters.

E. Evaluation metrics and protocol

Tracking accuracy is quantified by the root mean square cross track error

$$d_{\mathrm{RMSE}} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} d_c^2(n)}, \tag{15}$$

where N is the number of recorded control steps. Steering demand is reported as the mean absolute realised rudder angle

$$E_\delta = \frac{1}{N}\sum_{n=1}^{N} |\delta(n)|. \tag{16}$$

The quantities are reported separately because d_{RMSE} measures geometric path following accuracy, whereas E_δ captures average actuator deflection. Combining them into one score would impose an arbitrary weighting on quantities with different physical meanings. Moreover, E_δ is a steering demand indicator rather than a direct measure of actuator energy or wear because it does not capture how rapidly the rudder moves.

All numerical comparisons use a common evaluation procedure for trajectory generation, disturbance application, and metric calculation. The policies retain their respective training configurations, while the plant model, actuator limits, observations, reward definition, and evaluation metrics remain common. Because one checkpoint is evaluated per controller, the results are interpreted descriptively rather than as repeated seed estimates of algorithm performance.

For physical validation, a 1:75.5 scale KCS model was used. Propulsion was provided by a brushless DC motor and rudder actuation by a stepper motor. Position and motion were measured using an Ardusimple RTK2B SBC GPS system and an SBG Ellipse A IMU and fused using a Kalman filter. Propeller speed remained constant while the learned policy commanded the rudder. The available test area constrained the comparison to an approximately 40 m by 40 m square path, with a waypoint acceptance region used to switch between successive segments. This arrangement evaluates whether the policy retains its simulated behaviour when exposed to unmodelled dynamics, sensing uncertainty, and mild environmental disturbances.

TABLE II
CALM WATER RESULTS FOR THE THREE CONTROLLERS. LOWER VALUES ARE BETTER.

	DQN 3		DQN 5		PPO	
Path	RMSE	E_δ	RMSE	E_δ	RMSE	E_δ
$(10, 10)$	**0.75**	0.19	0.91	0.16	0.84	**0.12**
$(-10, 10)$	**2.22**	0.23	2.48	0.22	2.44	**0.19**
$(-10, -10)$	**2.06**	0.40	2.29	0.27	2.34	**0.18**
$(10, -10)$	**0.73**	0.31	0.90	0.21	0.82	**0.12**
Cardioid	0.35	0.27	0.47	0.22	**0.31**	**0.12**
Astroid	1.23	0.33	**1.20**	0.28	1.21	**0.21**
Sine	**0.89**	0.26	0.97	0.25	0.91	**0.13**
Star	**1.14**	0.30	1.44	0.26	1.55	**0.14**
Circle	1.08	0.37	**0.61**	0.34	0.82	**0.27**
Mean	**1.16**	0.30	1.25	0.25	1.25	**0.16**

TABLE III
WIND AND WAVE CONDITIONS USED IN THE DISTURBANCE EVALUATION.

Case	Path	V_w/U	β_w	W_h	T_w	ψ_w
I	Eight	6	$\pi/2$	3	10	$\pi/4$
II	Eight	6	$-\pi/2$	3	10	$-\pi/2$
III	Cardioid	3	0	2	15	0
IV	Cardioid	5	$\pi/4$	2	15	0
V	Sine	6	π	3	8	π
VI	Sine	3	$\pi/2$	3	8	$\pi/2$
VII	Astroid	4	$3\pi/4$	2	10	$3\pi/4$
VIII	Astroid	4	$\pi/4$	3	20	$-\pi/4$

III. RESULTS

A. Calm water path following

The calm water comparison contains nine manoeuvres: four single destination paths in different quadrants, cardioid, astroid, sine, star, and circle. Table II reports the tracking error and mean rudder demand for each controller. Across the nine manoeuvres, DQN 3 has the lowest mean cross track RMSE at $1.16L$. DQN 5 and PPO have nearly identical means of $1.25L$. DQN 3 also gives the smallest RMSE on six of the nine individual manoeuvres. A coarse action space therefore does not necessarily reduce nominal path tracking accuracy.

The control demand shows the opposite pattern. PPO gives the lowest E_δ on every calm water manoeuvre. Its mean value is 0.16, compared with 0.30 for DQN 3 and 0.25 for DQN 5. Relative to DQN 3, PPO reduces the mean absolute rudder angle by 44.41%; relative to DQN 5, the reduction is 32.99%. Fig. 1 summarises the corresponding tracking and rudder demand across the nine manoeuvres. The result indicates that the continuous policy achieves comparable tracking with substantially smaller average rudder deflection.

The circle provides the clearest example of why action resolution matters. DQN 3 has an RMSE of $1.08L$, whereas DQN 5 reduces the error to $0.61L$ and PPO gives $0.82L$. A

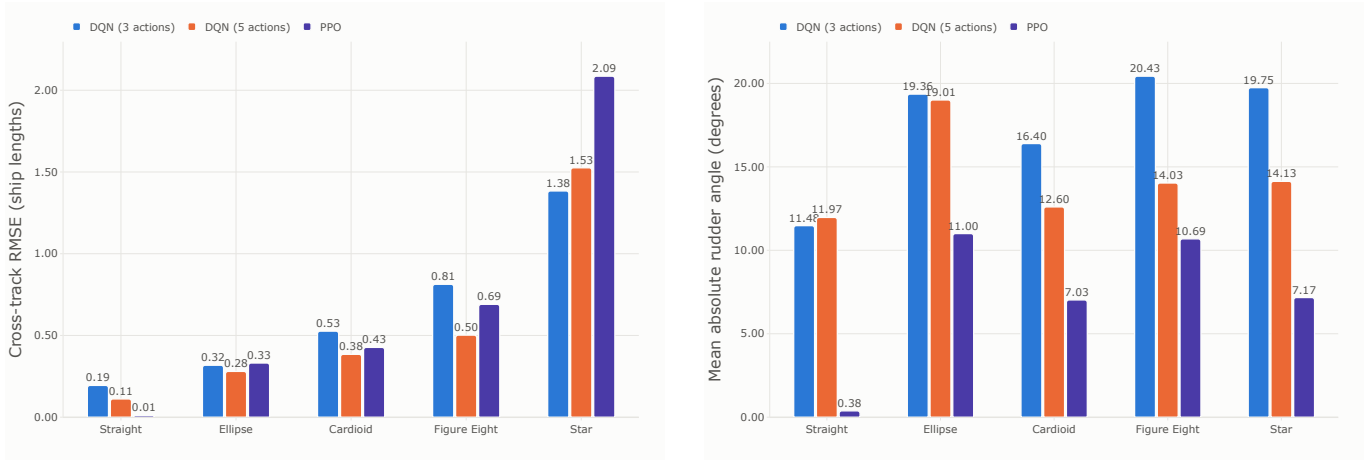

Fig. 1. Calm water comparison over the nine evaluated manoeuvres. Left: cross track RMSE. Right: mean absolute realised rudder angle. Lower values indicate better performance.

circle requires a sustained nonzero steering demand. DQN 3 can only alternate between midships and hard rudder to approximate such a demand. DQN 5 can hold $\pm 20°$, while PPO can select any intermediate value. The improvement therefore follows directly from the additional steering resolution rather than from a larger rudder limit.

B. Response to wind and wave disturbances

The disturbance study contains eight cases with different path geometries, wind directions and speeds, wave heights, wave periods, and wave directions. The cases are listed in Table III. They include figure eight, cardioid, sine, and astroid paths and deliberately vary both the magnitude and direction of environmental loading. The policies were trained in calm water and do not observe wind or wave variables. Their response to disturbances is therefore feedback driven through changes in d_c, χ_e, d_{wp}, and r, rather than through explicit feedforward compensation.

Table IV shows that the ranking in tracking accuracy becomes less consistent under disturbances. PPO gives the smallest RMSE in four cases, DQN 5 in three cases, and DQN 3 in one case. Averaged across the eight conditions, PPO gives the lowest RMSE at $0.73L$, followed by DQN 3 at $0.76L$ and DQN 5 at $0.79L$. The differences in mean tracking accuracy are modest compared with the difference in control demand.

PPO again gives the smallest E_δ in every disturbance case. Its mean value of 0.17 is 42.04% lower than DQN 3 and 29.38% lower than DQN 5. Thus, the most consistent advantage of the continuous action policy is not that it always gives the smallest path error, but that it achieves similar or better disturbance rejection with less average rudder deflection.

C. Transfer from simulation to model scale experiments

The simulation and experimental RMSE values for the model scale square path are reported in Table V, while Fig. 2 shows the measured trajectories of the three controllers during the model scale experiments. DQN 5 gives the smallest simulation error, but PPO gives the smallest experimental error. The experimental

TABLE IV
RESULTS IN THE EIGHT WIND AND WAVE DISTURBANCE CASES. LOWER VALUES ARE BETTER.

	DQN 3		DQN 5		PPO	
Case	RMSE	E_δ	RMSE	E_δ	RMSE	E_δ
I	0.72	0.36	**0.64**	0.26	0.65	**0.22**
II	0.70	0.37	0.74	0.26	**0.69**	**0.22**
III	0.38	0.24	0.47	0.22	**0.31**	**0.13**
IV	0.35	0.25	0.40	0.23	**0.26**	**0.14**
V	**0.59**	0.26	0.80	0.20	0.66	**0.14**
VI	0.94	0.26	0.94	0.24	**0.90**	**0.13**
VII	1.20	0.34	**1.13**	0.28	1.19	**0.21**
VIII	1.22	0.32	**1.20**	0.28	1.23	**0.22**
Mean	0.76	0.30	0.79	0.25	**0.73**	**0.17**

TABLE V
SQUARE PATH CROSS TRACK RMSE IN SIMULATION AND EXPERIMENT.

Controller	Simulation	Experiment	Increase
DQN 3	0.64	0.83	29.68%
DQN 5	**0.62**	0.86	38.06%
PPO	0.64	**0.79**	**22.63%**

RMSE increases relative to simulation for all three controllers: 29.68% for DQN 3, 38.06% for DQN 5, and 22.63% for PPO. The result shows that the nominal simulation ranking is not preserved after transfer to the physical vessel.

The course angle response in simulation follows the experimental switching pattern closely, whereas the cross track response differs more strongly after waypoint changes. Two effects are plausible. First, the MMG model and the physical vessel do not have identical transient dynamics. Second, the physical experiments contain environmental disturbances and sensing effects that are not reproduced exactly in the nominal simulation. The lower relative degradation of PPO suggests that smoother continuous steering may be less sensitive to these discrepancies, but repeated physical trials are required before treating this as a general algorithmic advantage.

Fig. 2. Model scale square path experiments for DQN 3, DQN 5, and PPO from left to right. White markers denote the prescribed waypoints and the red curve denotes the measured vessel trajectory.

D. Tightening spiral stress test

A tightening spiral provides a different stress condition because the required curvature increases continuously. The path contains 40 waypoints with an acceptance radius of $0.5L$. DQN 3 reaches 32 waypoints, PPO reaches 28, and DQN 5 reaches 27. This ordering differs from the control effort results and illustrates why the controllers should not be reduced to a single overall rank.

The three action DQN makes comparatively aggressive corrections and therefore retains strong waypoint capture as the spiral tightens, but this behaviour is accompanied by larger rudder demand. PPO follows with smaller average actuator deflection but reaches fewer waypoints in this specific stress case. The result exposes a genuine design tradeoff: applications that prioritise waypoint capture under rapidly increasing curvature may prefer a different controller characteristic from applications that prioritise smooth actuation and reduced steering activity.

IV. Discussion

The results show that action resolution affects closed loop behaviour in several distinct ways. Coarse actions can favour rapid path recovery: DQN 3 gives the smallest mean calm water RMSE and the highest spiral waypoint count. Its large available commands permit decisive correction when cross track error develops, but this behaviour is accompanied by the largest average rudder deflection. Additional action resolution is more beneficial when the path requires sustained curvature. On the circle, DQN 5 reduces RMSE by approximately 43% relative to DQN 3, which is consistent with the availability of the $\pm 20^{\circ}$ actions for maintaining a moderate turn without repeatedly switching between midships and hard rudder.

Continuous control provides the most consistent reduction in rudder demand. PPO has the smallest mean absolute rudder angle in every calm water manoeuvre and every disturbance case. Importantly, this advantage does not imply that continuous control must minimise geometric error on every path. The learning objective, policy parameterisation, and training trajectory distribution still determine how the available action authority is used. Action resolution should therefore be interpreted as one element of controller design rather than as a monotonic measure of controller quality.

The benefit of additional action resolution depends strongly on path geometry. DQN 3 gives the lowest RMSE for all four single destination manoeuvres, whereas the ranking changes on curved paths: PPO is best on the cardioid, DQN 5 on the astroid and circle, and DQN 3 on the sine and star paths. More available rudder commands therefore do not produce a monotonic improvement in tracking accuracy. Short corrective manoeuvres can benefit from decisive large commands, whereas sustained curvature benefits from a directly available moderate rudder command.

The disturbance cases reinforce this path dependent behaviour. The lowest RMSE is split across all three controllers, with PPO best in four cases, DQN 5 in three, and DQN 3 in one. In contrast, PPO gives the lowest E_{δ} in all eight cases. Because wind and wave quantities are absent from the observation vector, none of the policies performs explicit disturbance compensation. Their response arises from the effect of the disturbance on cross track error, course error, waypoint distance, and yaw rate. The results therefore describe feedback robustness of the learned steering policies rather than adaptation to measured environmental inputs.

The transfer results also suggest that action representation may influence sensitivity to model mismatch. A policy that repeatedly requests extreme discrete commands can depend more strongly on actuator lag and feedback timing than a policy that can request a moderate angle directly. The present experiments do not establish this mechanism causally; confirming it would require logging command switching frequency, realised rudder rate, and sensor to actuator delay in both simulation and physical trials.

The comparison suggests that controller selection should depend on the dominant operational requirement. If rapid recovery from path error or waypoint capture under strongly increasing curvature is most important, the aggressive corrections produced by a coarse discrete policy can be advantageous. If sustained turning is common, intermediate or continuous commands provide a more direct representation of the required steering demand. If lower average rudder deflection is important, PPO provides the clearest benefit in the present results. These behaviours arise from the interaction between policy action resolution and the first order rudder dynamics: the physical rudder remains rate limited in all cases, but a coarse policy can realise moderate steering only while moving between a small number of target angles.

The results also argue against ranking controllers with a single aggregate score. Tracking RMSE, waypoint completion, rudder demand, disturbance response, and transfer degradation capture different aspects of closed loop performance. This is visible in the present study: DQN 3 gives the lowest mean nominal RMSE and the highest spiral waypoint count, DQN 5 performs best on the circle, and PPO gives the lowest mean rudder demand and the smallest relative degradation in the model scale square path experiment. A useful evaluation procedure should therefore first verify task completion and safety, and then compare tracking and actuation metrics separately according to the intended operating conditions.

The comparison of simulation with experiment adds a further qualification. DQN 5 gives the smallest simulated RMSE on the square path but the largest experimental RMSE, whereas PPO shows the smallest relative increase after transfer. Therefore, nominal simulation performance alone is insufficient for controller selection. Physical validation should be treated as a separate stage because model mismatch, sensor uncertainty, actuator imperfections, and unmodeled environmental effects can alter the relative ordering observed in simulation.

A. Limitations and future studies

The present study provides a comparative evaluation of rudder action resolution across calm water simulations, environmental disturbances, model scale experiments, and a tightening spiral stress test. The results show consistent differences in tracking behaviour and steering demand among the three controllers. Nevertheless, the conclusions should be interpreted within the scope of the evaluated policies and operating conditions. Each controller is represented by one trained checkpoint and retains its respective training configuration. Future work could therefore repeat training across multiple seeds and conduct repeated evaluations to quantify variability and provide confidence intervals.

Physical experiments complement the simulation analysis by examining how the controllers behave after transfer to a physical vessel. The current experiments focus on a square reference path. Extending the evaluation to additional geometries, particularly trajectories combining sustained curvature with abrupt heading changes, would provide a broader assessment of transfer performance. Repeated trials with calm water and disturbed trials on identical paths would also allow the effect of environmental loading to be quantified more directly.

The mean absolute rudder angle E_δ provides an interpretable measure of the average steering demand, but does not capture the temporal characteristics of the rudder motion. Controllers with similar values of E_δ may still differ in switching frequency, total rudder travel, or time spent near the actuator limits. Future evaluations could therefore include additional measures such as total realised rudder variation, command switching frequency, or the integral of $|\dot{\delta}|$. These measures would help distinguish sustained steering from frequent corrective actions.

A further extension would be to separate the influence of action resolution more explicitly from the learning algorithm. In the present study, discrete control is represented by DQN and continuous control by PPO, reflecting the evaluated controller designs rather than a fully controlled ablation. A dedicated study could use matched network capacity, training budget, observations, reward formulation, and evaluation conditions while varying only the action parameterisation. This would provide a clearer estimate of the contribution of rudder resolution itself.

The commanded and realised rudder responses could also be examined in greater detail. All controllers are subject to the same first order actuator dynamics and rudder rate limit, so differences in commanded action resolution do not translate directly into equivalent differences in physical rudder motion. Analysing switching behaviour, saturation, rudder rate, and transient response after waypoint changes would provide further insight into the mechanisms underlying the observed tracking and steering characteristics.

V. Supplementary material

A maintained version of the code is available at https://github.com/Shaadalam9/compare_rl.

References

[1] R. Deraj, R. S. S. Kumar, M. S. Alam, and A. Somayajula, "Deep reinforcement learning based controller for ship navigation," *Ocean Engineering*, vol. 273, p. 113937, 2023.

[2] S. Sivaraj, S. Rajendran, and L. P. Prasad, "Data driven control based on deep q-network algorithm for heading control and path following of a ship in calm water and waves," *Ocean Engineering*, vol. 259, p. 111802, 2022.

[3] T. N. Larsen, H. Ø. Teigen, T. Laache, D. Varagnolo, and A. Rasheed, "Comparing deep reinforcement learning algorithms' ability to safely navigate challenging waters," *Frontiers in Robotics and AI*, vol. 8, p. 738113, 2021.

[4] E. Meyer, H. Robinson, A. Rasheed, and O. San, "Taming an autonomous surface vehicle for path following and collision avoidance using deep reinforcement learning," *IEEE Access*, vol. 8, pp. 41 466–41 481, 2020.

[5] J. Woo, C. Yu, and N. Kim, "Deep reinforcement learning-based controller for path following of an unmanned surface vehicle," *Ocean Engineering*, vol. 183, pp. 155–166, 2019.

[6] S. Wang, X. Yan, F. Ma, P. Wu, and Y. Liu, "A novel path following approach for autonomous ships based on fast marching method and deep reinforcement learning," *Ocean Engineering*, vol. 257, p. 111495, 2022.

[7] D.-H. Chun, M.-I. Roh, H.-W. Lee, J. Ha, and D. Yu, "Deep reinforcement learning-based collision avoidance for an autonomous ship," *Ocean Engineering*, vol. 234, p. 109216, 2021.

[8] J. Liu, C. Xiao, H. Yuan, C. Li, and Q. Li, "Unmanned surface vessels path following control using improved transferring reinforcement learning: Simulation and experiment," *Ocean Engineering*, vol. 320, p. 120346, 2025.

[9] W. Wang, M. Li, B. Wang, J. Hu, and T. Zhang, "A review of ship path planning for autonomous navigation: From model-driven methods to deep reinforcement learning," *Journal of Marine Science and Engineering*, vol. 14, no. 16, p. 1477, 2026.

[10] H. Yasukawa and Y. Yoshimura, "Introduction of MMG standard method for ship maneuvering predictions," *Journal of Marine Science and Technology*, vol. 20, no. 1, pp. 37–52, 2015.

[11] A. Guha and J. Falzarano, "Estimation of hydrodynamic forces and motion of ships with steady forward speed," *International Shipbuilding Progress*, vol. 62, no. 3–4, pp. 113–138, 2015.

[12] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.

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