Robust Formation Control of Wheeled Mobile Robots Using DDPG and TD3 Algorithms

Saraswathi N[1], Manjeet Rege[2], Shyam Krishan Joshi[3], Peplluis Esteva[4], Hemachandran K[5], Raul V. Rodriguez[6]

***Abstract*—The rigid formation control of Wheeled Mobile Robots (WMRs) in a cluttered environment is presented in this paper using a residual Reinforcement Learning (RL) framework. Unicycle robots form three shapes (a triangle, a square, and an X) and coordinate them simultaneously to different goals. A structured guidance controller is supplemented with a learned corrective policy, and the geometry of the rigid template is enforced with a Kabsch-based geometric projection, ensuring inter-agent and inter-formation collision avoidance is applied throughout. This framework is used to implement two state-of-the-art continuous-control algorithms (DDPG and TD3) and compare their performance through controlled multi-episode evaluation and component ablations under nominal and stochastic disturbance conditions. Results demonstrate that TD3 outperforms DDPG in terms of consistency and performance of the learned policies when the robot operates under nominal conditions, that the residual structure significantly improves the safety, formation accuracy and the training stability, and that the two algorithms show complementary behavior when they are disturbed. The novelty of the work is the first use of Kabsch-based rigid-template enforcement in a residual Deep RL architecture for simultaneous multi-formation navigation, where the usefulness of principled hybrid architectures that integrate classical control structure with modern RL for cooperative Multi-Robot Systems (MRS) is highlighted.**

***Index Terms*—Formation control, deep reinforcement learning, DDPG, TD3, wheeled mobile robots**

I. Introduction

Formation control of WMRs is a key issue in cooperative robotics, with applications in surveillance, autonomous transportation, environmental monitoring, exploration, and distributed sensing. Classical formation control strategies such as the scheme of leaders-follower, virtual structures, and consensus-based methods are widely studied [1], [2]. These methods have good theoretical assurances, however depend on accurate system models and tend to perform poorly when the dynamics are nonlinear, modeling uncertainties, and external perturbations.

Recent progress in DRL has offered an attractive alternative by enabling robots to learn control policies directly from the environment. Actor-critic models like DDPG [3] and TD3 [4] are likely to be especially useful in robotics because of their ability to handle high-dimensional continuous state and action spaces. They have already been applied to numerous types of tasks in robotic control, such as locomotion, manipulation, and coordination of multiple robots with each other [5]. However, learning-only controllers have a high probability of safety concerns during the training process, slow training, hyperparameter sensitivity, and slowness to adversarial examples or unmodeled dynamics.

A sub-area of hybrid control is referred to as residual RL; a standard model-based controller is first created, and then a learned correction policy is used to compensate for discrepancies between the original controller and the dynamic behavior observed in real time [6]. This also maintains the inherent stability of existing controller designs while simultaneously preserving the safety aspects of those designs [7]. As the safety-critical nature of MRS limits the use of pure exploration, this paper proposes a residual DRL approach for implementing rigid formation control using multiple WMRs.

The key findings of this paper can be outlined as follows:
(i) Novel safety-gated DRL framework:A residual DRL framework is proposed for multi-robot formation control, augmenting nominal guidance with learned corrective actions to enhance formation accuracy and convergence.
(ii) Kabsch-integrated DRL framework: This mechanism was developed to enable multi-robot simultaneous navigation to distinct goals while preserving inter-robot geometric configuration and ensuring collision avoidance throughout traversal.
(iii) Comparative evaluation of disturbances:This study benchmarks residual DDPG and TD3 control policies for multi-robot rigid formation navigation under both deterministic and stochastic disturbance conditions.
(iv) Multi-formation multi-goal navigation: The framework concurrently coordinates three distinct rigid formations pursuing separate goals within a shared obstacle-laden environment, a significantly more complex problem than those addressed in prior single-formation or single-goal studies.
The other sections are organised as follows. Section II explains the related work. The problem formulation is described in Section III. The methodology presented in section IV. In Section V the control architecture and algorithms is explored.

*This work was not supported by any organization

[1]Saraswathi Nayakanti is Research scholar in School of Technology, Woxsen University, Telangana, India saraswathinayakanti7@gmail.com

[2]Manjeet Rege is Professor and Chair of the Department of SE and DS; Director of the Center for Applied AI, University of St. Thomas, St. Paul, Minnesota 55105, USA rege@stthomas.edu

[3]Shyam Krishan Joshi is Assistant Professor, AIRC, Woxsen University, Telangana, India shyam.ee.iitd@gmail.com

[4]Peplluis Esteva De La Rosa is CU in Systems Engineering and Automatics, Campus de Montilivi, s/n 17071 GIRONA josepllu is.delarosa@udg.edu

[5]Hemachandran K is the Director of AIRC , Woxsen University, Telangana, India hemachandran.k@woxsen.edu.in

[6]Raul Villamarin Rodriguez is Vice President , Woxsen University, Telangana, India raul.rodriguez@woxsen.edu.in

Section VI analyses the results and discussion, and finally Section VII concludes the manuscript.

II. Related Work

This section provides a summary of recent developments most closely aligned with rigid formation control of non-holonomic WMRs, multi- robot coordination through DRL techniques, as well as mixed model-based and model-free control frameworks.

A. Classical Formation Control and Its Limitations

The work described in [8] demonstrated a global consensus-based formation control approach for nonholonomic WMRs with time-varying communication delays, without velocity feedback. [9] demonstrated formation control for WMRs disturbed by a combination of stochastic noise and non-random disturbance signals using distributed model predictive control (MPC) with an extended state observer. [10] demonstrated formation tracking laws for multi-robot systems with range constraints. [11] provided an embedded formation-controlling technique validated on the cooperative transport of multiple WMRs. While these approaches have provided theoretically robust solutions, they are subject to practical limitations for substantial mismatching between actual systems and their nominal models.

B. DRL for Multi-Robot Formation

The prominent learning-based methods for high-dimensional multi-robot cooperation have become continuous-control actor-critic algorithmic methodologies like DDPG [3] and TD3 [4]. [12] formulated multi-robot formation control in a constrained Markov game environment and proposed safe multi-agent policy-optimisation algorithms that constrain the probability of collision occurrence throughout training. [13] used DRL techniques for use with networked autonomous surface vehicles to achieve formation tracking. [14] applied RL methodology to optimal formation control of many robotic rollers. These studies typically analyse a single formation type; the ability to simultaneously coordinate many rigid formations towards separate objectives, in presence of stochastic process disturbances, remains largely unexamined in the literature.

C. Obstacle-Aware and Distributed Formation Learning

Recent Research examines this challenge using many different methods. [8] decomposed the problem into two parts: a centralised collision avoidance guidance module and a decentralised coordination layer for maintaining formation. [15] proposed an approach based on a bio-inspired neural-dynamic model to mitigate the speed jumps experienced by a robot during execution of learned policies. To further develop theory, [17] has performed an vision based path following following predictive control scheme. [16] developed a scalable multi-robot RL architecture. [11] addressed a related problem in cooperative transportation. In the past, formation rigidity and obstacle avoidance have been linked through reward shaping and potential fields without explicitly analysing how these things are affected by stochastic disturbances together.

D. Research Gap

The present work is different from the reviewed literature in four points.
(i) Residual structure: residual RL has been used mostly for single robot manipulation, but has been extended here to a multi-formation problem with twelve robots, where the [6], [7] trained residual is used to complement a guidance controller, not to stand alone.
(ii) Rigid projection: guidance-module approaches like enforce pairwise-distance constraints [8], the Kabsch-based projection used here provides a rotation- and translation-invariant error measure and drift is corrected at each step.
(iii) Disturbance handling: in addition to explicit observers, model-based schemes need a disturbance model, whereas the residual policy [9] is trained directly under stochastic process noise without a disturbance model.
(iv) Multi-formation objectives: previous DRL formation studies [12]– [14] consider only one formation to pursue one task, but there is no previous work on coordinating three different formations which pursue three different tasks in the same obstacle environment with stochastic disturbance.
It is these differences that drive the residual-DRL framework of the next section.

III. PROBLEM FORMULATION

Consider a team of N WMRs indexed by $i \in \{1, 2, ..., N\}$, operating in a two dimensional bounded planar workspace which has static obstacles. Each WMR has been modeled as a non-holonomic unicycle system, which is a standard kinematic representation of a WMR in formation control and cooperative navigation due to its ability to capture non-holonomic motion constraints while remaining computationally tractable. The following assumptions bound the scope of the proposed framework:
(i) Local observability during execution: Each robot observes only its own position, relative positions of robots within sensing radius r_s, and detected obstacle positions. No inter-robot communication is required at execution time.
(ii) No communication during execution: Coordination emerges implicitly through the shared reward structure and the per-formation Kabsch projection; no explicit message passing between robots is required.
(iii) Synchronous time steps: All robots update their states synchronously at each time step $\Delta t = 0.2s$.
(iv) Static environment: Obstacles are static and fully visible within the sensing radius, dynamic obstacles and adversarial robots are not considered.

1) Wheeled Mobile Robot Dynamics: The discrete-time kinematics of WMR of i are given by Equation (1) and Equation (2),

Without disturbances:

$$\begin{aligned} x_{i,k+1} &= x_{i,k} + v_{i,k} cos(\theta_{i,k})\Delta t \\ y_{i,k+1} &= y_{i,k} + v_{i,k} sin(\theta_{i,k})\Delta t \\ \theta_{i,k+1} &= \theta_{i,k} + \omega_{i,k}\Delta t \end{aligned} \tag{1}$$

With disturbance:

$$\begin{aligned} x_{i,k+1} &= x_{i,k} + v_{i,k} cos(\theta_{i,k})\Delta t + \omega^x_{i,k} \\ y_{i,k+1} &= y_{i,k} + v_{i,k} sin(\theta_{i,k})\Delta t + \omega^y_{i,k} \\ \theta_{i,k+1} &= \theta_{i,k} + \omega_{i,k}\Delta t + \omega^\theta_{i,k} \end{aligned} \tag{2}$$

Where $(x_{i,k}, y_{i,k}) \in R^2$ and $\theta_{i,k} \in (-\pi, \pi]$ denote the position and heading of WMR i at time step k, and $v_{i,k}, \omega_{i,k}$ are the linear and angular velocity control inputs, respectively. The control inputs are bounded such that $|v_{i,k}| \leq v_{max}$ and $|\omega_{i,k}| \leq \omega_{max}$ commonly considered for WMRs.

The terms $\omega^x_{i,k}, \omega^y_{i,k}$, and $\omega^\theta_{i,k}$ represent additive stochastic process disturbances, modeled as zero-mean Gaussian noise. This disturbance model captures the effects of actuator imperfections, wheel slippage, sensing noise, and unmodeled dynamics, and is routinely adopted in robust control and learning-based robotic systems.

2) Multi-Formation Structure and Group Motion: The robot team is partitioned into multiple rigid formation groups $\mathcal{G}$,

$$\mathcal{G} = \{\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3\},$$

corresponding to three rigid formations- a triangle, a square, and an X-formation. Each group $\mathcal{G}_m$ is assigned a distinct goal position represented with $p^{goal}_m \in R^2$, reflecting multi-task or cooperative mission scenarios encountered in surveillance, exploration, and distributed transportation.

To decouple global navigation from internal formation maintenance, the motion of each formation is characterized by its centroid $c_{m,k}$,

$$c_{m,k} = \frac{1}{|\mathcal{G}_m|} \sum_{i \in \mathcal{G}_m} \begin{bmatrix} x_{i,k} \\ y_{i,k} \end{bmatrix}$$

Centroid-based abstractions in formation control are enabled for scalable coordination while preserving distributed implementation.

3) Rigid Formation and Error Quantification: Each formation $\mathcal{G}_m$ is associated with a predefined rigid geometric template $\mathcal{T}_m$,

$$\mathcal{T}_m = r_{m,1}, r_{m,2}, ..., r_{m,|\mathcal{G}_m|}, \quad r_{m,j} \in R^2,$$

specified in a local reference frame. Formation rigidity is through distance preservation, and for inter-robot distances remain invariant as given by the equation (3)

$$||p_{i,k} - p_{j,k}|| = ||r_{m,i} - r_{m,j}||, \forall i, j \in \mathcal{G}_m \tag{3}$$

Such distance-based rigidity constraints from the foundation of rigid formation control theory [18]. Rather than enforcing all pairwise constraints explicitly, formation deviation is measured by aligning the template $\mathcal{T}_m$ with the current robot positions using an optimal rigid-body transformation:

$$e^{\text{form}}_m(k) = \frac{1}{|\mathcal{G}_m|} \sum_{i \in \mathcal{G}_m} \left\| \mathbf{p}_{i,k} - \left(\hat{\mathbf{R}}_m \mathbf{r}_{m,i} + \hat{\mathbf{t}}_m \right) \right\|^2 \tag{4}$$

Where R_m is planer rotation, $\hat{\mathbf{t}}_m$ is the translation from centroid, and desired template $\mathbf{r}_{m,i}$. This least-squares alignment problem admits a closed-form solution via the Kabsch algorithm, which computes the optimal rotation and translation minimizing the mean-squared error between two point sets. Further the concept can be explored by [19]. Each robot is then moved a fraction $\alpha = 0.8$ toward its target, which smooths motion while contracting the per robot deviation from the template by a factor of $(1 - \alpha)^2$ per step. Computed per formation from local positions and applied after the residual, the projection needs no communication and constrains the learner's output rather than being learned.

4) Safety and Environmental Constraints: Safe operation requires satisfaction of the following constraints:

Inter-robot collision avoidance:

$$||p_{i,k} - p_{j,k}|| \geq d_{min}, \quad \forall i \neq j,$$

Inter-formation separation:

$$||c_{m,k} - c_{n,k}|| \geq d^g_{min} rp, \quad \forall m \neq n,$$

Obstacle avoidance:

$$||p_{i,k} - o_l|| \geq r_l + d_{safe},$$

for each obstacle l with center o_l and radius r_l. These geometric safety constraints are standard in multi-robot navigation and formation control under complex environment.

5) Control Objective: The objective as in Eq. (5), is to design decentralized control policies,

$$u_{i,k} = \pi_i(\mathcal{O}_{i,k}) \tag{5}$$

based on local observations $\mathcal{O}_{i,k}$, such that:

1) each formation centroid converges to its assigned goal,
2) rigid formation geometry is preserved,
3) collisions and constraint violations are avoided,
4) performance remains robust under stochastic disturbances.

This results in a multi-objective control problem balancing formation accuracy, convergence speed, safety, and control effort.

6) Residual Reinforcement Learning: The model-based controllers provide structure and stability but lack adaptability, while purely learning-based approaches often struggle to enforce rigid geometric constraints. To address this, the control input is decomposed as Eq. (6),

$$u_{i,k} = u^{guide}_{i,k} + u^{res}_{i,k} \tag{6}$$

Where $u^{guide}_{i,k}$ is a model-based guidance control ensuring baseline safety and convergence, and $u^{res}_{i,k}$ is a learned residual policy. This residual RL paradigm constraints learning around

a stabilising controller, improving safety, convergence, and robustness. The residual policy is trained to maximize the expected discounted return in Equation (7),

$$\mathbb{E}\left[\sum_{k=0}^{\infty}\gamma^k r_k\right] \tag{7}$$

where the reward encodes formation error, centroid-to-goal distance, collision penalties, and control effort. Continuous-action actor-critic algorithms such as DDPG and TD3 are employed due to their suitability for high-dimensional continuous control and improved stability under uncertainties [3], [4].

A. Reward Function Design

The reward as in Equation (8) balances formation accuracy, goal attainment, safety constraints and control regularization to be able to achieve stable and efficient multi-robot navigation.

$$\begin{aligned} r_t = & \underbrace{-\alpha \sum_{m=1}^{M} e_m^{\text{form}}(t)}_{r_f} \underbrace{- \beta \sum_{m=1}^{M} \left\| c_{m,t} - p_m^{\text{goal}} \right\|}_{r_g} \\ & \underbrace{- \eta \sum_{i<j} \mathbb{I}(\|p_i - p_j\| < d_{\min})}_{r_i} \\ & \underbrace{- \zeta \sum_{i,l} \mathbb{I}(\|p_i - o_l\| < r_l + d_{\text{safe}})}_{r_o} \underbrace{- \rho \sum_{i=1}^{N} \|u_i\|^2}_{r_c} \end{aligned} \tag{8}$$

Where r_f is the formation reward, r_g is the goal reaching reward,r_i is the inter robot collision penalty, r_o is the obstacle avoidance penalty, and r_c is the control effort regularization. The scalar weighing coefficient ζ, η are the obstacle and inter-robot collision penalties, the formation error weight is α and goal progression weight is β, and control effort coefficient is ρ that determine the relative importance of each reward component during learning.

IV. METHODOLOGY

The framework follows the centralised-training, decentralised-execution (CTDE) paradigm [20] . During training, a shared policy is optimised using experience aggregated across all robots in the joint system, providing sample-efficient learning. At execution, each robot acts on its 18 dimensional local observation $O_{i,k}$ own position (3 elements), the relative positions of the two nearest teammates in the formation (4 elements), the centroid-to-goal vector in the formation (2 elements), and the relative positions of the three nearest obstacles (9 elements), without communicating with the other robots, both during training and at execution. This separation brings together the benefits of sample efficiency of centralised learning and decentralised control of deployment.

Rigid formation maintenance is enforced via a per-formation Kabsch projection applied after each policy update. Off-policy actor-critic policies (DDPG and TD3) are used to learn the residual policy, enabling stable continuous-action space learning and comparative benchmarking. The sample efficiency is ensured by centralisation of training, whereas the execution is decentralised with local observations, which makes it scalable to multi-formation navigation in cluttered environments.

V. CONTROL ARCHITECTURE AND ALGORITHMS

A. Residual Control Structure

The control input for robot i is defined as in Eq. (6). The gating factor $\lambda \in [\lambda_{min}, 1]$ is adaptively scheduled based on the minimum obstacle clearance and the distance to the goal, reducing the influence of the learned residual near safety-critical regions and allowing stronger learning corrections in free space. This structure constrains exploration around a stabilizing baseline controller, improving safety and training stability under disturbances.

B. Reinforcement Learning Algorithms (DDPG and TD3)

Two off-policy actor-critic algorithms are used for residual policy learning. DDPG employs a deterministic policy gradient with a single critic network for value estimation. TD3 extends DDPG by incorporating

1) twin critic networks to mitigate overestimation bias,
2) target policy smoothing by adding clipped noise to target actions, and
3) delayed policy updates to stabilize actor learning.

Actor–critic methods with continuous actions lack closed-form convergence guarantees even in the single-robot case [3], [4], and the non-stationarity of the multi-robot setting precludes standard convergence proofs; we therefore make no formal convergence claim.

C. Training and Execution Algorithm

The rigid formation control of WMRs with the DDPG algorithm is explained with Algorithm [1].

Algorithm 1 DDPG for Rigid Formation Control of WMRs

Require: $\mathcal{E}$, goals $\{\mathbf{g}_m\}$, formations, DDPG params, disturbance flag δ
Ensure: Trained policy π_θ
Init π_θ, Q_ϕ, targets, buffer $\mathcal{D}$, OU noise
for epoch $= 1..E$ **do**
  for $t = 1..T$ **do**
    $\mathbf{u}_t \leftarrow \mathbf{u}_t^g + \lambda_t(\pi_\theta(\mathbf{o}_t) + \mathcal{N}_{OU})$
    Step dynamics (+ noise if δ=1); rigid projection
    Store $(\mathbf{o}_t, \mathbf{u}_t, r_t, \mathbf{o}_{t+1}, d_t)$
  end for
  for $k = 1..K$ **do**
    Update Q_ϕ, π_θ; soft-update targets
  end for
end for
return π_θ

Similarly, the rigid formation of WMRs using TD3 is explained below with Algorithm [2]. In the next section, the simulation configuration and results are explored.

Algorithm 2 TD3 for Rigid Formation Control of WMRs

Require: Env $\mathcal{E}$, TD3 params, disturbance flag δ
Ensure: Policy π_θ
Init π_θ, Q_{ϕ_1}, Q_{ϕ_2}, targets, buffer $\mathcal{D}$
for epochs **do**
  for steps **do**
    Observe $\mathbf{o}_t$; compute guidance $\mathbf{u}_t^g$
    $\mathbf{u}_t^{res} = \pi_\theta(\mathbf{o}_t) + \epsilon$, $\epsilon \sim \mathcal{N}$
    Gate λ_t; $\mathbf{u}_t = \mathbf{u}_t^g + \lambda_t \mathbf{u}_t^{res}$
    Step unicycle dynamics (+noise if $\delta = 1$); Kabsch projection
    Store $(\mathbf{o}_t, \mathbf{u}_t, r_t, \mathbf{o}_{t+1})$ in $\mathcal{D}$
  end for
  Sample minibatch; target smoothing; update Q_{ϕ_1}, Q_{ϕ_2}
  Delayed actor update; soft-update targets
end for
return π_θ

D. Stability and Safety of the Residual Architecture

Formal Lyapunov guarantees for deep residual RL in multi-robot continuous-action systems remain unresolved; therefore, a bounded-perturbation interpretation is adopted. The guidance controller $u_{i,k}^{\text{guide}}$ provides practical stability for the unicycle dynamics through goal-attraction, obstacle-repulsion, and separation terms that reduce goal error while enforcing collision avoidance. The residual action $u_{i,k}^{\text{res}}$, bounded by $(v_{\max}, \omega_{\max})$, is scaled by an adaptive gate $\lambda \in [\lambda_{\min}, 1]$, such that

$$u_{i,k} = u_{i,k}^{\text{guide}} + \lambda u_{i,k}^{\text{res}}, \qquad \|u_{i,k} - u_{i,k}^{\text{guide}}\| \le \lambda u_{\max}^{\text{res}}. \tag{9}$$

As $\lambda \to \lambda_{\min} = 0.2$ near obstacles and goals, the policy remains within a bounded neighbourhood of the stabilising baseline, limiting unsafe exploration.

VI. SIMULATIONS AND RESULTS

The simulation set is shown in Table. I, where 12 WMRs are split into three formations with triangle(3 WMRs), Square(4 WMRs), and X(5 WMRs). Each formation has an individual goal and three obstacles in the environment. The rigid formation configuration of the WMRs is attained using the Kabsch projection.

TABLE I: Simulation Setup (DDPG/TD3)

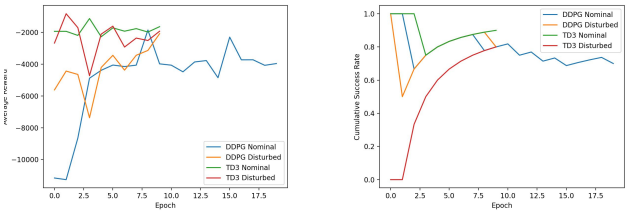
Item	Value
Workspace	$[-10, 10] \times [-10, 10]$ with three obstacles
Robots	$N = 12$ (Triangle:3, Square:4, X:5)
Dynamics	Discrete-time unicycle, $\Delta t = 0.2$ s
Actuation limits	$v_{\max} = 1.4$ m/s, $\omega_{\max} = 2.2$ rad/s
Formations	Rigid templates + Kabsch projection ($\alpha = 0.8$)
Goals	Per-formation goals; $r_g : 0.50 \to 0.18$ m (curriculum)
Horizon	$T = 640$ steps / episode
Residual gate	$\lambda_{\min} = 0.2$, scale $= 2.5$ (min rule)
Disturbance	Gaussian noise: $\sigma_p = 0.005$ m, $\sigma_\theta = 0.003$ rad
Algorithms	Residual DDPG, Residual TD3 (twin critics, delay=2)
Training	Replay buffer 10^6, $\gamma = 0.995$, $\tau = 0.005$
Metrics	Reward, success rate, formation error, time-to-goal

The framework is formation-agnostic: the same guidance law, residual policy, and Kabsch projection apply to all three templates without task-specific modification. The hyperparameters in Table. II follow established practice for off-policy continuous control [3], [4] and are held identical across DDPG and TD3. The discount factor $\gamma = 0.995$ suited to 640 step horizon, the target-update rate $\tau = 0.005$, replay buffer 10^6, and batch size $256 - 512$. The actor and critic learning rates reflect the faster critic timescale required before policy becomes reliable. TD3 specific settings target policy smoothing noise $\sigma = 0.2, \text{clip} = 0.5$ and update delay=2 are adopted from [4]

TABLE II: Hyperparameters of DDPG and TD3

Hyperparameter	DDPG	TD3
Actor learning rate	3×10^{-4}	3×10^{-4}
Critic learning rate	8×10^{-4}	8×10^{-4}
Discount factor(γ)	0.995	0.995
Target network update rate(τ)	0.005	0.005
Replay buffer size	10^6	10^6
Batch size	256-512	256-512
Exploration noise	OU noise	Gaussian noise
Policy noise	–	$\sigma = 0.2$
Noise clip	–	$c = 0.5$
Twin critics	No (single Q)	Yes (Q_1, Q_2)
Policy update delay	1 (every step)	2 (delayed)
Warm-up steps	10^3-5 $\times$ 10^3	10^3-5 $\times$ 10^3
Target policy smoothing	No	Yes
Residual gate floor	$\lambda_{\min} = 0.2$	$\lambda_{\min} = 0.2$
Residual gate scale	2.5	2.5
Optimizer (actor/critic)	Adam	Adam

The performance characteristics are plotted in Fig. 1 compares the average rewards and success rate of DDPG and TD3 algorithm.



(a) Average reward of DDPG and TD3 (b) Success rate of DDPG and TD3

Fig. 1: DDPG and TD3 under nominal and disturbances

TABLE III: DDPG Nominal vs. Disturbed (mean $\pm$ std)

Setting	Success (%)	Avg. Reward	Steps (succ.)	Time to Goal (s)	Form. Error (m)
Nominal	48.0 ± 50.0	-1170 ± 713	193 ± 69	38.7	0.033 ± 0.002
Disturbed	88.0 ± 32.5	-1645 ± 550	250 ± 100	50.0	0.026 ± 0.003

All the results are tabulated in the Table. III, Table. IV, and Table. V shows that the TD3 algorithm performs better than the DDPG algorithm in terms of success rate, average reward, and convergence towards the goal.

TABLE IV: TD3 Nominal vs. Disturbed (mean $\pm$ std)

Setting	Success (%)	Avg. Reward	Steps (succ.)	Time to Goal (s)	Form. Error (m)
Nominal	76–84 $\pm$ 37–43	-5259 to -8898	243–278 $\pm$ 60–125	48–56	0.033–0.034 $\pm$ 0.002–0.004
Disturbed†	40.0 ± 49.0	rescaled†	526 ± 56	105	0.117 ± 0.027

TABLE V: Comparing DDPG vs. TD3 configurations

Algorithm	Success (%)	Avg. Reward	Steps	Form. Error (m)	Collisions/ep
DDPG	48.0 ± 50.0	-1170 ± 713	193 ± 69	0.033 ± 0.002	0.40 ± 0.75
TD3	76–84 $\pm$ 37–43	-5259 to -8898	243–278	0.033 ± 0.002–0.004	2–8

To isolate the contribution of each architectural component, four TD3 configurations were evaluated under nominal conditions over ten episodes each, with one component disabled at a time while keeping the trained policy fixed (Table VI). The results show that the Kabsch projection is critical for geometric safety, as its removal preserves goal-reaching but increases formation error by more than 50 times and raises collisions from single digits to hundreds per episode.

TABLE VI: TD3 Ablation (10 episodes per configuration)

Configuration	Success (%)	Form. Error (m)	Collisions/ep	Effect
Full system	90–100	0.033 ± 0.002	7–11	baseline
No Kabsch projection	90–100	1.85–2.17	92–298	formation collapse
No adaptive gating ($\lambda = 1$)	0	0.033	3–16	no convergence
Guidance only (no residual)	0–10	0.037	0	RL needed

The adaptive gate is similarly essential for convergence, since enforcing $\lambda = 1$ reduces success to 0%, demonstrating that gate-constrained exploration stabilises learning around the guidance controller. Finally, the guidance-only baseline achieves at most 10% success, indicating that the model-based controller alone is insufficient for the multi-formation, multi-goal task and that the TD3 residual policy provides the required adaptive capability.

VII. CONCLUSION

In this paper, a safety-gated residual DRL framework is proposed to solve the rigid multi-formation control problem for WMRs in the presence of nonlinear dynamics, obstacle constraints and stochastic disturbances. The resulting geometry is maintained in a rigid manner by the Kabsch projection, while the baseline safety is ensured by the guidance controller, and the learned correction is provided by the adaptively gated residual policy. In addition, experiments with simultaneous triangle, square and X-formation revealed that TD3 outperforms DDPG in terms of stability and consistency of performance, and ablations verified the critical role of adaptive gating in achieving convergence and Kabsch projection in ensuring geometric safety. The study focuses only on the simulation using static obstacles, parameters that are manually tuned and stability guarantees that are of an empirical nature, not formal. In future, we will work with real robots, dynamic obstacles, partial observability, communication delays, and automatic optimisation of the reward and guidance parameters.

REFERENCES

[1] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: Algorithms and theory," *IEEE Transactions on Automatic Control*, vol. 51, no. 3, pp. 401-420, 2006. https://doi.org/10.1109/TAC.2005.864190

[2] Jorge Cortés, S. Martínez, T. Karatas, and F. Bullo,"Coverage control for mobile sensing networks," *IEEE Transactions on Robotics and Automation*, vol. 20, no. 2, pp. 243-255, 2004. https://doi.org/10.1109/TRA.2004.824698

[3] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," in *Proc. International Conference on Learning Representations (ICLR)*, 2016. https://doi.org/10.48550/arXiv.1509.02971

[4] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. International Conference on Machine Learning (ICML)*, 2018. https://doi.org/10.48550/arXiv.1802.09477

[5] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1238-1274, 2013. https://doi.org/10.1177/02783649134957

[6] S. Johannink *et al.*, "Residual reinforcement learning for robot control," *IEEE Robotics and Automation Letters*,vol. 4, no. 2, pp. 423-430, 2019. https://doi.org/10.48550/arXiv.1812.03201

[7] D. Silver *et al.*, "Residual policy learning," *arXiv preprint arXiv:1812.06298*, 2018. https://doi.org/10.48550/arXiv.1812.06298

[8] G. Wang, X. Wang, and S. Li, "A guidance module based formation control scheme for multi-mobile robot systems with collision avoidance," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 1, pp. 382–393, 2022. https://doi.org/10.1109/TASE.2022.3228186.

[9] J. Peng et al., "Consensus formation control of wheeled mobile robots with mixed disturbances under input constraints," *Journal of the Franklin Institute*, vol. 361, no. 17, p. 107300, 2024. https://doi.org/10.1016/j.jfranklin.2024.107300

[10] Cao, X., Li, K., and Li, Y., "Robust adaptive formation control for nonlinear multi-agent systems with range constraints," *Nonlinear Dynamics*, vol. 112, no. 3, pp. 1917–1929, 2024. https://doi.org/10.1007/s11071-023-09118-x.

[11] Q. Wu, X. Wang, and X. Qiu, "Embedded technique-based formation control of multiple wheeled mobile robots with application to cooperative transportation," *Control Engineering Practice*, vol. 150, p. 106002, 2024. https://doi.org/10.1016/j.conengprac.2024.106002

[12] S. Gu *et al.*, "Safe multi-agent reinforcement learning for multi-robot control," *Artificial Intelligence*, vol. 342, pp. 103-128, 2023. http://doi.org/10.1016/j.artint.2023.103905

[13] Ding, T.-F., et al., "Reinforcement learning formation tracking of networked autonomous surface vehicles with bounded inputs via cloud-supported communication," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 469–480, 2023. http://doi.org/10.1109/TIV.2023.3323767

[14] Y.-H. Wei, J.-W. Wang, and Q. Zhang, "Reinforcement learning-based optimal formation control of multiple robotic rollers," *Robotics and Autonomous Systems*, vol. 189, 2025. http://doi.org/10.1016/j.robot.2025.104947

[15] Xu, Z., Yan, T., Yang, S. X., Gadsden, S. A., & Biglarbegian, M. "Distributed robust learning based formation control of mobile robots based on bioinspired neural dynamics," *IEEE Transactions on Intelligent Vehicles*, vol. 20, pp. 1180–1189, 2024. https://doi.org/10.1109/TIV.2024.3380000

[16] Gong, Y., Xiong, H., Li, M., Wang, H., and Nian, X., "Reinforcement learning for multi-agent formation navigation with scalability," Applied Intelligence, vol. 53, no. 23, pp. 28207–28225, 2023. https://doi.org/10.1007/s10489-023-05007-3

[17] N. Aldana, et al., "Homography-Based Predictive Control for Smooth Visual Path Following in Autonomous Robots," *IFAC-PapersOnLine*, vol. 59, no. 17, pp. 335–340, 2025. https://doi.org/10.1016/j.ifacol.2025.10.186

[18] B. D. O. Anderson, C. Yu, B. Fidan and J. M. Hendrickx, "Rigid graph control architectures for autonomous formations," in *IEEE Control Systems Magazine*, vol. 28, no. 6, pp. 48-63, Dec. 2008. https://doi.org/10.1109/MCS.2008.929280

[19] Kabsch, Wolfgang. "A solution for the best rotation to relate two sets of vectors," *Foundations of Crystallography*, vol. 32, no. 5, pp. 922-923, 1976. https://doi.org/10.1107/S0567739476001873

[20] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in Proc. 31st Conf. *Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 6380–6391. https://doi.org/10.48550/arXiv.1706.02275