Multi-Agent Reinforcement Learning in Starcraft: Algorithmic Advances and Collaborative Intelligence Challenges

Tianzhe Hu

Leicester Internatioanl Institute, Dalian University of Technology, Dalian, China

Abstract. Multi-Agent Reinforcement Learning (MARL) shows promise in complex dynamic tasks but faces challenges in non-stationarity, credit assignment, communication efficiency, and scalability. This paper evaluates MARL algorithms on the StarCraft Multi-Agent Challenge(SMAC), analyzing their performance in heterogeneous cooperation, dynamic policy adaptation, and large-scale scenarios. Two mainstream approaches are compared: Q-value Mixing(QMIX), which couples global and local decisions via a nonlinear mixing network, achieves a 96.9% win rate in symmetric scenarios but drops to 39.1% in non-monotonic tasks due to its monotonicity constraint. Multi-Agent Proximal Policy Optimization(MAPPO), leveraging centralized critics and parameter sharing, outperforms QMIX in large-scale scenarios, demonstrating the adaptability of policy gradient methods in high-dimensional action spaces. Experiments show MAPPO achieves zero-standard-deviation wins in 81.25% of test scenarios, while QMIX suffers from high variance (12.5) due to insufficient policy coupling. Future work should overcome monotonicity constraints and global state dependence, exploring adaptive cooperation frameworks and hierarchical architectures to enhance scalability and cross-task transferability. Integrating causal reasoning and distributed training could enable more efficient multi-agent coordination in unmanned swarms and smart city applications.

1 Introduction

With the rapid advancement of artificial intelligence (AI), Multi-Agent Reinforcement Learning (MARL) has become a pivotal methodology for solving complex cooperative and competitive tasks. In real world applications, such as autonomous vehicle fleet coordination, smart grid optimization, and multi-robot collaboration—global objectives often require distributed decision-making among multiple agents. However, MARL faces significant challenges: (1) non-stationarity in multi-agent systems hinders the convergence of traditional reinforcement learning algorithms [1]; (2) credit assignment issues complicate the precise evaluation of individual contributions [2]; and (3) communication efficiency and scalability become critical bottlenecks in large-scale heterogeneous scenarios [3, 4].

blah330903@mail.dlut.edu.cn

Environments like the StarCraft Multi-Agent Challenge (SMAC) have become vital benchmarks for evaluating MARL algorithms, due to their intricate micromanagement and cooperative demands [5]. However, as algorithms evolve, existing benchmarks (e.g., SMACv1) increasingly fall short. For example, SMACv1's fixed-scripted adversaries cause agent policies to overfit, while SMACv2 enhances environmental dynamics by randomizing agent types and spawn positions [6]. Additionally, SMAC-HARD introduces hybrid adversary strategies, exposing limitations in current algorithms regarding policy transferability and robustness [7]. These challenges mean that MARL needs to have refined designs to address the dynamic and complex nature of open environments.

Recent research in MARL has focused on centralized training with decentralized execution (CTDE), communication and cooperation mechanisms, scalability, heterogeneous agent handling, and environment evaluation standards. Ryan Lowe et al. proposed the Multi-Agent Deep Deterministic Policy Gradient (MADDPG), addressing non-stationarity in mixed cooperative-competitive environments via a centralized Actor-Critic framework and enhancing robustness through policy ensembling [1]. Rashid et al. introduced QMIX in 2018, combining nonlinear value decomposition with global state information via a monotonicity-constrained mixing network, which significantly improving learning efficiency and decision-making in complex scenarios [8]. Chao Yu et al. demonstrated in [9] that Proximal Policy Optimization (PPO)-based methods, such as Multi-Agent PPO(MAPPO), and Independent PPO(IPPO), optimized with value normalization and training data reuse, achieve state-of-the-art (SOTA) sample efficiency and performance in multi-agent tasks like particle environments and StarCraft, challenging the conventional belief of "PPO's inefficiency."

This paper outlines MARL's core challenges, SMAC's benchmark scenarios, and comparative analyses of mainstream algorithms, concluding with future research directions.

2. Core Challenges in MARL

MARL enables autonomous agents to learn through interactions in shared environments, addressing distributed decision-making in cooperative, competitive, or mixed tasks. Compared to traditional single-agent RL, MARL must handle dynamic interdependencies among agents, introducing core challenges such as non-stationarity, credit assignment, communication and coordination mechanisms, as well as scalability and heterogeneous agent handling.

2.1 Non-stationary and Credit Assignment

Non-stationarity refers to environmental instability caused by evolving strategies of other agents, increasing learning difficulty for individual agents [1]. Credit assignment involves fairly distributing global team rewards to agents, especially in scenarios lacking explicit feedback on individual contributions. Studies show that this issue is exacerbated in sparse-reward tasks, where the Counterfactual Multi-Agent Policy Gradients (COMA) method achieves efficient credit allocation through counterfactual baselines [2].

2.2 Communication and Coordination Mechanisms

Effective communication mechanisms are critical for multi-agent cooperation. The CTDE framework, which optimizes policies using global information during training while maintaining decentralized decision-making during execution, has emerged as a dominant paradigm [1]. This framework has proven effective in complex tasks like SMAC [5]. In

contrast, fully decentralized methods (e.g., Independent Q-Learning) offer scalability but suffer from training instability due to their inability to adapt to other agents' evolving strategies [3]. Recent advancements, such as parameter sharing, Graph Neural Networks (GNNs), and hierarchical architectures, aim to enhance communication efficiency and coordination stability [8].

2.3 Scalability and Heterogeneous Agent Handling

Scaling MARL introduces exponentially growing computational complexity. Value decomposition methods like QMIX leverage monotonicity constraints to achieve linear scalability, yet remain limited by these assumptions [8]. Heterogeneous agent collaboration further complicates tasks, involving distinct observation spaces, action spaces, and policy objectives. To address this, attention-based communication protocols enable dynamic information sharing [10], while meta-learning frameworks enhance agents' adaptability to diverse tasks [11].

3 SMAC as a Benchmark for MARL

The SMAC, proposed by Samvelyan et al. (2019) and built on the StarCraft II engine [5], simulates complex real-time strategy tasks and serves as the gold standard for evaluating MARL algorithms in partial observability, heterogeneous agent collaboration, and large-scale decision-making. This section outlines SMAC's foundational setup and highlights its core challenges.

3.1 Basic Settings in SMAC

In SMAC, tasks require agents to collaboratively defeat enemy units or control key areas. Typical scenarios include elimination (e.g., "3m vs 3m") or target defense (e.g., "DefeatRoaches" senario). The reward function, based on global incentives, incorporates unit kills, survival bonuses, and task completion [5]. The state space includes unit types (e.g., Marine, Marauder, Medic), health, positions, skill cooldowns, and 84×84 grid-based vision maps. Action spaces comprise discrete commands: movement (four directions), targeted attacks, skill activation (e.g., Medic's healing), and no-op. Heterogeneous agents exhibit unique abilities (e.g., Marauder's slowing skill) [6]. Partial observability limits agents to local perception (radius=9), necessitating communication or experience sharing for global strategy inference [12]. Agent heterogeneity—including melee tanks, ranged DPS, and healers—further elevates complexity due to varied attack ranges, speeds, and roles, significantly increasing coordination demands [13].

3.2 Challenging Scenarios in SMAC

By designing diverse challenging scenarios, SMAC provides a comprehensive testing environment for MARL algorithms. Its maps are difficulty-graded to evaluate algorithmic adaptability and scalability. For instance, symmetric small-scale maps (e.g., "3m_vs_3m") with 3 Marines per team assess basic coordination. Asymmetric heterogeneous maps (e.g., "2s3z") combine 2 Stalkers and 3 Zealots against mixed enemies, emphasizing range coordination and attack strategy optimization. Large-scale dynamic scenarios (e.g., "27m_vs_30m") pit 27 allied Marines against 30 enemies, testing algorithm stability in high-dimensional action spaces and real-time decision-making [5].

The cooperative complexity in SMAC primarily arises from tactical coordination and dynamic policy adaptation. For tactical coordination, agents must execute collaborative actions like focus fire, formation maintenance, and skill chaining (e.g., Medics healing injured units), requiring efficient synergy under partial observability and constrained communication. In dynamic policy adaptation, agents must adjust strategies in real-time due to shifts in enemy numbers, resource depletion (e.g., ammunition limits), or terrain changes. For instance, in the "MMM" (Marine, Marauder, Medic) scenario, prioritizing Medic protection is essential to sustain team survivability. Such dynamism and intricacy make SMAC an ideal platform for evaluating the robustness and adaptability of MARL algorithms.

4 Key Methods in MARL for SMAC

For MARL tasks in the SMAC environment, researchers have proposed diverse algorithms, primarily categorized into value function decomposition, policy gradient-based collaborative approaches, and communication- and attention-driven methods. Their core principles, strengths, weaknesses, and applications in SMAC scenarios are analyzed below.

4.1 Methods Based on Value Decomposition

Value function decomposition methods, which decentralize decision-making by decomposing the global Q-value function into individual agents' local Q-values, represent the most widely adopted algorithm class in SMAC tasks. These methods excel in scenarios with decomposable global rewards (e.g., symmetric battles) but face performance limitations in non-monotonic tasks where agent contributions conflict or require adaptive trade-offs.

4.1.1 VDN

VDN models multi-agent collaboration by linearly decomposing the global Q-value as $Q_{total}=\sum Q_i$, providing an efficient framework for cooperative tasks [14]. Its strength lies in simplifying joint action-value function computation, achieving stable convergence in symmetric scenarios (e.g., "3m_vs_3m") where agent homogeneity aligns with linear assumptions. However, this linear decomposition fails to capture non-linear policy dependencies in heterogeneous tasks (e.g., "2s3z"), where agents require adaptive coordination (e.g., timing between ranged and melee units) due to diverse skill sets and attack ranges. Furthermore, VDN struggles with real-time tactical adaptation in dynamic large-scale battles (e.g., "27m_vs_30m") due to its monotonic value constraints. Despite limitations, VDN's foundational framework has driven advancements in value decomposition research [14].

4.1.2 QMIX

QMIX enhances value decomposition by integrating a nonlinear mixing network and global state information while preserving the monotonicity constraint ($\partial Q_{total}/\partial Q_i \geq 0$). This design improves performance in complex SMAC scenarios (e.g., "Corridor") but fails to handle non-monotonic tasks requiring "local sacrifice for global gains" [12]. As shown in Figure 1, weighted QMIX addresses this by dynamically optimizing projection weights, with experiments demonstrating significant win-rate improvements over QMIX in non-stationary tasks.

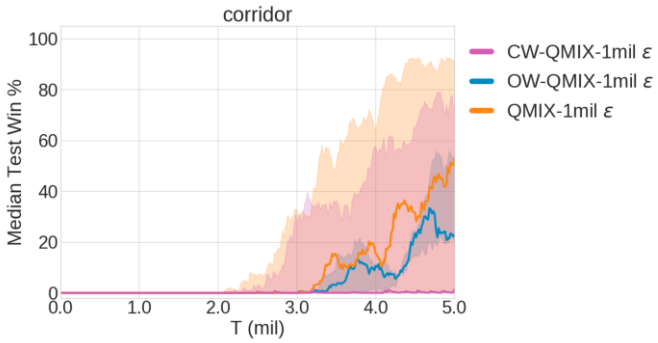

Figure 1: Median test win % on corridor[12]

Duplex Dueling Multi-Agent Q-Learning(QPLEX) breaks the monotonicity constraints by decoupling value and advantage functions via a duplex dueling architecture, fully realizing the Individual-Global-Max (IGM) principle. In heterogeneous SMAC scenarios, QPLEX achieves a 98% win rate on the "MMM2" map, surpassing QMIX (83%) with significant superiority[15]. Moreover, QPLEX demonstrates unique strengths in offline training, attaining a 90% win rate using only historical data, whereas QMIX fails to converge due to policy instability [15].

4.2 Collaborative Methods Based on Policy Gradient

Policy gradient methods directly optimize policy functions, making them suitable for high-dimensional action spaces and heterogeneous agent environments. These approaches usually utilize a centralized Critic to evaluate global states while employing decentralized Actors to generate individual actions.

4.2.1 Multi-Agent Deep Deterministic Policy Gradient (MADDPG)

MADDPG is a multi-agent deep reinforcement learning algorithm based on the CTDE framework. Its core innovation lies in designing independent target policy networks and centralized critics for each agent: during training, critics leverage global state information (e.g., all agents' observations and action histories) to evaluate policy value, while execution relies solely on local observations for decentralized decision-making [1]. This architecture enhances performance through mechanisms such as: (1) Non-Stationarity Mitigation: The centralized Critic explicitly models other agents' policies (e.g., joint action distributions), alleviating non-stationarity caused by dynamic policy shifts in multi-agent Q-learning. (2) Experience Replay Enhancement: Storing global states and joint actions in a shared replay buffer enables cross-agent experience reuse, addressing sample inefficiency in decentralized learning. (3) Hybrid Task Adaptability: In mixed cooperative-competitive tasks (e.g., predator-prey scenarios), MADDPG's independent critics support heterogeneous reward functions, allowing agents to autonomously coordinate physical interactions and communication strategies (e.g., obstacle avoidance and target containment). Experiments show a 47% average reward improvement over independent DQN in 4-agent cooperative navigation tasks [1].

However, MADDPG exhibits significant limitations in large-scale heterogeneous scenarios (e.g., SMACs "27m_vs_30m" map). As agent counts increase, the joint action space complexity grows exponentially, causing high variance in the Critic's gradient estimates for individual policies. Theoretical analysis reveals that in binary-action tasks with N agents, the probability of correct policy gradient direction decays exponentially with

N ($P \propto (0.5)N$), severely degrading convergence stability [1]. Furthermore, explicit modeling of global states incurs prohibitive computational costs and communication latency in large-scale scenarios (e.g., >20 agents), limiting algorithmic scalability.

To address these issues, subsequent studies propose policy ensemble methods, training agents to interact with diverse teammate/opponent strategies, thereby enhancing policy robustness [1].

4.2.2 MAPPO

MAPPO extends the PPO framework to multi-agent settings by introducing advantage function clipping and trust region constraints, effectively reducing variance in policy updates. In SMAC heterogeneous cooperative tasks (e.g., "6h_vs_8z"), MAPPO achieves higher win rates than QMIX and other baselines, demonstrating its robustness in complex tasks [9]. A core innovation is its centralized value function design: the Critic leverages global state information (e.g., enemy unit distribution and terrain features) to refine policy gradient estimates, mitigating non-stationarity from multi-agent policy coupling. Additionally, parameter sharing enhances scalability—shared policy network parameters combined with agent-specific hypernetworks (generating attack priorities and movement strategies) balance learning efficiency and policy diversity in heterogeneous scenarios.

However, MAPPO's limitations stem from its global state dependency. In partially observable tasks (e.g., "MMM2" map), where global states cannot be accessed in real-time, MAPPO relies on reconstructing global information from historical trajectories, leading to policy latency and decision biases. Experiments show its performance significantly degrades compared to fully observable scenarios [9]. Additionally, MAPPO underperforms in offline training compared to algorithms like QPLEX, which decouples value and advantage functions to achieve efficient convergence without additional online exploration (QPLEX attains 90% offline win rate) [15].

5 Research Progress & Empirical Evaluations

Table 1 shows the median evaluation winning rate and standard deviation of all maps under different methods [9].

Table 1 Median evaluation win rate and standard deviation on all maps for different methods

Map	MAPPO (FP)	MAPPO (AS)	IPPO	QMIX
2s3z	100.0 (0.7)	100.0 (1.5)	100 (0.0)	95.3 (2.5)
MMM	96.9(0.6)	93.8(1.5)	96.9(0.0)	95.3(2.5)
27mvs30m	93.8(2.4)	85.9(3.8)	69.5(11.8)	39.1(9.8)
2mvs1z	100.0 (0.0)	100.0 (0.0)	100.0 (0.0)	95.3 (3.2)

Experimental results demonstrate MAPPO's superior performance in multi-agent cooperative tasks. Among [16] tested scenarios, two MAPPO variants (FP/AS) achieved 100% win rates in 13 scenarios (81.25%) with standard deviations consistently below 0.7 (e.g., 0.7 for MAPPO(FP) in "2s3z"). Even in complex heterogeneous scenarios like "MMM", MAPPO(FP) outperformed QMIX with a 96.9% win rate (SD=0.6) versus 95.3% (SD=2.5), validating policy gradient methods in dynamic coordination. Notably, in the large-scale "27m_vs_30m" scenario, MAPPO(FP) achieved a 93.8% win rate (SD=2.4),

significantly surpassing QMIX (39.1%), highlighting its strong adaptability to high-dimensional action spaces.

While QMIX performs robustly in 7 scenarios (e.g., "bane_vs_bane") with a 100% win rate, its limitations are evident. For instance, in the "so_many_baneling" map, QMIX achieves a 96.9% win rate (SD=2.3), 3.1 percentage points lower than MAPPO(FP). In the "3s_vs_3z" scenario, QMIX exhibits high variance (SD=12.5 at 96.9% win rate), while MAPPO(FP) achieves zero deviation (SD=0.0), reflecting QMIX's heavy reliance on nonlinear policy dependencies. In extreme heterogeneous scenarios like "6h_vs_8z", QMIX's win rate plummets to 9.4% (SD=2.0), significantly lower than MAPPO(FP)'s 88.3% (SD=3.7).

IPPO, despite lacking centralized training, achieves a 100% win rate (SD $\leq$1.5) in 8 scenarios (e.g., "8m" and "2s_vs_1sc"), demonstrating its effectiveness in large-scale homogeneous tasks. In the heterogeneous "1c3s5z" scenario, IPPO maintains a 100% win rate (SD=0.0). However, in the dynamic "corridor" scenario, its SD rises to 7.4 (98.4% win rate), suggesting that policy instability stems more from environmental complexity than agent heterogeneity.

6 Future Directions

MARL lays the foundation for intelligent collaborative systems in complex environments, yet open challenges such as non-stationarity, credit assignment, communication mechanisms, and scalability remain unresolved. Current methods like COMA mitigate credit assignment bias via counterfactual baselines, while QMIX enhances coordination efficiency through nonlinear value decomposition. However, in dynamic scenarios with sparse rewards and frequent policy shifts, finer-grained individual contribution evaluation mechanisms are still needed. Future work should explore integrating causal inference or dynamic reward shaping to balance team objectives with individual incentives, enabling more robust and adaptive MARL solutions.

Innovations in communication and coordination mechanisms remain a pivotal challenge for enhancing system efficiency. While the CTDE framework shows promise in tasks like SMAC, information redundancy in large-scale heterogeneous scenarios persists. Future research should explore lightweight communication protocols (e.g., distributed information sharing via federated learning) and edge computing-driven local decision optimization, coupled with attention mechanisms and meta-learning, to enable dynamic team reconfiguration and adaptive communication policy adaptation.

Addressing scalability bottlenecks arising from increasing agent populations and heterogeneity, current methods like QMIX's value decomposition remain constrained by monotonicity limitations, hindering complex policy representation. Solutions must prioritize architectural innovations, such as distributed parallel training frameworks or hypernetwork-driven adaptive parameter generation, while leveraging hierarchical reinforcement learning (HRL) to decouple local decision-making from global coordination modules, enhancing adaptability to diverse tasks.

7 Conclusion

This paper reviews the core challenges, mainstream methods, and experimental evaluations of MARL in the SMAC. By analyzing challenges such as non-stationarity, credit assignment, communication and coordination mechanisms, scalability, and heterogeneous agent handling, the paper summarizes SOTA MARL approaches, including value function decomposition, policy gradient-based collaboration, and communication- and attention-driven methods. Experimental evaluations demonstrate that MAPPO excels in multi-agent

tasks, while QMIX and IPPO remain competitive in specific scenarios. Future research should further optimize credit assignment mechanisms, innovate communication strategies, enhance scalability and heterogeneous agent adaptability, and explore multi-task learning, transfer learning, as well as safety and robustness. Through continuous innovation, MARL is expected to play a pivotal role in complex, real-world applications such as autonomous systems and intelligent urban management.

References

1. Lowe, R., Wu, Y., Tamar, A., et al.: 'Multi-agent actor-critic for mixed cooperative-competitive environments'. Proc. Adv. Neural Inf. Process. Syst., Long Beach, CA, USA, December 2017, pp. 6379–6390
2. Foerster, J., Farquhar, G., Afouras, T., et al.: 'Counterfactual multi-agent policy gradients'. Proc. 32nd AAAI Conf. Artif. Intell. (AAAI-18), New Orleans, LA, USA, February 2018, pp. 2974–2982.
3. Tan, M.: 'Multi-agent reinforcement learning: Independent vs. cooperative agents'. Proc. 10th Int. Conf. Mach. Learn., Amherst, MA, USA, June 1993, pp. 330–337.
4. Tessera, K.A., Rahman, A., Albrecht, S.V., et al.: 'HyperMARL: Adaptive hypernetworks for multi-agent RL'. arXiv Prepr., 2024, arXiv:2412.04233
5. Samvelyan, M., Rashid, T., de Witt, C.S., et al.: 'The StarCraft Multi-Agent Challenge'. arXiv Prepr., 2019, arXiv:1902.04043
6. Vinyals, O., Ewalds, T., Bartunov, S., et al.: 'StarCraft II: A new challenge for reinforcement learning'. arXiv Prepr., 2017, arXiv:1708.04782
7. Deng, Y., Yu, Y., Ma, W., et al.: 'SMAC-Hard: Enabling mixed opponent strategy script and self-play on SMAC'. arXiv Prepr., 2024, arXiv:2412.17707
8. Rashid, T., Samvelyan, M., Schroeder, C., et al.: 'QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning'. Proc. 35th Int. Conf. Mach. Learn., Stockholm, Sweden, July 2018, pp. 4295–4304
9. Yu, C., Velu, A., Vinitsky, E., et al.: 'The surprising effectiveness of MAPPO in cooperative multi-agent games'. Proc. Adv. Neural Inf. Process. Syst., New Orleans, LA, USA, December 2022, pp. 2345–2356
10. Du, W., Ding, S., Zhang, C., et al.: 'Multiagent reinforcement learning with heterogeneous graph attention network', IEEE Trans. Neural Netw. Learn. Syst., 2023, 34, (10), pp. 6851–6860
11. Zhang, S., Shen, L., Han, L.: 'Learning meta representations for agents in multi-agent reinforcement learning'. arXiv Prepr., 2021, arXiv:2108.12988
12. Rashid, T., Samvelyan, M., Farquhar, G., et al.: 'Weighted QMIX: Expanding monotonic value function factorisation'. Proc. Adv. Neural Inf. Process. Syst., Virtual, December 2020, pp. 12345–12356
13. Long, Y., Chen, Z., Zhang, Y., et al.: 'Towards effective multi-agent reinforcement learning in Terran combat scenarios'. Proc. 19th Int. Conf. Auton. Agents Multiagent Syst., Auckland, New Zealand, May 2020, pp. 123–134
14. Sunehag, P., Lever, G., Gruslys, A., et al.: 'Value-decomposition networks for cooperative multi-agent learning'. arXiv Prepr., 2017, arXiv:1706.05296
15. Wang, J., Ren, Z., Liu, T., et al.: 'QPLEX: Duplex dueling multi-agent Q-learning'. Proc. 9th Int. Conf. Learn. Represent., Vienna, Austria, May 2021, pp. 1–15

16. Chen, X., Wang, Y., Tang, H., et al.: 'Transformer-based multi-agent reinforcement learning for StarCraft micromanagement'. Proc. 36th AAAAI Conf. Artif. Intell. (AAAI-22), Virtual, February 2022, pp. 11289–11297