

State-of-the-art in Robot Learning for Multi-Robot Collaboration: A Comprehensive Survey

Bin Wu and C. Steve Suh

Abstract—With the continuous breakthroughs in core technology, the dawn of large-scale integration of robotic systems into daily human life is on the horizon. Multi-robot systems (MRS) built on this foundation are undergoing drastic evolution. The fusion of artificial intelligence technology with robot hardware is seeing broad application possibilities for MRS. This article surveys the state-of-the-art of robot learning in the context of Multi-Robot Cooperation (MRC) of recent. Commonly adopted robot learning methods (or frameworks) that are inspired by humans and animals are reviewed and their advantages and disadvantages are discussed along with the associated technical challenges. The potential trends of robot learning and MRS integration exploiting the merging of these methods with real-world applications is also discussed at length. Specifically statistical methods are used to quantitatively corroborate the ideas elaborated in the article.

Index Terms—Multi-robot, Cooperation, robot learning

I. INTRODUCTION

With the advancement of technology and the development of artificial intelligence [1], [2], [3], robot learning has become one of the key factors driving the progress of robotic technology. Especially in the field of MRS [4], robot learning has shown great potential and application value. MRS complete complex tasks by coordinating the actions of multiple robots, offering higher efficiency, reliability, and flexibility compared to single robot systems (SRS) [5], [6]. However, with the diversification of application scenarios and the increase in task requirements, the learning mechanisms in MRS face many challenges, including collaborative learning, communication constraint, environmental adaptability, and algorithm's ability to generalize [7], [8].

This article reviews the latest research in robot learning within MRS, including theoretical foundations, key technologies, applications, challenges faced and paths being explored. An extensive review of existing literature allows the core issues and proper solution strategies to be identified along with the performance and limitations of different learning methods for practical applications.

First, the basic concepts of MRS and robot learning are introduced and their importance in current technological context is stated. Next, the design principles of learning mechanisms in MRS is explored in-depth, including, but not limited to, technologies such as Reinforcement Learning (RL), Transfer Learning (TL), and Imitation Learning (IL). Much insight is gained from evaluating the advantages and disadvantages of different learning strategies essential to

inspiring future research and charting the path moving forward. Moreover, successful case studies of MRS in specific applications are examined, such as automated warehousing, search and rescue, environmental monitoring, and precision agriculture, to showcase the effectiveness and applicability of robot learning technologies in solving real-world problems. The main challenges encountered in implementing multi-robot learning systems, including resource allocation, task decomposition, learning efficiency, and system scalability are also discussed. Finally, the article looks forward to the various directions of future development of robot learning in MRS, emphasizing the importance of interdisciplinary collaboration and how the integration of artificial intelligence, machine learning, control theory, and cognitive science will elevate the field to a greater level of development. The survey paper should provide researchers in the related fields with a comprehensive reference framework, inspire more innovative research and application, and jointly promote the progress of MRS and robot learning.

II. DEFINITION AND SCOPE

In this section, more detailed definitions of the key concepts addressed in this article are provided along with delineation of the scope of current issues. This is done to help establish a unified cognitive basis and also facilitate a deeper discussion of the issues based on the basis.

A. Multi-robot Systems

From a system-level perspective, the advantages of MRS over SRS are evident [4], [9]. Figure 1 indicates the fundamental differences of the two systems. An MRS refers to a system composed of two or more robots that can complete specific tasks or objectives through collaboration or competition. In such systems, each robot may have unique capabilities and limitations, and they achieve collective intelligence and action through communication and coordination. Research on MRS mainly focuses on how to effectively design and implement interaction, communication, and collaboration mechanisms among robots to improve the efficiency and effectiveness of the entire system [9].

An MRS can be represented as a tuple (N, S, A, T, R, C, G) .

- **Robot Set:** Let $N = \{1, 2, \dots, n\}$ be the set of robots in an MRS, where each i represents an independent robot entity.
- **State Space:** For each robot i , its state can be represented by an element in the state space S_i . The state

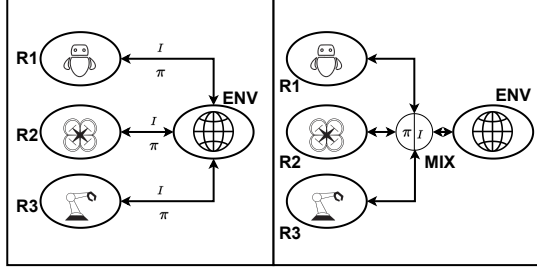


Fig. 1: Single robot vs multi-robot interaction decision framework

space of the entire system is the Cartesian product of these individual state spaces, i.e., $S = S_1 \times S_2 \times \dots \times S_n$.

- **Action Space:** Each robot i can perform a series of actions, defined by the action space A_i . Similarly, the action space of the entire system is the Cartesian product of the individual action spaces, i.e., $A = A_1 \times A_2 \times \dots \times A_n$.
- **Transition Function:** The transition function $T : S \times A \rightarrow S$ defines how the system state changes based on the combination of actions performed. For a given current state and combination of actions, the transition function returns the next state of the system.
- **Reward Function:** $R : S \times A \rightarrow \mathbb{R}$ assigns a real number reward value to each state and action combination, reflecting the effectiveness of that combination in achieving the system's goals.
- **Communication Model:** In an MRS, communication between robots can be represented by the communication model C , which defines how robots exchange information with other robots or the system.
- **Goal Function:** MRS usually have one or more goals, which can be represented by the goal function $G : S \rightarrow \mathbb{R}$, evaluating the extent to which the system achieves its goals in a specific state.

B. Multi-robot Cooperation

The concept and mathematical definition of MRC problems can be further developed based on the foundational framework of MRS in the last section [5], [6]. MRC problems as seen in Figure 2 involve a group of robots that share information, coordinate actions, and make decisions together to achieve a common task or goal. A more detailed application is provided by Figure 3. The primary objective of cooperation is to utilize the collective capabilities of multiple robots to accomplish tasks that are not possible or inefficient for a single robot. This cooperation may include aspects such as task allocation, joint decision-making, resource sharing, coordinated actions, and goal sharing. To achieve this, a common goal function ($G : S \rightarrow \mathbb{R}$) can be defined to measure and observe the performance of the entire system in achieving the common goal.

- **Collaboration Strategy:** Define a set of strategies $\Pi = \{\pi_1, \pi_2, \dots, \pi_n\}$, where each $\pi_i : S \rightarrow A_i$ is a decision rule guiding robot i in choosing actions in a given state.

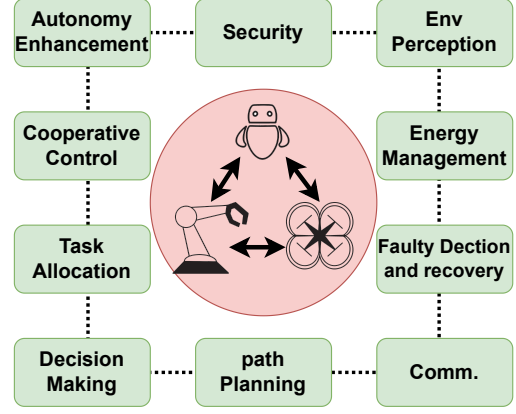


Fig. 2: Sub-problems of multi-robot systems

- **Communication Model Extension:** Extend the communication model C to $C_e : S \times N \rightarrow \mathcal{P}(N)$, where $\mathcal{P}(N)$ is the power set of the robot set N , indicating which groups of robots can communicate in a given state.
- **Joint Action:** Define a joint action mapping $\alpha : S \times \Pi \rightarrow A$, combining the strategies of all the robots and current state to produce the action of the entire system.
- **Collaborative Benefit:** Introduce a utility function $U : S \times A \rightarrow \mathbb{R}$ to evaluate the performance of the system as a whole under a given state and action.
- **Constraints:** Consider the constraints of interactions and cooperation among robots, such as communication range, resource limitation, and task dependency.

C. Robot Learning

When discussing robot learning, researchers often first consider it as an interdisciplinary field bringing together machine learning and robotics, which is undoubtedly correct. However, before clarifying the essence of robot learning, what learning is must be clarified. In sociology [10], learning is usually defined as the process by which individuals or groups acquire social behavior patterns through the process of socialization. This includes the internalization of social norms, values, language, skills, and behavior patterns. Sociologists emphasize the impact of social structures and culture on the individual learning process. The definition of learning in anthropology [11] emphasizes the process of cultural transmission and adaptation. Learning is seen as the way individuals acquire knowledge, skills, beliefs, and behavior patterns from their cultural environment. Anthropologists often focus on the differences in learning across cultures or ethnic groups and how these differences affect the adaptation and development of individuals and communities. In psychology [12], learning is typically defined as a relatively permanent change in behavior or thought patterns produced by experience. This definition emphasizes an individual's response and adaptation to environmental stimuli, including cognitive learning, emotional learning, and behavioral learning.

In traditional disciplines, there are far more definitions of

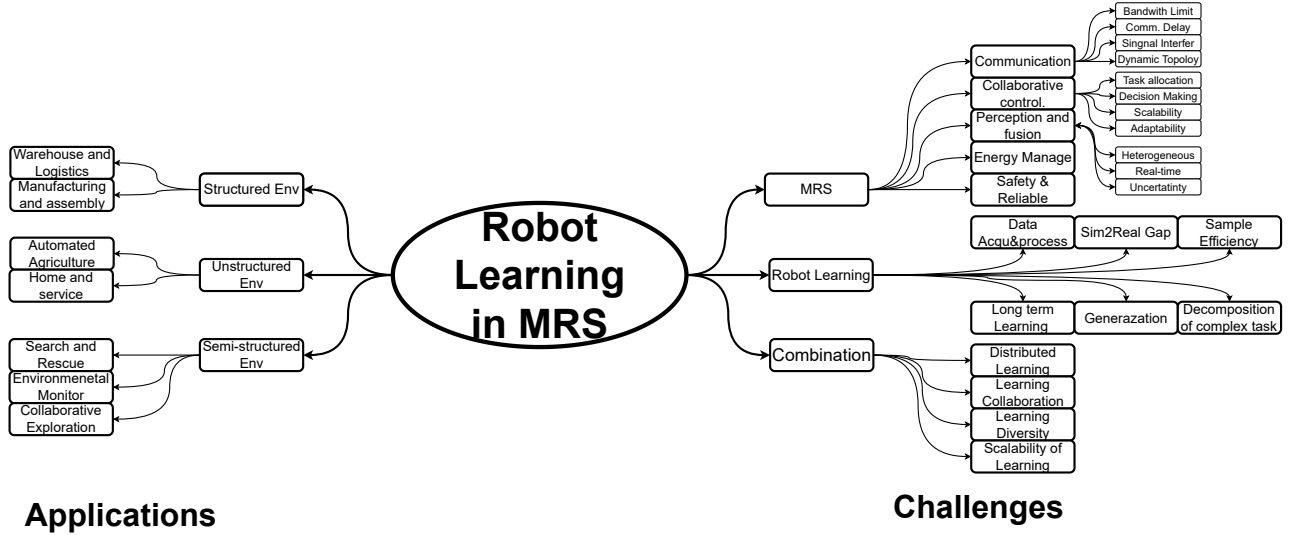


Fig. 3: Robot learning in the context of MRS: applications and challenges

learning than those listed above, but this does not prevent one from finding some commonalities. In this article, the concept of learning from four dimensions is discussed:

- **Utilization of Experience:** Learning is a process of acquiring knowledge or skills through experience.
- **Behavioral and Cognitive Changes:** Learning leads to changes in behavior or ways of thinking.
- **Environmental Adaptation:** Learning is considered a way for individuals or groups to adapt to their environment.
- **Process Nature:** Learning is viewed as a continuous process, rather than a single event. This process involves the constant acquisition, processing, and application of information, as well as gradual adaptation and change over time.

Returning to the definition of robot learning, it refers to the process where a robot merges hardware (such as sensing, processing, and executing components) with software (specific learning algorithms) to utilize data, experience, or interactions with the environment to acquire new knowledge or skills, thereby improving its performance [13], [14], [15]. This process involves perception (acquiring information through sensors), decision-making (making decisions based on learned knowledge), and action (executing tasks). The goal of robot learning is to enable robots to autonomously adapt to new tasks and environments, enhancing their flexibility and efficiency. Suppose a robot's strategy ($\pi : S \rightarrow A$) is a mapping from state s to action a , guiding the robot in choosing actions in a given state, then, the learning algorithm is used to extract patterns and knowledge from the data to improve the strategy (π).

D. Robot Learning in MRC

In the section above, the concepts of MRC and robot learning were separately introduced. Building on this foundation, it naturally leads to another question: what is the

difference between robot learning in MRC and single robot learning? Before answering this question, an assumption needs to be made that, regardless if MRS or SRS is being considered, each robot makes decisions based on the local information it acquires. Below, the question is addressed from two aspects: 1. Utilization of information, where robots in a MRS exchange information. 2. Joint decision-making, whether it's competitive or cooperative decision-making, there is a coordination mechanism among the robots in a MRS to promote more efficient operation of the overall system. These two basic points are both the advantage and challenge of multi-robot learning. Exchange of information (experience) provides the robots with more learning materials but also increases learning burden. Joint decision-making can create a synergy effect, while also imposing more restrictions on each robot's decision-making, increasing the difficulty of learning.

III. LEARNING METHOD

The concept of machine learning was inspired by the ways human and animal learn [16], [17]. Researchers design algorithms and machines by simulating how the human brain works, hoping that machines can acquire knowledge and skill through observation and experience. The classification of learning methods discussed in this section is also based on a logical division, mirroring human or animal learning methods. Afterward various robot learning methods is discussed in the context of MRS.

A. Carbon-based vs. Silicon-based

First, one can draw from psychology and neuroscience the following classifications of human and animal learning methods:

- **Classical Conditioning:** Conducted by the Russian physiologist Pavlov [18], who discovered how animals learn

through associating stimuli with responses via experiments.

- **Operant Conditioning:** Introduced by B.F. Skinner in his book [19], the concept involves increasing or decreasing the frequency of specific behaviors through rewards and punishments.
- **Observational Learning:** Individuals learn by observing others' behaviors and their consequences [12].
- **Cognitive Learning:** Emphasizes the role of exploration and problem-solving in the learning process [20].
- **Sociocultural Learning:** Discusses how social interactions influence cognitive development [21].
- **Affective Learning:** Explores how emotions affect learning and memory, especially the neural mechanisms of fear responses [22].

Although these do not cover all known ways of learning in human and animal, however, they do provide a brief classification basis for robot learning in MRS. Mapping the learning methods of human and animal to those of machine learning provides an interesting perspective on how robots emulate natural learning processes, as indicated in Figure 4. The classification of learning methods based on carbon-based life forms such as human and animal can be mapped onto the principles of the learning processes of existing machine learning *silicon – based* methods.

1) *Classical Conditioning:* Classical conditioning involves learning through the association between stimuli. In the field of machine learning, the mechanism most similar to this associative learning is supervised learning [23], [24]. This type of machine learning involves mapping between inputs (similar to conditioned stimuli) and outputs (similar to conditioned responses). Training data includes inputs and their corresponding outputs, and the model learns from these data to predict the output of new inputs. For example, when using neural networks for image recognition, the model learns the relationship between patterns in images and labels, similar to the associative learning between stimuli in classical conditioning.

2) *Operant Conditioning:* Operant conditioning focuses on the relationship between behaviors and consequences, which is very similar to the concept of RL [25], [26]. In RL, an agent learns behavior strategies through interaction with the environment to maximize certain cumulative rewards. This learning process involves exploration (trying new behaviors to discover effective strategies) and exploitation (using known strategies to obtain rewards). The mechanism of RL is similar to operant conditioning, relying on the consequences of behaviors (rewards or punishments) to form or change behaviors.

3) *Observational Learning:* In machine learning methods analogous to observational learning, unsupervised learning [23], [27] and imitation learning [15], [28] stand out as particularly representative. Observational learning involves observing the behaviors and outcomes of others and learning from them. If one likens certain aspects of observational learning to feature learning or clustering in the field of machine learning, then the process in unsupervised learning,

which involves identifying patterns and structures from data without explicit labels or feedback, shares strong similarities with observational learning. Inference learning (IL) is a method that allows robots or software agents to observe and mimic the behaviors of human experts or other agents. The key components of IL include: 1. Demonstrations - Behavioral demonstrations observed by the learner, usually provided by human experts or other advanced agents. 2. Behavior Cloning - A method of learning behavior acquired directly from demonstration, without the need for explicit modeling of the environment. 3. Inverse RL - Inferring a reward function through observation of demonstrations and then using this reward function to guide the learning process. It is evident that IL and observational learning of animals share a notable connection, with their principles being similar in many ways. Both learning methods involve learning from the observation of others' behaviors and imitating or replicating these observed behaviors in future actions.

4) *Cognitive Learning:* In machine learning, there are many methods that correspond to cognitive learning. For example, Transfer Learning (TL) [29], [30] allows a model to apply existing knowledge (usually learned on one task) to another related but different task. This method is particularly useful in situations with limited data, as it can reduce the amount of data and computational resources needed to train on a new task. TL typically involves learning feature representations from a source task and then adapting these representations to a target task.

Causal inference learning (CIL) [31], [32] focuses on learning the causal relationships between variables from data, rather than just correlations. This involves using statistical methods, experimental design, and computational models to determine whether one event (the cause) directly leads to another event (the effect) and how to quantify this impact. Both cognitive learning and CIL focus on understanding causal relationship. In cognitive learning, individuals understand causal relationships through observation, experience, and reasoning, while in CIL, algorithms identify and validate causal relationships through data analysis.

ML [33] is defined as the process of "learning how to learn." The goal of ML is to enable machine learning models to optimize and improve the learning process through previous experiences, allowing them to adapt and learn more quickly when faced with new tasks. Both cognitive learning and ML involve higher-level learning on top of the basic learning process. In cognitive learning, this is manifested as metacognitive abilities [34], i.e., understanding and managing one's own learning process. In ML, this is manifested as the ability to learn how to learn more effectively.

Ensemble learning (EL) [35], [36] improves the accuracy and stability of predictions by combining multiple learning models. The basic idea behind this method is that individual models may have their limitations, but when the predictions of multiple models (such as decision trees, neural networks, etc.) are combined, the overall predictive performance is enhanced through diversity and complementarity. In cognitive learning, individuals may use multiple sources of information

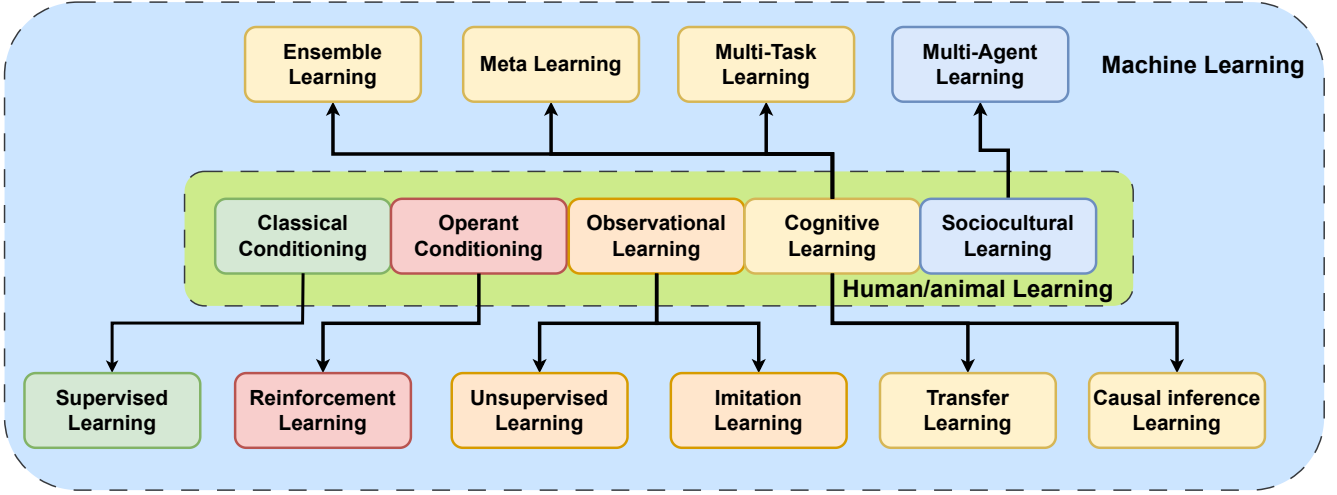


Fig. 4: Mapping of human/animal learning methods to machine learning methods

and strategies to enhance learning effectiveness. EL improves predictive performance by combining multiple models. Both can potentially utilize a diversified perspective and methods to optimize results.

Multitask learning [37], [38] enhances the model's generalization ability by learning multiple related tasks simultaneously during the training process. The core idea behind this method is that multiple tasks share some common representations or features so that while learning one task, the model can also learn useful information from other tasks. Both cognitive learning and multi-task learning may involve simultaneously dealing with multiple related tasks to improve efficiency and performance.

5) *Sociocultural Learning*: Sociocultural learning theory emphasizes that individuals learn and develop within the context of social interaction and cultural background. This type of learning naturally adapts to robot learning in MRS [39], [40], especially in decentralized systems where each robot can make independent decision. Multi-agent learning focuses on how multiple intelligent agents learn and interact in the same or interdependent environment. These agents learn the best strategies by observing the behavior of the environment and other agents to achieve their goals, which may involve cooperation or competition. Key issues in multi-agent learning include coordination, competition, communication, and the sharing of learning strategies. The similarity between sociocultural learning and multi-robot learning lies in the emphasis on interactions with other agents and the influence of the environment on learning.

6) *Affective Learning*: Affective learning in robots is a profound and complex research direction [41], but this article will not delve into it extensively. However, when there is a human in the loop in MRS [42], [43], the system may face issues related to recognizing, interpreting, processing, and simulating human emotions. These types of issues can be categorized under affective computing [44]. How to properly handle this information to improve the naturalness

and efficiency of human-computer interaction is a promising research direction.

B. Reinforcement Learning

The application of RL in MRS has been a research hotspot in the last decade. It is defined as a framework where each agent learns its behavior strategy through interaction with the environment to maximize certain cumulative rewards. Specifically, this learning process can be mathematically described as an extension of the Markov Decision Process (MDP), commonly known as the Multi-Agent Markov Decision Process (MAMDP) [45], [46]. A standard MAMDP can be defined as a tuple (S, A, P, R, γ) , where: S represents the state space, encompassing all possible environmental states. $A = A^1 \times A^2 \times \dots \times A^n$ represents the joint action space, with each A^i being the action space for agent i . $P : S \times A \times S \rightarrow [0, 1]$ is the state transition probability function, indicating the probability of transitioning to the next state given the current state and joint actions. $R : S \times A \times S \rightarrow \mathbb{R}$ is the reward function, which could be the sum of rewards for each agent or some other form of aggregation. γ is the discount factor used to calculate the present value of future rewards, with its value ranging from $0 \leq \gamma < 1$. In a multi-agent environment, the goal of each agent is to learn a policy $\pi^i : S \rightarrow A^i$, aiming to maximize its expected cumulative discounted reward. Unlike single-agent RL, in multi-agent RL, each agent must consider the impact of the behaviors of other agents on the environment and on its own rewards.

For an agent, the objective can be mathematically defined as maximizing the expected cumulative discounted reward $V^\pi(s)$ or $Q^\pi(s, a)$, where s represents the state and a represents the action. State-value function $V^\pi(s)$ represents the expected return under policy π in state s . It is defined as the sum of expected rewards for all possible paths, where each reward is multiplied by the power of the discount factor γ , indicating the proximity in time:

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s \right] \quad (1)$$

Action-value function $Q^\pi(s, a)$ represents the expected return of taking action a in state s , and then following policy π :

$$Q^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s, A_0 = a \right] \quad (2)$$

The goal of maximizing the reward function can be achieved by optimizing the aforementioned value functions. In a multi-agent environment, this goal is more complex because the optimal strategy of each agent may depend on the strategies of other agents. Therefore, agents need to learn a policy π^* that maximizes their expected cumulative discounted reward while considering the possible strategies of other agents. For single-agent environment, this can be simplified to finding an optimal policy π^* such that for all states s , $V^{\pi^*}(s) \geq V^\pi(s)$ for all π . Thus, the mathematical definition of maximizing rewards involves finding an optimal policy π^* such that: For all states s , $V^{\pi^*}(s) = \max_\pi V^\pi(s)$, or for all states s and actions a , $Q^{\pi^*}(s, a) = \max_\pi Q^\pi(s, a)$. This is usually achieved through iterative algorithms such as dynamic programming, the Monte Carlo method, Temporal Difference learning, or deep learning, which gradually approximates the optimal policy π^* .

The basic idea of RL is to learn through interaction, which means that compared to traditional model-based methods, RL does not require detailed prior knowledge about the environment. Robots can learn effective strategies through trial and error, which is particularly useful in unknown or uncertain environments. For example, in [47], [48], RL is used to solve the navigation problems of robot teams in complex dynamic environment. At the same time, RL can improve the adaptability and flexibility of MRS, such as helping aerial robot swarms deal with complex turbulent flows [47], or MRS tracking of moving targets [49]. RL can also handle multi-objective optimization problems, allowing each robot in an MRS to consider the overall system's optimal performance while pursuing individual goals. A continuous RL method introduced in [50] can adapt to multi-objective optimization functions to guide robots' movement in dynamic environment. In the previous MAMDP definition, it is assumed that each robot is fully observable, but in reality, partially observable situations are more common. Fortunately, MAMDP can be easily extended to MAPOMDP, and in [51], [52], [53], [54], all are based on the assumption of partial observability to apply RL to MRS. As a hot research topic in recent years, there are a rich set of references available for multi-robot reinforcement learning [40], [55], [56], [57], [58].

On the other hand, RL also has unique disadvantages [7], [52], [59], [60]. RL usually requires a large number of interaction samples to learn effective strategies, which may be impractical in real-world MRS due to the high

cost and time consumption of physical experiments. For this reason, [61] discusses various aspects of the simulation to reality (Sim-to-Real) transfer problem in robotics Deep Reinforcement Learning (DRL). In [62], an actor-critic algorithm combined with experience replay is introduced to improve sample utilization. By reusing past experiences, this method can learn effective strategies with fewer samples required. In complex multi-robot environment, ensuring the stability and convergence of learning algorithms is a challenge, especially in the case of continuous action spaces or multi-agent interactions. [63] introduces the Deep Deterministic Policy Gradient (DDPG) algorithm, a method that combines deep learning and reinforcement learning to solve control problems in continuous action space.

C. Imitation Learning

In the context of MRC, IL, as an effective learning strategy, aims to accelerate the learning process by observing and imitating the behavior of experts or other robots [64], [65], [66], [67]. While mathematical definition may vary depending on the specific method (such as Behavioral Cloning, Inverse RL, etc.) and the application scenario considered, a general mathematical framework can be stated: in the context of MRS where the followings are assumed, a state space S , an action space A , and a transition function $T : S \times A \rightarrow S$, representing the probability of system state transition under a given state and action. The goal is to learn a policy $\pi : S \rightarrow A$, that is, given the current state, to determine the action to be taken, in order to imitate the behavior of an expert (another robot or human). The expert's behavior is given by a set of trajectories $D = \{(s_1, a_1), (s_2, a_2), \dots, (s_N, a_N)\}$, where s_i represents the state and a_i represents the action taken by the expert in that state. The goal of IL is to minimize the difference between the learning policy and the expert policy. This can be quantified in different ways, for example, by minimizing the distance between the policy output action and the expert action, or by maximizing the similarity of the trajectories generated by the learning policy to the expert trajectories.

IL has clear advantages and disadvantages in MRS. A unique advantage is that by observing and imitating robots or humans who have effectively performed specific tasks, other robots can quickly learn new skills, reducing the time needed to learn from scratch through trial-and-error. In [66], a new active IL framework is proposed, where a teacher-student interaction model is utilized to significantly leverage the advantages of IL. Experiments on the MetaDrive benchmark and Atari 2600 games demonstrate that this method is more efficient in achieving performance close to that of experts compared to previous methods. In MRS, individuals can learn the importance of cooperation and cooperative strategies more quickly by observing the behavior of other cooperating robots, thereby promoting collaborative work throughout the group. In [68] it is pointed out that the emergence of cooperative behavior can be explained through understanding the co-evolution process of cooperators' core and betrayers' periphery, emphasizing the role of partner

selection and imitation strategies in promoting cooperative behaviors, without assuming the presence of underlying communication or reputation mechanism. In this way, the article provides a unified framework to study imitation-based cooperation in dynamic social networks.

The disadvantages of IL [69] are also obvious. They can be summarized into the followings: 1. Dependence on high-quality demonstrations: If the quality of expert demonstrations is not high, or if they do not match the current task environment, the learned strategy may lead to poor performance or even incorrect behavior. 2. Limitations in generalization capability: Since IL relies on specific demonstrations, the learned strategy may exhibit insufficient generalization capability when encountering unseen environment or task. 3. Lack of self-exploration: Over-reliance on imitation may lead to a lack of self-exploration and innovation capability in robots, preventing them from autonomously discovering solutions that are superior to the demonstrations. 4. Lack of diversity: In an MRS, if all robots imitate the same demonstration, it may lead to homogenization of behaviors, reducing the system's adaptability and robustness. With the development of IL, methods to address flawed demonstrations have been proposed in [70], [71]. A novel IL framework introduced in [72] expands the applicability of IL by incorporating the concepts of hindsight information embedding and contextual strategies, demonstrating superior performance across multiple tasks and settings.

D. Transfer Learning

In the context of MRS, the concept of TL typically involves transferring knowledge learned on one robot or a group of robots to other robots or robot groups, in order to improve learning efficiency, reduce the amount of training data needed, or enhance performance in new tasks [73], [74], [75], [76], [77]. While there are detailed mathematical models and definitions for specific applications, a general mathematical framework for TL may involve the following key elements: 1. Source Task T_S and Target Task T_T , which are defined through task-related data distributions, objective functions, etc. 2. Source Domain D_S and Target Domain D_T , each domain consists of a feature space and a marginal probability distribution (i.e., $P(X)$). In MRS, different robots may encounter different environment (domain). 3. Transfer Function f , which is the mapping from the source task to the target task. The purpose of this function is to enable the effective application of the knowledge learned on the source task to the target task. [78] introduces a framework for multi-robot TL from a dynamical system perspective.

From the perspective of multi-robot TL, TL can assist in the learning process of robots, reducing the training time and data needed to achieve excellent performance, while also promoting the sharing of knowledge and experience. This enables robots to learn not only from their own experiences but also from the successes and failures of other robots which is similar to IL. Especially under resource-constrained conditions (such as computing power, storage space, etc.), TL can reuse existing knowledge. As for shortcomings, if the source

task and target task are not sufficiently similar, or if the method of TL is not properly implemented, negative transfer may occur [79], [80]. This means the knowledge learned from other robots could actually decrease the performance of the robot on the target task. Moreover, to achieve effective knowledge transfer, communication and coordination among robots are necessary, which might increase the complexity and overhead of the system.

E. Causal Inference Learning

Causal inference learning refers to the process of using data to infer the causal relationships between variables. In the context of MRS, causal inference can help comprehend and predict the interactions between different robot behaviors and environmental factors [81], [82], [83], [84]. Although the concept of causal inference has a precise mathematical definition in statistics, its application in MRS remains an active research area, involving complex dynamic systems and interactions. In causal inference research, a core concept is the Potential Outcomes Framework, also known as the Rubin Causal Model (RCM). Additionally, methods based on Graphical Models play an important role in causal inference, especially in describing complex causal relationships between variables. Under the Potential Outcomes Framework, for each individual and each possible treatment (or intervention), there is a potential outcome. Causal effect is defined as the difference in potential outcomes under different interventions. Mathematically, if $Y_i(t)$ represents the potential outcome of individual i under intervention t , then the causal effect for individual i can be expressed as $Y_i(t_1) - Y_i(t_0)$, where t_1 and t_0 represent different intervention states. Graphical models use graphs (typically Directed Acyclic Graphs, DAGs) to represent the causal relationships between variables. In these models, nodes represent variables, and directed edges represent causal relationships. Through the analysis of the graph, one can identify conditional independency, causal pathway, and possible intervention effect. In MRS, CIL usually focuses on how to infer the causal relationships between robot behaviors based on observed data (e.g., the behavior of robots and changes in the environment) or the causal effects of robot behaviors on environmental changes. This includes understanding how the behavior of one robot might affect the behavior of other robots or the overall state of the system.

MRS typically involve multiple autonomous robots interacting in a shared environment to complete various tasks, such as collaborative transport, search and rescue, and automated monitoring. CIL has a unique advantage in explainable robot learning methods, as it can help one understand how the behavior of one robot affects the behavior of other robots or the overall state of the system. This is crucial for designing highly coordinated and efficient MRS. Furthermore, by identifying and understanding causal relationships, system designer can better formulate intervention measures (such as adjusting task assignments, communication strategies, etc.) to optimize the performance of the entire system. When robots clarify the various causal structures between themselves and

TABLE I: Analysis of literature on robot learning methods used in MRSs in the article

Ref	Methodology	Application	Topology	Joint Decision making	Limitations
[46]	Q-Learning	Non-specific	Decentralized	Based on the reputations of others, derived from past interactions.	May converge to non-efficient outcomes
[51]	DRQNs	Underwater vehicles	Decentralized	Distillation of single-task policies into a unified policy that performs well across multiple tasks, without explicit task identification during operation.	Complexity of training
[55]	Attention-Based DRL	Surveillance	Hybrid	By the DRL model	Complexity of training
[49]	CE-PG	Search and Rescue	Hybrid	Learned policy and their current state	Complexity of implementing the scheme
[56]	PG	Non-specific	Hybrid	Manipulated incentives and policies designed	Complexity of implementing the scheme
[52]	MA-DEC	Non-specific	De/Centralized	Structured observation framework	Scalability, complexity of implementing the scheme
[47]	GCNN, DRL	Aerial Operation	Decentralized	Sharing sensor measurements between nearby robots	Sim2Real gap
[85]	IRL, MCTS	Automated vehicles	Decentralized	Mimic human-like behavior in traffic	Complexity in learning effectively
[86]	IRL, transformer	Non-specific	Decentralized	Local observations, global communication inputs, and a learning-based stochastic tie-breaking strategy	Complexity in implementation scheme, scalability
[68]	EGT	Non-specific	Decentralized	Imitating the strategies of their neighbors in the network	Complexity of implementing the scheme
[78]	ODM	Quadrotor	De/Centralized	The emphasis is on the transfer learning process between individual robots rather than collective decision-making.	Poor generalization ability
[74]	DRL, HCP	Manipulation and Locomotion	De/Centralized	Optimizing individual policies based on the hardware characteristics of each robot.	Poor generalization ability
[75]	AC ILC	Trajectory Tracking	Centralized	Transferring learned control inputs for accurate trajectory tracking	Poor generalization ability
[87]	DDQL	Monitoring and Surveillance	Decentralizedshared	Experiences and a common learning coordinating implicitly through the distributed reinforcement learning framework to achieve collective objectives.	complexity of computing, scalability
[83]	PM	Search and Detection	Hybrid	Trust levels are made jointly through the aggregation of direct and indirect experiences.	Poor in limited communication
[82]	CI, RL	Football Strategy	Decentralized	Causal relationships between players and opponents	Sim2Real gap, complexity of implementing the scheme
[84]	DIPM	Search and Rescue	Decentralized	Conflicting goals are resolved through a single-bid auction mechanism among locally communicating robots	Complexity of implementing the scheme
[88]	MARL, RL	Non-specific	Decentralized	Reinforcement learning policies that evolve from individual experiences.	Complexity of implementing the scheme
[89]	DWPI	Non-specific	Decentralized	Inferring other agents' preferences	Poor generalization ability
[90]	DE-DRL	Air Traffic Control	Decentralized	Arbitrating between the decisions made by the local kernel-based RL model and the wider-reaching deep RL model, leveraging the strengths of both methods.	Complexity of implementing the scheme and training
[91]	e2e-CEL	Non-specific	De/Centralized	By learning to select and aggregate predictions from a subset of base learners.	Complexity of implementing the scheme and training
[92]	CA-PE	Overcooked Environment	Decentralized	Context-aware module that predicts the partner's policy level,	Complexity of implementing the scheme and training
[93]	NMEM	Monitoring	Decen/Hybrid	Combination of local collaboration signals and a macroscopic model	Scalability
[94]	SML	UAVs	Decentralized	Dynamic Stackelberg game	Complexity of implementing the scheme
[95]	MGRL	power grids, traffic tolling systems	Hybrid	Individual reward functions modified by incentives from the central planner,	Sim2Real gap, high computing source
[96]	ML	UAVs	Centralized	Enables an individual UAV to autonomously detect unsafe situations and replan its trajectory.	Sim2Real gap, Complexity of implementing the scheme
[97]	Dif-MAML	Monitoring	Decentralized	Achieve consensus on a common "launch model"	Poor in limited communication

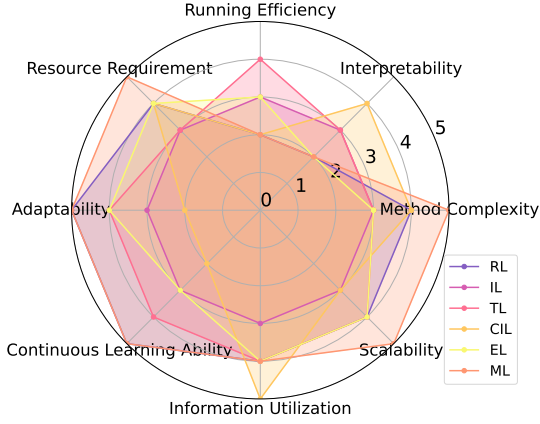


Fig. 5: Evaluation of six robot learning methods across eight dimensions

the environment, they can greatly reduce the dependency on large-scale data, which is significant in situation where data is scarce or the cost of data acquisition is high. On the other hand, the dynamism and interaction complexity of MRS make constructing accurate causal models very challenging. Therefore, the identification and verification of causal relationships require precise model design and complex data analysis [88], [98]. Although causal inference can reduce the dependency on large amounts of data, it still requires high-quality data to identify true causal relationships. Collecting and organizing such data in MRS is also very challenging. Conducting experiments in MRS (such as randomized controlled trials) to verify causal relationships may be impractical, especially in real-world applications, such as operating in unstable or uncontrollable environment [89], [99].

F. Ensemble Learning

In the context of MRC, Ensemble Learning (EL) can be seen as multiple robots (or agents) working together to improve the effect of overall task execution, where each robot can be considered as a base learner. The mathematical definition of EL usually relates to a specific ensemble method, such as Bagging, Boosting, or Stacking. Generally, the goal of EL is to combine the predictions of multiple models to reduce generalization error. Mathematically, EL can be defined as suppose there is a set of base learners $\{h_1, h_2, \dots, h_T\}$, each learner h_i can give an output $h_i(x)$ for a given input x . The goal of EL is to combine these predictions through a certain strategy (such as simple averaging, weighted averaging, or voting) to form the final ensemble prediction $H(x)$:

$$H(x) = f(h_1(x), h_2(x), \dots, h_T(x)) \quad (3)$$

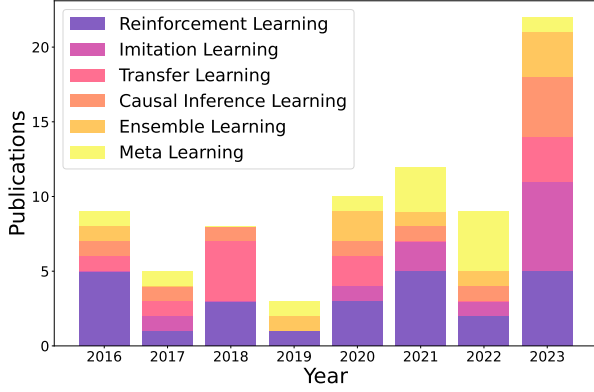
where f is the combining strategy, and T is the total number of base learners. In the context of MRC, the decision or prediction of each robot can be considered as the output of a base learner, and the ensemble method is used to coordinate the decisions among these robots, in order to improve the overall efficiency or accuracy of task execution.

In MRS, the greatest advantage of EL stems from its fundamental characteristic of improving prediction accuracy, robustness, and generalization ability by combining multiple models. By adjusting EL strategies, MRS can flexibly adapt to changes in task requirements or robot capabilities [92], [90], [100], [91], [93]. Supriyo Ghosh [90] combined the strengths of both kernel-based and deep multi-agent RL policies. This combination allows the system to leverage fine-grained local policies and more global policies efficiently, improving the decision-making process in air traffic control scenarios. [100] presents a decentralized EL approach that leverages sample exchange among multiple agents to improve performance in multi-agent systems. The method allows for decentralized data handling by having agents exchange data samples to enhance their collective predictive abilities. The benefits of using this ensemble method include increased accuracy through collaborative learning, competitive performance with state-of-the-art methods while maintaining data decentralization, and efficient utilization of local data resources by each agent. [91] introduces a novel framework for EL through differentiable model selection, integrating machine learning with combinatorial optimization. Despite the many advantages that EL offers in MRS, realizing these advantages also requires addressing a series of challenges, including how to effectively integrate data and decisions from different robots [101], [102].

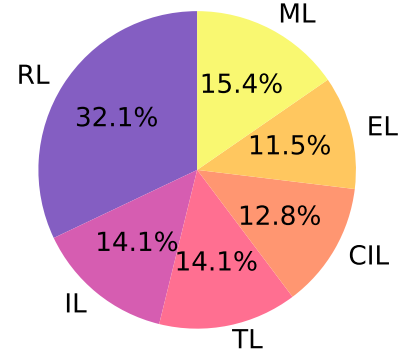
G. Meta Learning

Meta learning (ML), in the field of machine learning, refers to the process of building models that learn how to learn. It enables models to use past experiences to accelerate the learning process for new tasks, or to improve performance on new tasks [103], [94], [95], [104], [96]. Although the concept of ML is relatively intuitive, its mathematical definition may vary depending on different research background and application scenario. In the context of MRC, the goal of ML is to enable robots to quickly adapt to new cooperative tasks or environments drawing from previous cooperative experiences. In this scenario, ML often focuses on learning how to effectively share information, make decisions, and adapt to the behaviors of cooperative partners. A summary of all the methods mentioned above and their corresponding references can be found in Table I.

Mathematically, ML in MRS can be defined as finding a learning algorithm \mathcal{A} , which use the learning experiences from past tasks $\{T_1, T_2, \dots, T_n\}$ to improve learning efficiency and performance in a new task T_{new} . Specifically, consider a system containing multiple robots, each robot i has a parameter vector θ_i for task T , and the system's goal is to minimize a common loss function $L(T, \{\theta_i\})$, which measures the performance of the entire robot team in task T . The process of ML can be seen as finding an optimization algorithm \mathcal{A} , which adjusts the learning strategy of each robot so that the team can adapt more quickly to new tasks. This can be formalized by minimizing the expected loss over all tasks, as below:



(a)



(b)

Fig. 6: Statistics on the number of cited papers: a. citation counts of different robot learning method articles by year; b. Proportion of different robot learning methods

$$\min_{\mathcal{A}} \mathbb{E}_{T \sim \mathcal{T}} [L(T, \{\theta_i^*\})] \quad (4)$$

where $\{\theta_i^*\}$ is the set of robot parameter vectors obtained by applying algorithm \mathcal{A} to task T , and \mathcal{T} is the distribution of tasks.

Applying ML in MRS offers a series of unique advantages. For instance, MRS can quickly adjust their strategies to adapt to new environments or tasks through ML, reducing the exploration time in unknown environments [105], [106], [97]. Moreover, once a single robot learns a new skill or strategy, this knowledge can be rapidly disseminated throughout the entire robot group via ML mechanisms, improving the overall learning efficiency. This also facilitates the sharing of strategy and experience among robots, enabling the entire system to perform complex tasks cohesively. In summary, ML allows robots to learn from past experiences to adapt quickly to new task or environment, which is especially important for MRS, as these systems often need to operate collaboratively in dynamically changing environment. [107], [108]

IV. DISCUSSION

A. Technical Challenges

Technical challenges of robot learning in the context of MRS mainly come from three aspects: 1. The complexity of MRS themselves [5], [8], [52], [109], 2. The intrinsic complexity of specific robot learning methods [13], [69], and 3. Potential new problems that arise from the complexity resulted from merging the first two. [7], [59], [60].

MRS face numerous challenges, especially in complex environments such as underground, underwater, or remote areas, where communications can be severely limited [9], [110]. This includes issues like communication delay, signal attenuation, and data loss, which pose challenges to real-time coordination and control. Moreover, effectively allocating tasks and coordinating control to optimize use of resource,

improve efficiency of task execution, and adapt to dynamically changing environments remains a difficult problem [84], [111]. In terms of perception [112], compared to SRS, MRS face more difficult challenges in dealing with the heterogeneity, uncertainty, and temporality of different sensor data, as well as the challenge of processing large volumes of data in real-time with limited computing resources. In certain complex environments, robots in an MRS need to understand the intentions and behaviors of other robots to work effectively together. This requires not only advanced communication capability but also the ability to engage in complex social interaction and collaborative decision-making. A lot of works remain required to address other issues such as safety, reliability, and scalability still.

In the previous section on learning methods, the respective weakness and challenge of each robot learning method were discussed. Challenges that are common across these robot learning methods [59], [62] are considered in the present section. First is data and sample efficiency. Robot learning is often limited by the amount of available training data. Collecting a large volume of labeled data in the real-world is both expensive and time-consuming. Therefore, improving learning algorithms' data efficiency, such as through transfer learning, learning from simulation, and sample-efficient strategy in RL, is a significant challenge. Moreover, the generalization ability of learning methods is also an essential issue [72], [113]. Trained robot models need to perform well in unseen environment and situation. Enhancing generalization ability requires an algorithm not only to learn patterns in the training environment but also to adapt to new and dynamically changing environment. Real-time decision-making and control are essential [114], [115], as robots must be able to respond quickly to environmental changes and make decision and take action. This requires learning algorithms to achieve efficient data processing and decision-making with limited computing resources.

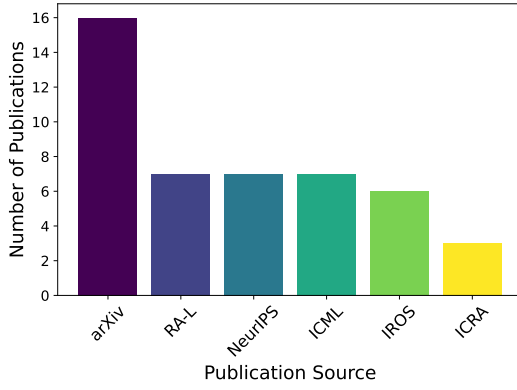


Fig. 7: Top six sources of cited literature

Lastly, whether these learning methods can autonomously learn and adapt is critical [42], [116], [117]. In complex and constantly changing environment, robots need to have the ability to learn and self-optimize in real-time. This demands the development of learning mechanisms that can automatically adapt to new task and environment without direct human intervention or disruption.

B. Applications

Applications of Robot Learning in MRS can be categorized based on the known extent of the environment into three types: structured environment, semi-structured environment, and unstructured environment, see figure3. This classification helps in understanding the challenges and requirements of robot learning and collaboration in different settings.

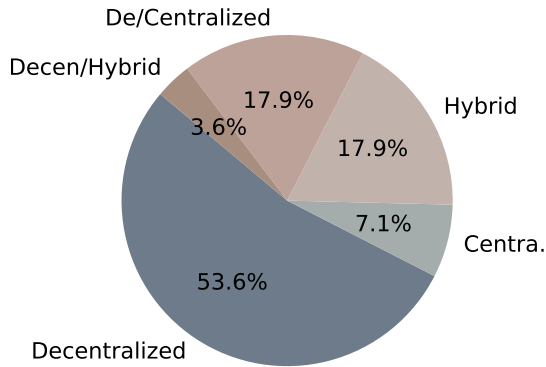


Fig. 8: Proportion of different communication topologies

In structured environment, layout and condition are relatively invariant and predictable. Robots can rely on accurate maps and predefined rules for operations. For example, in warehouse logistics such as cargo handling, sorting, and distribution [57], the environments are relatively unchanging, with shelf and aisle layouts being predetermined and known.

On manufacturing lines [118], where tasks and operational procedures are standardized, robots can learn specific assembly skill and collaboration pattern. Unstructured environments are highly dynamic, unpredictable, and lacking set rules and structures, demanding higher autonomy and learning capability from robots. For instance, in the aftermath of earthquake or flood where the environmental condition is highly complex and uncertain, robots would be required to autonomously explore and carry out rescue missions [112], [119]. Similarly, monitoring in natural settings such as forest, ocean, or polar region involves complex and variable environmental conditions. Exploring unknown or hazardous environments like deep sea, underground cave, or outer space, where the environment is entirely unknown, necessitates autonomous learning and collaboration by robots [120], [121]. Semi-structured environment lie between structured and unstructured environments, featuring some predictability but also variability and uncertainty, requiring robots to have a degree of adaptability and learning capability. A good example is automated agriculture [122], where the basic layout of a farmland is a constant, but factors such as crop growth condition, weed emergence, and weather condition introduce variability. Similarly, domestic and household service robots have to deal with a constant basic layout but face significant variability in the placement of daily items and the activities of family members [123]. Through the application of MRS in real-world scenarios, the specific requirements and challenges of Robot Learning in different environments must be defined and available to enable targeted design and optimization of learning strategies and collaboration mechanisms.

In the context of MRC, robot learning faces a series of unique challenges, reflecting issues from different dimensions such as distributed learning, collaborative learning, diversity in learning, to scalability of learning. The following is a discussion about these four aspects of challenges: In distributed learning [48], [84], [124], each robot may operate in different environments, collecting data with different characteristics and distribution needs. How to handle this data heterogeneity to achieve effective distributed learning is a problem. In collaborative learning [93], [112], robots in an MRS need to coordinate their actions to achieve a common goal. How to design learning algorithms to discover and implement efficient collaboration strategies to adapt to complex tasks and environments is a key challenge. In learning diversity [95], robot systems need to be able to adapt to dynamic changes in the environment. This requires learning algorithms to handle the uncertainty and variability of the environment, while maintaining adaptability to diversity. In terms of algorithm and system scalability [114], [125], [126], as the number of robots increases, how to maintain the scalability of learning algorithms and systems becomes a major challenge. Algorithms need to be able to efficiently process the data generated by a large number of robots, and the system design should support the addition of more robots without degrading performance.

C. Quantitative Analysis

In Figure 5, eight dimensions are used to gauge the applicability of various robot learning methods in the context of MRC: running efficiency, interpretability, method complexity, scalability, information utilization, continuous learning ability, adaptability, and resource requirement. Each dimension is rated on a scale of 1-to-5 to indicate its strength or weakness. RL and ML perform well in six of these dimensions while showing shortcomings in running efficiency and interpretability. Albeit having a balanced evaluation across all the eight dimensions, nevertheless, IL fails to excel in any of the specific areas. CIL has the best interpretability and information utilization among all the methods, however, it performs worst in running efficiency. TL's adaptability falls between RL and IL. EL shows deficiencies in interpretability and continuous learning ability compared to other dimensions.

In Figure 6a, the statistics on the year of publication of the articles cited in the paper and the corresponding number of robot learning articles are compiled. It's important to note that the statistics for 2016 include both 2016 and the preceding years, and the statistics for 2023 include both 2023 and 2024. From the figure, it's evident that focus is given to articles published in the last seven years, especially those published in 2020 and later, to better illustrate current research progress. In the pie chart seen in Figure 6b, the proportion of articles on the six most discussed robot learning methods in this text is calculated. It is evident that articles on RL register the largest number, accounting for 32.1% of the total, while the least is ensemble learning, which accounts for 11.5%. The numbers for other articles are relatively close. In Figure 7, the top six sources of the most applied articles in this text are listed. Arxiv preprints and conference papers dominate due to their quick publication cycles, which is closely related to the faster pace of research progress in robotics and machine learning compared to other fields in recent years. Figure 8 shows the proportion of communication topology structures in robot learning methods for MRS reviewed in the previous sections. Methods that are decentralized account for more than 50%, while hybrid methods and those supporting both decentralized and centralized approaches each account for 17.9%, with centralized methods being the least, at only 7.1%. This aligns with one of the assumptions mentioned earlier that individuals in an MRS can make decisions independently. In summary, the proportion of the articles cited in this paper exploring decentralized methods reaches 92.9%.

D. Research Trending

Robot learning in the context of MRC is facing a trend of rapid development and ongoing evolution. Existing methods have already achieved good results in many customized tasks, but there are still many challenges in generalized task requirement and under restricted condition. While predicting the future is difficult, nevertheless, a few key driving factors for robot learning in MRS from a realistic perspective are considered in the followings. Considering the balance

between generality and customization, large language models (such as the GPT series) offer powerful natural language processing capabilities, which can facilitate robots in understanding and executing more complex instructions. In the future, one can foresee these models being further customized to fit specific MRC scenarios while maintaining a degree of generality to flexibly handle various tasks. Moreover, as the capability of large language models in understanding and generating natural language continues to improve, interactions between robots as well as between robots and humans will become more natural and efficient. This will greatly enhance the collaborative efficiency of MRS in executing complex tasks. In practical complex tasks, a single robot learning method may not be sufficient to meet requirements, and combining different machine learning techniques can provide a more flexible and powerful learning mechanism. Especially in future's broader human-robot interactions, the greatly increased interpretability of robot learning brought by causal inference can help both robot systems and humans understand the causal relationships behind tasks more deeply, thereby making more reasonable and efficient joint decisions. As robot technology develops, how to effectively reduce energy consumption becomes an important topic. Optimizing algorithms and hardware design, as well as adopting more efficient learning methods, will be key to reducing the energy consumption of MRS.

V. REMARKS

This article performed a comprehensive survey on the wide range of mainstream methods and frameworks of robot learning within the context of MRS. Instead of directly discussing the classification and differences of learning methods from the perspective of machine learning, human and animal learning methods was first mapped to robot learning approaches. Especially in the current era of rapidly expanding intelligent robotics, the presented review provides a summary and blueprint for future research in robot learning, particularly in the context requiring collaborative learning. Moreover, considering the current trend of technology and development, robot systems are expected to be more widely applied in all aspects of human life. The frequency of interactions between robots and humans, animals, other robots, and the entire environment will increase. Promoting these interactions, robot learning technology is currently an important research topics with an implication to the foreseeable future.

REFERENCES

- [1] V. Sze, Y.-H. Chen, J. Emer, A. Suleiman, and Z. Zhang, "Hardware for machine learning: Challenges and opportunities," in *2017 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2017, pp. 1–8.
- [2] J. Peddie, *The History of the GPU-Eras and Environment*. Springer Nature, 2023.
- [3] A. Tyagi, S. Kukreja, M. N. Meghna, and A. K. Tyagi, "Machine learning: Past, present and future," *Neuroquantology*, vol. 20, no. 8, p. 4333, 2022.
- [4] T. Arai, E. Pagello, L. E. Parker, *et al.*, "Advances in multi-robot systems," *IEEE Transactions on robotics and automation*, vol. 18, no. 5, pp. 655–661, 2002.

- [5] Y. Rizk, M. Awad, and E. W. Tunstel, "Cooperative heterogeneous multi-robot systems: A survey," *ACM Computing Surveys (CSUR)*, vol. 52, no. 2, pp. 1–31, 2019.
- [6] R. Fierro, L. Chaimowicz, and V. Kumar, "Multi-robot cooperation," in *Autonomous Mobile Robots*. CRC Press, 2018, pp. 417–460.
- [7] S. Kapoor, "Multi-agent reinforcement learning: A report on challenges and approaches," *arXiv preprint arXiv:1807.09427*, 2018.
- [8] A. Rogers, S. Ramchurn, and N. Jennings, "Delivering the smart grid: Challenges for autonomous agents and multi-agent systems research," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 26, no. 1, 2012, pp. 2166–2172.
- [9] A. Gautam and S. Mohan, "A review of research in multi-robot systems," in *2012 IEEE 7th international conference on industrial and information systems (ICIIS)*. IEEE, 2012, pp. 1–5.
- [10] P. Berger and T. Luckmann, "The social construction of reality: A treatise in the sociology of knowledge anchor books: Usa," *Bogart, LM, Wagner, G., Galvan, FH, & Banks, D.(2010). Conspiracy beliefs about HIV are related to antiretroviral treatment non-adherence among African American men with HIV. Journal of acquired immune deficiency syndromes (1999)*, vol. 53, no. 5, p. 648, 1967.
- [11] J. Lave and E. Wenger, *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.
- [12] A. Bandura, "Social learning theory general learning press," *New York*, 1977.
- [13] O. Kroemer, S. Niekum, and G. Konidaris, "A review of robot learning for manipulation: Challenges, representations, and algorithms," *The Journal of Machine Learning Research*, vol. 22, no. 1, pp. 1395–1476, 2021.
- [14] J. H. Connell and S. Mahadevan, *Robot learning*. Springer Science & Business Media, 2012, vol. 233.
- [15] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," *Annual review of control, robotics, and autonomous systems*, vol. 3, pp. 297–330, 2020.
- [16] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychological review*, vol. 65, no. 6, p. 386, 1958.
- [17] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.
- [18] P. I. Pavlov, "Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex," *Annals of neurosciences*, vol. 17, no. 3, p. 136, 2010.
- [19] B. F. Skinner, *The behavior of organisms: An experimental analysis*. BF Skinner Foundation, 2019.
- [20] J. Piaget, M. Cook, et al., *The origins of intelligence in children*. International Universities Press New York, 1952, vol. 8, no. 5.
- [21] L. S. forme avant 2007 Vygotskij and V. John-Steiner, *Mind in society: The development of higher psychological processes*. Harvard University Press, 1979.
- [22] J. E. LeDoux, *The emotional brain: The mysterious underpinnings of emotional life*. Simon and Schuster, 1998.
- [23] C. Bishop, "Pattern recognition and machine learning," *Springer google schola*, vol. 2, pp. 5–43, 2006.
- [24] I. Muhammad and Z. Yan, "Supervised machine learning approaches: A survey," *ICTACT Journal on Soft Computing*, vol. 5, no. 3, 2015.
- [25] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.
- [26] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.
- [27] H. U. Dike, Y. Zhou, K. K. Deveerasetty, and Q. Wu, "Unsupervised learning based on artificial neural network: A review," in *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*. IEEE, 2018, pp. 322–327.
- [28] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 1.
- [29] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
- [30] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.
- [31] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, and A. Zhang, "A survey on causal inference," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 15, no. 5, pp. 1–46, 2021.
- [32] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.
- [33] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.
- [34] A. Salles, J. Ais, M. Semelman, M. Sigman, and C. I. Calero, "The metacognitive abilities of children and adults," *Cognitive Development*, vol. 40, pp. 101–110, 2016.
- [35] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000, pp. 1–15.
- [36] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Frontiers of Computer Science*, vol. 14, pp. 241–258, 2020.
- [37] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.
- [38] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," *Advances in neural information processing systems*, vol. 31, 2018.
- [39] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, pp. 1–49, 2022.
- [40] N. Anastassacos, J. García, S. Hailes, and M. Musolesi, "Cooperation and reputation dynamics with reinforcement learning," *arXiv preprint arXiv:2102.07523*, 2021.
- [41] O. S. Quick, "Empathizing and sympathizing with robots: implications for moral standing," *Frontiers in Robotics and AI*, vol. 8, p. 791527, 2022.
- [42] N. Churamani, S. Kalkan, and H. Gunes, "Continual learning for affective robotics: Why, what and how?" in *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2020, pp. 425–431.
- [43] B. Abendschein, C. Edwards, A. Edwards, V. Rijhwani, and J. Stahl, "Human-robot teaming configurations: A study of interpersonal communication perceptions and affective learning in higher education," *Journal of Communication Pedagogy*, vol. 4, pp. 123–132, 2021.
- [44] A. Appriou, A. Cichocki, and F. Lotte, "Modern machine-learning algorithms: for classifying cognitive and affective states from electroencephalography signals," *IEEE Systems, Man, and Cybernetics Magazine*, vol. 6, no. 3, pp. 29–38, 2020.
- [45] L. Busoni, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 38, no. 2, pp. 156–172, 2008.
- [46] L. Panait and S. Luke, "Cooperative multi-agent learning: The state of the art," *Autonomous agents and multi-agent systems*, vol. 11, pp. 387–434, 2005.
- [47] D. Patiño, S. Mayya, J. Calderon, K. Daniilidis, and D. Saldaña, "Learning to navigate in turbulent flows with aerial robot swarms: A cooperative deep reinforcement learning approach," *IEEE Robotics and Automation Letters*, 2023.
- [48] T. Fan, P. Long, W. Liu, and J. Pan, "Distributed multi-robot collision avoidance via deep reinforcement learning for navigation in complex scenarios," *The International Journal of Robotics Research*, vol. 39, no. 7, pp. 856–892, 2020.
- [49] H. Guo, Z. Liu, R. Shi, W.-Y. Yau, and D. Rus, "Cross-entropy regularized policy gradient for multirobot nonadversarial moving target search," *IEEE Transactions on Robotics*, 2023.
- [50] K. Zhang, S. McLeod, M. Lee, and J. Xiao, "Continuous reinforcement learning to adapt multi-objective optimization online for robot motion," *International Journal of Advanced Robotic Systems*, vol. 17, no. 2, p. 1729881420911491, 2020.
- [51] S. Omidshafiei, J. Papis, C. Amato, J. P. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *International Conference on Machine Learning*. PMLR, 2017, pp. 2681–2690.
- [52] D. Foster, D. J. Foster, N. Golowich, and A. Rakhlin, "On the complexity of multi-agent decision making: From learning in games to partial monitoring," in *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 2023, pp. 2678–2792.
- [53] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable mdps," in *2015 aai fall symposium series*, 2015.

- [54] H. Wu, A. Ghadami, A. E. Bayrak, J. M. Smereka, and B. I. Epureanu, "Impact of heterogeneity and risk aversion on task allocation in multi-agent teams," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7065–7072, 2021.
- [55] R. Wang, D. Zhao, and B.-C. Min, "Initial task allocation for multi-human multi-robot teams with attention-based deep reinforcement learning," *arXiv preprint arXiv:2303.02486*, 2023.
- [56] S. Nikkhoo, Z. Li, A. Samanta, Y. Li, and C. Liu, "Pimbot: Policy and incentive manipulation for multi-robot reinforcement learning in social dilemmas," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 5630–5636.
- [57] A. Agrawal, A. S. Bedi, and D. Manocha, "Rtaw: An attention inspired reinforcement learning method for multi-robot task allocation in warehouse environments," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1393–1399.
- [58] Z. Fan, N. Peng, M. Tian, and B. Fain, "Welfare and fairness in multi-objective reinforcement learning," *arXiv preprint arXiv:2212.01382*, 2022.
- [59] R. Nair, M. Tambe, M. Yokoo, D. Pynadath, and S. Marsella, "Taming decentralized pomdps: Towards efficient policy computation for multiagent settings," in *IJCAI*, vol. 3, 2003, pp. 705–711.
- [60] X. Zhang, Z. Liu, J. Liu, Z. Zhu, and S. Lu, "Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 825–18 838, 2021.
- [61] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2020, pp. 737–744.
- [62] Z. Wang, V. Bapst, N. Heess, V. Mnih, R. Munos, K. Kavukcuoglu, and N. de Freitas, "Sample efficient actor-critic with experience replay," *arXiv preprint arXiv:1611.01224*, 2016.
- [63] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.
- [64] H. M. Le, Y. Yue, P. Carr, and P. Lucey, "Coordinated multi-agent imitation learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1995–2003.
- [65] X. Wang, Y. Zhou, and W. Jin, "Distributed differentiable dynamic game for multi-robot coordination," 2023.
- [66] X.-H. Liu, F. Xu, X. Zhang, T. Liu, S. Jiang, R. Chen, Z. Zhang, and Y. Yu, "How to guide your learner: Imitation learning with active adaptive expert involvement," *arXiv preprint arXiv:2303.02073*, 2023.
- [67] T. Ablett, B. Chan, and J. Kelly, "Learning from guided play: Improving exploration for adversarial imitation learning with simple auxiliary tasks," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1263–1270, 2023.
- [68] J. Bara, P. Turrini, and G. Andrighetto, "Enabling imitation-based cooperation in dynamic social networks," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 2, p. 34, 2022.
- [69] N. Rajaraman, L. Yang, J. Jiao, and K. Ramchandran, "Toward the fundamental limits of imitation learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2914–2924, 2020.
- [70] Z. Li, T. Xu, Z. Qin, Y. Yu, and Z.-Q. Luo, "Imitation learning from imperfection: Theoretical justifications and algorithms," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [71] G.-H. Kim, S. Seo, J. Lee, W. Jeon, H. Hwang, H. Yang, and K.-E. Kim, "Demodice: Offline imitation learning with supplementary imperfect demonstrations," in *International Conference on Learning Representations*, 2021.
- [72] J. Liu, L. He, Y. Kang, Z. Zhuang, D. Wang, and H. Xu, "Ceil: Generalized contextual imitation learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [73] D. Schwab, Y. Zhu, and M. Veloso, "Zero shot transfer learning for robot soccer," in *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, 2018, pp. 2070–2072.
- [74] T. Chen, A. Murali, and A. Gupta, "Hardware conditioned policies for multi-robot transfer learning," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [75] K. Pereida, M. K. Helwa, and A. P. Schoellig, "Data-efficient multirobot, multitask transfer learning for trajectory tracking," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 1260–1267, 2018.
- [76] S. Chen, Q. Sun, H. You, T. Yang, and J. Hao, "Transfer learning based agent for automated negotiation," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023, pp. 2895–2898.
- [77] L. Smith, J. C. Kew, T. Li, L. Luu, X. B. Peng, S. Ha, J. Tan, and S. Levine, "Learning and adapting agile locomotion skills by transferring experience," *arXiv preprint arXiv:2304.09834*, 2023.
- [78] M. K. Helwa and A. P. Schoellig, "Multi-robot transfer learning: A dynamical system perspective," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 4702–4708.
- [79] G. Vrbancic and V. Podgorelec, "Transfer learning with adaptive fine-tuning," *IEEE Access*, vol. 8, pp. 196 197–196 211, 2020.
- [80] L. Gui, R. Xu, Q. Lu, J. Du, and Y. Zhou, "Negative transfer detection in transductive transfer learning," *International Journal of Machine Learning and Cybernetics*, vol. 9, pp. 185–197, 2018.
- [81] Y. Luo, J. Peng, and J. Ma, "When causal inference meets deep learning," *Nature Machine Intelligence*, vol. 2, no. 8, pp. 426–427, 2020.
- [82] S. Wang, Y. Pan, Z. Pu, B. Liu, and J. Yi, "Deconfounded opponent intention inference for football multi-player policy learning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 8054–8061.
- [83] Y. Guo, X. J. Yang, and C. Shi, "Enabling team of teams: A trust inference and propagation (tip) model in multi-human multi-robot teams," *arXiv preprint arXiv:2305.12614*, 2023.
- [84] A. J. Smith and G. A. Hollinger, "Distributed inference-based multi-robot exploration," *Autonomous Robots*, vol. 42, pp. 1651–1668, 2018.
- [85] K. Kurzer, M. Bitzer, and J. M. Zöllner, "Learning reward models for cooperative trajectory planning with inverse reinforcement learning and monte carlo tree search," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 22–28.
- [86] Y. Wang, B. Xiang, S. Huang, and G. Sartoretti, "Scrip: Scalable communication for reinforcement-and imitation-learning-based multi-agent pathfinding," *arXiv preprint arXiv:2303.00605*, 2023.
- [87] F. Venturini, F. Mason, F. Pase, F. Chiariotti, A. Testolin, A. Zanello, and M. Zorzi, "Distributed reinforcement learning for flexible uav swarm control with transfer learning capabilities," in *Proceedings of the 6th ACM workshop on micro aerial vehicle networks, systems, and applications*, 2020, pp. 1–6.
- [88] H. Wu, A. Ghadami, A. E. Bayrak, J. M. Smereka, and B. I. Epureanu, "Evaluating emergent coordination in multi-agent task allocation through causal inference and sub-team identification," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 728–735, 2022.
- [89] J. Lu, "Preference inference from demonstration in multi-objective multi-agent decision making," *arXiv preprint arXiv:2304.14126*, 2023.
- [90] S. Ghosh, S. Laguna, S. H. Lim, L. Wynter, and H. Poonawala, "A deep ensemble method for multi-agent reinforcement learning: A case study on air traffic control," in *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 31, 2021, pp. 468–476.
- [91] J. Kotary, V. Di Vito, and F. Fioretto, "End-to-end optimization and learning for multiagent ensembles," *arXiv preprint arXiv:2211.00251*, 2022.
- [92] X. Lou, J. Guo, J. Zhang, J. Wang, K. Huang, and Y. Du, "Pecan: Leveraging policy ensemble for context-aware zero-shot human-ai coordination," *arXiv preprint arXiv:2301.06387*, 2023.
- [93] V. Edwards, T. C. Silva, B. Mehta, J. Dhanoa, and M. A. Hsieh, "On collaborative robot teams for environmental monitoring: A macroscopic ensemble approach," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 11 148–11 153.
- [94] Y. Zhao and Q. Zhu, "Stackelberg meta-learning for strategic guidance in multi-robot trajectory planning," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 11 342–11 347.
- [95] J. Yang, E. Wang, R. Trivedi, T. Zhao, and H. Zha, "Adaptive incentive design with multi-agent meta-gradient reinforcement learning," *arXiv preprint arXiv:2112.10859*, 2021.
- [96] E. Yel, S. Gao, and N. Bezzo, "Meta-learning-based proactive online planning for uavs under degraded conditions," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 320–10 327, 2022.
- [97] M. Kayaalp, S. Vlaski, and A. H. Sayed, "Dif-maml: Decentralized

- multi-agent meta-learning,” *IEEE Open Journal of Signal Processing*, vol. 3, pp. 71–93, 2022.
- [98] J. Xu, K. Yin, J. M. Gregory, and L. Liu, “Causal inference for debiasing motion estimation from robotic observational data,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3008–3014.
- [99] Z. Wang, K. Mülling, M. P. Deisenroth, H. Ben Amor, D. Vogt, B. Schölkopf, and J. Peters, “Probabilistic movement modeling for intention inference in human–robot interaction,” *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 841–858, 2013.
- [100] Y. Yu, J. Deng, Y. Tang, J. Liu, and W. Chen, “Decentralized ensemble learning based on sample exchange among multiple agents,” in *Proceedings of the 2019 ACM International Symposium on Blockchain and Secure Critical Infrastructure*, 2019, pp. 57–66.
- [101] T. N. Rincy and R. Gupta, “Ensemble learning techniques and its efficiency in machine learning: A survey,” in *2nd international conference on data, engineering and applications (IDEA)*. IEEE, 2020, pp. 1–6.
- [102] A. Mohammed and R. Kora, “A comprehensive review on ensemble deep learning: Opportunities and challenges,” *Journal of King Saud University-Computer and Information Sciences*, 2023.
- [103] C. Finn, A. Rajeswaran, S. Kakade, and S. Levine, “Online meta-learning,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 1920–1930.
- [104] M. Schrum, M. J. Connolly, E. Cole, M. Ghetiya, R. Gross, and M. C. Gombolay, “Meta-active learning in probabilistically safe optimization,” *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 713–10 720, 2022.
- [105] F. G. Mohammadi, M. H. Amini, and H. R. Arabnia, “An introduction to advanced machine learning: Meta-learning algorithms, applications, and promises,” *Optimization, Learning, and Control for Interdependent Complex Networks*, pp. 129–144, 2020.
- [106] Y. Hu, M. Chen, W. Saad, H. V. Poor, and S. Cui, “Distributed multi-agent meta learning for trajectory design in wireless drone networks,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 3177–3192, 2021.
- [107] N. Saunshi, Y. Zhang, M. Khodak, and S. Arora, “A sample complexity separation between non-convex and convex meta-learning,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 8512–8521.
- [108] A. Gupta, M. Lanctot, and A. Lazaridou, “Dynamic population-based meta-learning for multi-agent communication with natural language,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 899–16 912, 2021.
- [109] Z. Yan, N. Jouandeau, and A. A. Cherif, “A survey and analysis of multi-robot coordination,” *International Journal of Advanced Robotic Systems*, vol. 10, no. 12, p. 399, 2013.
- [110] R. Doriya, S. Mishra, and S. Gupta, “A brief survey and analysis of multi-robot communication and coordination,” in *International conference on computing, communication & automation*. IEEE, 2015, pp. 1014–1021.
- [111] A. Khamis, A. Hussein, and A. Elmogy, “Multi-robot task allocation: A review of the state-of-the-art,” *Cooperative robots and sensor networks 2015*, pp. 31–51, 2015.
- [112] J. P. Queralta, J. Taipalmaa, B. C. Pullinen, V. K. Sarker, T. N. Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju, and T. Westerlund, “Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision,” *Ieee Access*, vol. 8, pp. 191 617–191 643, 2020.
- [113] K. Cobbe, O. Klimov, C. Hesse, T. Kim, and J. Schulman, “Quantifying generalization in reinforcement learning,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 1282–1289.
- [114] A. Bajcsy, S. L. Herbert, D. Fridovich-Keil, J. F. Fisac, S. Deglurkar, A. D. Dragan, and C. J. Tomlin, “A scalable framework for real-time multi-robot, multi-human collision avoidance,” in *2019 international conference on robotics and automation (ICRA)*. IEEE, 2019, pp. 936–943.
- [115] A. M. Derbas, K. M. Al-Aubidy, M. M. Ali, and A. W. Al-Mutairi, “Multi-robot system for real-time sensing and monitoring,” in *15th International Workshop on Research and Education in Mechatronics (REM)*. IEEE, 2014, pp. 1–6.
- [116] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, “Continual lifelong learning with neural networks: A review,” *Neural networks*, vol. 113, pp. 54–71, 2019.
- [117] K. Shaheen, M. A. Hanif, O. Hasan, and M. Shafique, “Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks,” *Journal of Intelligent & Robotic Systems*, vol. 105, no. 1, p. 9, 2022.
- [118] K.-C. Chen, S.-C. Lin, J.-H. Hsiao, C.-H. Liu, A. F. Molisch, and G. P. Fettweis, “Wireless networked multirobot systems in smart factories,” *Proceedings of the IEEE*, vol. 109, no. 4, pp. 468–494, 2020.
- [119] A. V. Nazarova and M. Zhai, “The application of multi-agent robotic systems for earthquake rescue,” *Robotics: Industry 4.0 Issues & New Intelligent Control Paradigms*, pp. 133–146, 2020.
- [120] E. Olcay, F. Schuhmann, and B. Lohmann, “Collective navigation of a multi-robot system in an unknown environment,” *Robotics and Autonomous Systems*, vol. 132, p. 103604, 2020.
- [121] Y. Bai, K. Asami, M. Svinin, and E. Magid, “Cooperative multi-robot control for monitoring an expanding flood area,” in *2020 17th International Conference on Ubiquitous Robots (UR)*. IEEE, 2020, pp. 500–505.
- [122] C. Lytridis, V. G. Kaburlasos, T. Pachidis, M. Manios, E. Vrochidou, T. Kalampokas, and S. Chatzistamatis, “An overview of cooperative robotics in agriculture,” *Agronomy*, vol. 11, no. 9, p. 1818, 2021.
- [123] A. Abou Allaban, M. Wang, and T. Padir, “A systematic review of robotics research in support of in-home care for older adults,” *Information*, vol. 11, no. 2, p. 75, 2020.
- [124] J. Choi, S. Oh, and R. Horowitz, “Distributed learning and cooperative control for multi-agent systems,” *Automatica*, vol. 45, no. 12, pp. 2802–2814, 2009.
- [125] O. F. Rana and K. Stout, “What is scalability in multi-agent systems?” in *Proceedings of the fourth international conference on Autonomous agents*, 2000, pp. 56–63.
- [126] C. D. Hsu, H. Jeong, G. J. Pappas, and P. Chaudhari, “Scalable reinforcement learning policies for multi-agent control,” in *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2021, pp. 4785–4791.