Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Dynamic joint resource allocation in maritime wireless communication networks: a meta-reinforcement learning approach based on knowledge embedding

Zhongyang MAO[†§1,2], Zhilin ZHANG[†‡§1,2], Faping LU[1,2], Xiguo LIU[1,2], Zhichao XU[1,2], Yaozong PAN[1,2], Jiafang KANG[1,2], Yang YOU[3]

[1]*Naval Aviation University, Yantai 264001, China*

[2]*Shandong Key Laboratory of Sea and Air Information Perception and Processing Technology, Yantai 264001, China*

[3]*PLA 91001 Unit, Beijing 100000, China*

[†]E-mail: freedom_mzy@163.com; zzl19970811@163.com

Received Jan. 3, 2025; Revision accepted May 28, 2025; Crosschecked Dec. 12, 2025

Abstract: As human exploration of the ocean expands, the demand for continuous, high-quality, and ubiquitous maritime communication is steadily increasing. However, the dynamic nature of the marine environment and resource constraints present significant challenges for traditional heuristic resource allocation methods, complicating the balance between high-quality communication and limited network resources. This results in suboptimal system throughput and an over-reliance on specific problem structures. To address these issues, in this paper, we introduce a joint resource allocation method based on knowledge embedding. The proposed approach includes an action distribution alignment module designed to improve resource utilization by preventing unreasonable action-output combinations. Furthermore, by integrating knowledge embedding with meta-reinforcement learning techniques, a physical guidance loss function is formulated, which effectively reduces the sample size required for model training, thereby enhancing the algorithm's generalization capabilities. Simulation results show that the proposed method achieves an increase in average system throughput of 31.19% compared to the model-agnostic meta-learning proximal policy optimization (MAML-PPO) algorithm and 80.91% compared to the RL^2 algorithm, across various channel environments.

Key words: Marine wireless communication; Resource allocation; Knowledge embedding; Meta-reinforcement learning
https://doi.org/10.1631/FITEE.2500007 **CLC number:** TN92

1 Introduction

With the increasing exploration of the ocean, the demand for high-quality wireless communication with full-time, all-encompassing coverage at sea is growing rapidly. This has become a core objective for the development of next-generation communication technologies, such as Beyond 5G (B5G) and 6G. Achieving this vision necessitates the integration of efficient transmission methods, wide-area coverage, and heterogeneous resource orchestration within maritime wireless communication networks. Such integration is crucial for delivering ubiquitous, intelligent information services that seamlessly combine communication, computing, and perception (Yin et al., 2023) to support complex tasks across space, air, sea, and specific regions.

However, the marine environment presents significant challenges in densely deploying and supporting high-power base stations. Mobile platforms such as ships, large drones, and aircraft are increasingly being

‡ Corresponding author
§ These two authors contributed equally to this work
ORCID: Zhongyang MAO, https://orcid.org/0000-0001-6279-1627; Zhilin ZHANG, https://orcid.org/0009-0006-1442-3735

used to overcome these limitations. While offering flexibility, these platforms also lead to scarce wireless network resources and limited service capabilities. Furthermore, the complex and dynamic nature of ocean channel environments, along with the constantly changing topology of network nodes, makes it difficult for any static resource allocation scheme to effectively address the communication needs of maritime networks. This highlights the urgent need for flexible wireless network resource allocation schemes capable of adapting to evolving conditions.

Human maritime activities often rely on various types of nodes, including ships, drones, aircraft, submarines, buoys, and satellites, to achieve high-quality communication over wide areas. This necessitates the unified management and allocation of heterogeneous resources. However, the vastness of the ocean, multiplicity of network nodes, and diversity of services—each with significantly varying resource demands—pose considerable challenges for efficient resource allocation. As a result, resource allocation in these settings requires high real-time performance, wide coverage, and the ability to handle complex and dynamic conditions. Existing time–frequency resource allocation technologies, which rely on fixed frequency bands and subcarriers, struggle to meet the evolving demands of maritime communication. Therefore, there is an urgent need for more flexible and scalable resource allocation technologies capable of accommodating the growing complexity and capacity requirements of maritime wireless communication networks.

Currently, orthogonal frequency division multiple access (OFDMA) (Yuan et al., 2023; Ferreira et al., 2024; Jha et al., 2024; Yan et al., 2024; Jin et al., 2025a) is one of the core wireless resource allocation technologies in land-based wireless transmission systems, offering flexible spectrum allocation and high spectral efficiency. The OFDMA framework provides a structured approach to resource allocation, supporting dynamic frequency selection, enabling the division of resource blocks across multiple subcarriers (Bossy et al., 2022), and allowing the management of modulation schemes and power levels for each subcarrier (Li et al., 2022). However, the solution space for allocation schemes is vast. Additionally, power and spectrum allocation in wireless communication networks is a typical NP-hard, non-convex problem, presenting significant challenges for traditional solution methods. As a result, resource allocation strategies in OFDMA-based wireless networks often rely on heuristic techniques (Gautam et al., 2019; Yang LW et al., 2022; Švedek et al., 2023; Wang T et al., 2023) or deep learning methods (Le et al., 2019; Tseng et al., 2023; Zhang et al., 2023; Tan et al., 2024). For example, Gautam et al. (2019) investigated relay selection and power allocation for simultaneous wireless information and power transfer (SWIPT) in multi-user OFDMA systems. Their proposed method optimized the power allocation ratio and relay assignment, significantly improving throughput and energy efficiency. Similarly, Yang LW et al. (2022) developed a power allocation strategy for macro/micro cellular heterogeneous networks using the heuristic bat algorithm, which enhances convergence accuracy and energy efficiency optimization. However, this approach may encounter challenges related to increased computational complexity when applied to larger networks or more complex interference models.

To tackle the energy efficiency resource allocation problem for OFDMA heterogeneous networks (HetNets), Le et al. (2019) applied the successive convex approximation method to approximate the optimal solution for resource block and power allocation, effectively addressing non-convex optimization challenges. Their proposed method satisfies both quality of service (QoS) and fairness constraints. Similarly, Tseng et al. (2023) used deep learning to address video transmission resource management in OFDMA non-orthogonal multiple access (NOMA) systems. By incorporating an additional penalty term into the loss function, they achieved improvements in average capacity and reductions in non-compliant resource allocations.

Thus, OFDMA is a promising technology for achieving seamless communication with wide-area coverage at sea. Experts and scholars have already embarked on exploring its applications across diverse environments, including land, sea, and air (Liu et al., 2021; Su et al., 2021; Han et al., 2022; Kim et al., 2023; Hu et al., 2024; Wang T and You, 2024; Wang XH et al., 2024; Yang SD et al., 2024). For instance, Hu et al. (2024) introduced a flexible aggregated federated learning approach grounded in OFDMA within an onshore wireless federated learning scenario.

This innovation addresses the challenge of optimizing client selection, subchannel allocation, and modulation techniques in resource-constrained wireless networks, thereby enhancing the convergence speed and accuracy of federated learning. Additionally, Kim et al. (2023) proposed a deep-learning-based spectrum sensing solution within an underwater acoustic cognitive radio network scenario that leverages OFDMA. This solution tackles the issue of improving spectrum sensing accuracy within limited sensing times and is well-suited for underwater equipment, albeit with relatively high computational complexity. The marine wireless communication environment exhibits distinct characteristics compared to terrestrial environments: marine nodes move at high speeds (Wang LY et al., 2024), are widely distributed, and exhibit significant diversity in types (Ning et al., 2023), accompanied by diverse service demands. Consequently, node energy consumption is constrained, communication channels show rapid time variability (Meister et al., 2024), and there is a high demand for real-time allocation strategies. Traditional heuristic methods often yield locally optimal solutions, struggling to adapt to dynamic or evolving scenarios. Meanwhile, pure deep reinforcement learning (DRL) methods are typically tailored for single tasks, requiring substantial training data and suffering from low sample efficiency. Therefore, the direct application of existing terrestrial OFDMA heuristic or deep learning resource allocation techniques (Jin et al., 2025b) to marine scenarios poses significant challenges.

Meta-learning, particularly in the context of reinforcement learning, has the potential to overcome the limitations of static and heuristic methods in such dynamic environments. The introduction of meta-reinforcement learning provides a novel perspective for addressing these issues. Several studies have explored the integration of communication resource allocation techniques with meta-learning approaches to address the challenges inherent in dynamic maritime communication networks. For instance, Letchford et al. (2020) introduced a meta-learning approach that balances fairness and spectral efficiency, addressing the resource allocation challenge in overloaded OFDMA systems, particularly in high-demand scenarios. Tefera et al. (2023) proposed a DRL-assisted optimization model for downlink OFDMA systems. By leveraging DRL, their method achieves superior throughput and signal-to-interference-plus-noise ratio (SINR) compared to conventional optimization techniques.

Hou et al. (2023) at Zhejiang University, China, combined the model-independent meta-learning algorithm and unsupervised learning mechanisms to offer a novel solution for rapidly and continuously optimizing resource allocation under variable channel state information distribution. Chen et al. (2022) introduced a cache-assisted collaborative task offloading and resource allocation strategy grounded in meta-reinforcement learning within the multi-access edge computing context. This strategy addresses the issue of resource wastage due to redundant task computation and transmission in Internet of Things (IoT) applications, significantly enhancing the quality of experience (QoE) for users. However, it encounters constraints in managing multi-user dynamic resource allocation and multi-user interference. Dhuheir et al. (2024) presented an unmanned aerial vehicle (UAV)-assisted energy harvesting framework using meta-reinforcement learning within the context of wireless energy harvesting for IoT devices in disaster-stricken areas. This framework addresses the challenge of maximizing energy harvesting for IoT devices under resource constraints, thereby improving energy harvesting efficiency and coverage, although it has a relatively high computational complexity.

In addition, Sun et al. (2022) presented a transfer learning framework based on multi-agent deep Q-networks (MADQNs), aimed at maximizing the aggregate rate for all users through dynamic allocation strategy adjustments. Wang XM et al. (2022) integrated knowledge distillation into the multi-agent algorithm transfer learning (TL)-MADQN, enhancing convergence speed and data rates. This approach addresses the challenge of integrating subcarrier and power allocation in a 5G multi-cell multiple input single output (MISO)-OFDMA system, effectively meeting substantial access demands and high data rate requirements. Meanwhile, Hou et al. (2023) combined the model-agnostic meta-learning (MAML) algorithm with an unsupervised learning mechanism, providing a novel method for optimizing resource allocation in environments with fluctuating channel state information (CSI). Shi et al. (2025) proposed a user-level configuration

adaptive framework based on meta-reinforcement learning in the scenario of real-time edge video analysis. This framework addresses the challenge of optimizing QoE under dynamic network and video content conditions. However, the efficacy of the proposed method depends heavily on the quantity and quality of the learning data.

While existing meta-DRL-based dynamic resource allocation methods have made notable progress, they still face significant challenges: (1) Meta-DRL approaches often involve high complexity when multiple algorithms are combined to generate output strategies, and their performance can be limited when allocating joint resources through combined output actions. (2) Existing meta-DRL methods typically require extensive pre-training on large datasets to achieve optimal results, which makes them difficult to apply in scenarios where data collection is limited. As a result, improving the efficiency of meta-DRL methods and reducing their dependency on extensive data remain pressing challenges.

To address these challenges, in this paper, we integrate meta-reinforcement learning with the OFDMA framework to optimize information energy efficiency per unit power for autonomous decision-making in offshore base stations. We propose a joint resource allocation method based on knowledge-embedding meta-reinforcement learning. This method improves the meta-DRL approach's ability to generate action combination strategies for joint resource allocation by aligning the distribution of multiple agent actions through dynamic transfer mapping. By incorporating knowledge embedding, we design a domain-knowledge-based physical guidance loss function to guide the meta-DRL model in allocating power and spectrum in accordance with the physical world's known rules, thereby reducing the model's dependency on large amounts of data. The proposed method aims to address dynamic resource demands in maritime wireless communication networks, offering a robust and scalable solution. The main contributions of this paper are as follows:

1. We propose a time–frequency resource management model and an objective optimization function for the OFDMA system in maritime wireless networks. Focusing on enhancing time–frequency resource utilization, a time–frequency resource block allocation model is developed to jointly optimize the bandwidth, power, and spectrum resources of nodes.

2. We propose a joint resource allocation method, knowledge-embedding model-agnostic meta-learning (KE-MAML), based on meta-reinforcement learning and knowledge embedding. This method consists of two networks: an outer meta-learning network and an inner DRL network. By incorporating knowledge embedding, a domain-specific physical guidance loss function is designed for the inner network. This function uses known domain knowledge as a soft constraint, generating a physical guidance term that steers the model's optimization and adjustment processes. This approach reduces the data dependency during model training and enhances the model's generalization capabilities.

3. We introduce a universal action distribution alignment module. This module uses various distribution mapping techniques based on the level of conflict within action combinations, enabling agents to generate actions that adhere to real-world constraints. By reducing decision conflicts when agents output action combinations, the module minimizes resource loss and strategic inconsistencies.

2 System model

In this paper, we explore a maritime wireless communication network scenario, which integrates air, space, and ground communication networks to provide diverse services for maritime mobile nodes (Fig. 1). Given the challenges of constructing large-scale communication infrastructure in the marine environment, maritime mobile nodes must assume additional

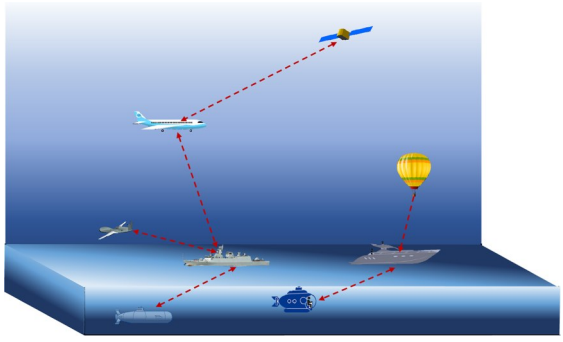

Fig. 1 Maritime wireless network architecture

roles to maintain the wireless communication network. For instance, these nodes can function not only as users but also as temporary base stations or relay nodes, supporting specific communication tasks.

For analytical simplicity, while maintaining generality, we abstract the nodes in our model as network nodes, categorizing them into low-speed nodes (e.g., ships, small drones) and high-speed nodes (e.g., medium to large drones, aircraft), based on their movement rates (Mao et al., 2024). These nodes move continuously and randomly at varying speeds across the sea area.

2.1 System structure

We consider a downlink maritime wireless communication orthogonal frequency division multiple access (MWC-OFDMA) system, where the central node (CN) receives data from terrestrial base stations or satellites and subsequently distributes them (e.g., commands) to multiple maritime mobile nodes (MNs) in OFDMA mode (Fig. 2). Given the large distances between mobile nodes, mutual interference among them is neglected.

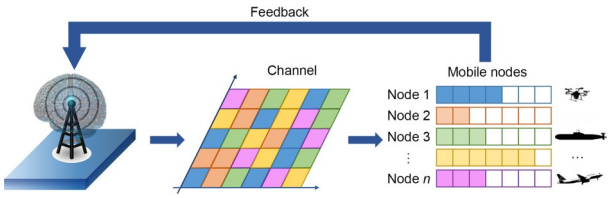

Fig. 2 Schematic of intelligent MWC-OFDMA scheduling

We assume that when a mobile node enters the CN's communication range, it can immediately establish a connection with the CN, and when it exits the communication range, the connection is promptly terminated. The network consists of one CN and n submarine node stations (SNS), forming the downlink of the OFDMA subsystem. The total bandwidth B is divided into K subcarrier resource blocks, each with a bandwidth of $f=B/K$. Let $\mathcal{N}=\{1, 2, \cdots, n\}$ and $\mathcal{K}=\{1, 2, \cdots, K\}$ denote the sets of SNS and subcarriers, respectively.

$$f_k = f_c \pm \left(k_s - \frac{K}{2} - 1\right)\Delta f, \tag{1}$$

where f_k is the frequency corresponding to the k_s^{th} subcarrier, f_c represents the center frequency, and Δf is the center frequency interval of subcarriers.

If node i is allocated an output power of P and n_i subcarrier resource blocks, the output power allocated to each block is P/n_i.

2.2 Optimization scheme for system resource allocation

Based on the channel transmission model in free space, we derive the total throughput model for the downlink from the base station to the mobile node. Next, by incorporating the unique characteristics of the marine environment and substituting the relevant environmental parameters, we derive the maximum energy efficiency resource allocation optimization scheme for the maritime wireless transmission environment, which is obtained by maximizing the total throughput index.

The logarithmic path loss model is used to calculate the transmission loss during maritime wireless communication.

$$\mathrm{PL}(d) = \mathrm{PL}(d_0) + 10n_{\mathrm{Loss}} \lg\left(\frac{d}{d_0}\right), \tag{2}$$

where d represents the distance between the node and the base station, d_0 is the reference distance, and n_{Loss} represents the path loss index, which is 2 in free space.

The relationship between the signal reception power P_r, transmission power P_t, path loss PL(d), fast fading characteristics, and channel gain is expressed as follows:

$$P_r(d) = \frac{P_t}{\mathrm{PL}(d)} \cdot \left|h\left(f_{d,i}, v_i\right)\right|^2 \cdot G\left(f_{d,i}\right), \tag{3}$$

where $h(f_{d,i}, v_i)$ is the fast fading characteristic and $G(f_{d,i})$ is the channel gain.

By substituting Eq. (3) into Shannon's capacity equation, the channel capacity C of a single node connected to the base station is obtained:

$$C = B\log_2\left(1 + \frac{\dfrac{P_t}{\mathrm{PL}(d)} \cdot \left|h\left(f_{d,i}, v_i\right)\right|^2 \cdot G\left(f_{d,i}\right)}{N_0 B + I}\right), \tag{4}$$

where N_0 is the noise power spectral density and I is the interference power spectral density.

The total throughput of the base station is then calculated as follows:

$$\begin{aligned}\max \Gamma &= \max \sum_{i=1}^{n_k} C_i \\ &= \max \sum_{i=1}^{n_k} B_i \\ &\quad \cdot \log_2\left(1+\frac{\dfrac{P_{t,i}}{\mathrm{PL}(d_i)}\cdot\left|h\left(f_{d,i},v_i\right)\right|^2\cdot G\left(f_{d,i}\right)}{N_0 B_i + I_i}\right).\end{aligned} \tag{5}$$

Given that the maximum bandwidth and output power of the base station remain constant, the OFDMA system divides the entire spectrum into n_k resource blocks, allocating them all to each time slot. The available bandwidth for each node is determined based on the number of allocated resource blocks. Thus, the problem of maximizing throughput is transformed into an optimal allocation problem for power and spectrum.

For maritime communication, the commonly used air–sea channel path loss model is based on the classical logarithmic path loss formula, with adjustments made for marine propagation environments and sea wave conditions using correction factors. The channel noise model is referenced from the International Telecommunication Union (ITU) radio noise (ITU, 2016), and the channel gain model was described by Wang J et al. (2018), as follows:

$$G_i\left(t,f_{c,i}\right) = L(\gamma,t) + 10 n_{\mathrm{Loss}} \lg\left(\frac{d_t}{d_{i,k,t}}\right) + \chi_\sigma^{i,k,t} + \varsigma F_t, \tag{6}$$

where n_{Loss} represents the path loss index, typically set to 1.1 (Wang J et al., 2018) due to the sea waveguide effect. $\chi_\sigma^{i,k,t}$ denotes shadow fading, which increases under more adverse sea conditions. To better account for the rapid movement of nodes, an adjustment parameter F_t is introduced, and ς is set to −1 when the node is far from the shore base and 1 otherwise. $d_{i,u,t}$ signifies the reference distance from node i to base station u, calculated as follows:

$$d_{i,u,t} = \sqrt{\left(x_{i,t}-x_{u,t}\right)^2+\left(y_{i,t}-y_{u,t}\right)^2+\left(z_{i,t}-z_{u,t}\right)^2}, \tag{7}$$

where $x_{i,t}$, $y_{i,t}$, and $z_{i,t}$ are the coordinates of node i along the x-axis, y-axis, and z-axis, respectively, and $x_{u,t}$, $y_{u,t}$, and $z_{u,t}$ are the coordinates of base station u along the x-axis, y-axis, and z-axis, respectively.

Given the significant distance between the nodes and the base station, a simplified three-ray path loss model is used, with the formula detailed by Wang J et al. (2018). The model includes the following components:

$$L\left(\gamma,t\right) = 10\lg\left(\left(\frac{d_t}{d_{i,k,t}}\right)^{\gamma}\left(P_{\mathrm{Los}}+P_{\mathrm{R1}}+P_{\mathrm{R2}}\right)\right), \tag{8}$$

where P_{Los}, P_{R1}, and P_{R2} represent the attenuation of the main path and the two reflection paths, and γ is the channel environment adjustment parameter, where a higher value indicates a poorer channel environment.

To ensure "relative fairness" in resource allocation among different users and prevent the allocation of excessive resources to some users, while leaving others underserved, the allocation for each user is constrained by minimum bandwidth and power limits. Ultimately, the minimum data transmission rate is used to meet user service quality demands, ensuring that all users can perform their tasks effectively.

In summary, the power–spectrum joint optimization constraint function aimed at maximizing the transmission rate is expressed as follows:

$$\begin{cases}\max\limits_{P_{t,i},B_i} \sum\limits_{i=1}^{n} B_i \log_2\left(1+\dfrac{\dfrac{P_{t,i}}{\mathrm{PL}(d_i)}\cdot\left|h\left(f_{d,i},v_i\right)\right|^2\cdot G_i\left(t,f_{d,i}\right)}{N_0 B_i+I_i}\right), \\ \text{s.t.}\quad \mathrm{C1}: B_{\min}\leqslant B_i\leqslant B_{\max}, \\ \qquad \mathrm{C2}: P_{\min}\leqslant P_{t,i}\leqslant P_{\max}, \\ \qquad \mathrm{C3}: C_i\geqslant C_{\min,i},\end{cases} \tag{9}$$

where $C_{\min,i}$ is the minimum data transmission rate of the i^{th} node, $B_{\min}$ is the minimum bandwidth, $B_{\max}$ is the maximum bandwidth, $P_{\min}$ is the minimum power, and $P_{\max}$ is the maximum power.

3 Meta-reinforcement learning method for joint resource allocation based on knowledge embedding

Given the resource constraints and complex dynamics of maritime wireless communication networks, it is impractical for agents to learn all environmental changes solely through training. Therefore, the generalization ability of agents is particularly crucial in such complex and dynamic environments. DRL benefits from the strong fitting capabilities of deep learning but is inherently limited by its data dependency. When an agent encounters an unfamiliar environment where obtaining a large amount of observation data is challenging, it must rely on in-depth analysis of limited data and empirical knowledge to adapt to the environment.

To conserve resources, we consider having the agent generate multiple resource allocation plans simultaneously. However, ensuring that these plans align with real-world constraints and logic can be difficult.

In response to these challenges, in this section, we introduce a meta-reinforcement learning approach for joint resource allocation, termed KE-MAML. This method is framed from both the loss function and action output perspectives. It comprises two main components: an outer component, which is a model network based on meta-learning, and an inner component, which is a strategy network derived from DRL.

3.1 Basic principles

The proposed method incorporates two loops—an inner loop and an outer loop—to optimize both the inner and outer model networks. The outer model network generates random task sets and initial strategies, which are then passed to the inner loop for learning and optimization. In the inner loop, the action distribution alignment module (DAM) is used to optimize strategies for each task. The module interacts with the environment to collect rewards, calculate task-specific losses, and iteratively update the initial strategy until the task concludes. Finally, the task-specific losses are backpropagated to the outer loop. After aggregating the outputs from the inner loop, the outer loop adjusts the global policy parameters in the model based on the feedback gradient, thereby improving the generalization and universality of the strategy. A structural block diagram of this method is shown in Fig. 3.

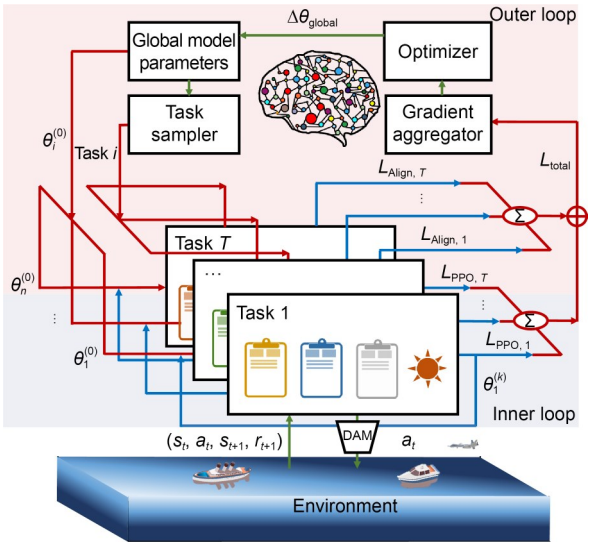

Fig. 3 Dynamic joint resource allocation method for maritime wireless communication based on knowledge embedding (DAM: distribution alignment module)

The outer loop optimizes the model's global initialization parameters based on the losses provided by the inner loop.

$$\theta^{(0)} \leftarrow \theta^{(0)} - \beta\nabla_{\theta^{(0)}}\sum_{i=1}^{m} L_{T_i}\left(\theta_i^{(k)}; D_i\right), \tag{10}$$

where β is the meta-learning rate, D_i represents the interaction experience of the i^{th} task with the environment, and L_{T_i} denotes the loss function for the i^{th} task.

The inner loop uses the initial strategy supplied by the outer loop to generate action combinations for different tasks via the DAM. It then interacts with the environment, accumulates experiences that include state, action, reward, and next state, and inputs these experiences into the loss function to compute gradients. The gradients are used to update the specific strategy parameters for the corresponding task. The physical guidance loss function based on domain knowledge and the DAM are described below.

3.1.1 Physical guidance loss function based on domain knowledge

When an agent simultaneously outputs multiple strategies, resulting in numerous action combinations

and a vast exploration space, it becomes necessary to steer its strategy optimization direction to align performance with expectations. The physical significance of the loss function lies in quantifying the deviation between the agent's behavior and the desired objective, aiding in guiding the agent's learning process and adjusting its strategy.

In response to this, we introduce and design a physical guidance loss function embedded with domain knowledge. This function uses empirical knowledge to construct a guidance term that incorporates implicit physical rules. The guidance term serves to reduce ineffective behavior during agent exploration, accelerate model convergence, and facilitate rapid iteration of the output strategy towards the desired objective.

The agent's physical guidance loss function, grounded in domain knowledge, comprises two components: the basic loss term $\mathrm{Loss}_{\mathrm{CLIP}}$ and the physical guidance loss term $\mathrm{Loss}_{\mathrm{knowledge}}$. The proportion of these components is regulated by the adjustment factor α. The central function of α is to balance the gradient contributions from the domain knowledge loss and the strategy optimization loss. Ideally, the gradient magnitudes of these two components should be comparable to prevent any single part from dominating the update direction.

Mathematically, this is expressed as

$$\mathrm{Loss}_{\mathrm{new}} = \mathrm{Loss}_{\mathrm{CLIP}} + \alpha \cdot \mathrm{Loss}_{\mathrm{knowledge}}, \tag{11}$$

where the value of α is

$$\alpha \approx \frac{\left\| \nabla \mathrm{Loss}_{\mathrm{CLIP}} \right\|}{\left\| \nabla \mathrm{Loss}_{\mathrm{knowledge}} \right\|}. \tag{12}$$

The basic loss term, $\mathrm{Loss}_{\mathrm{CLIP}}$, uses a clipped objective function that optimizes the surrogate loss. This loss function evaluates the ratio of the advantage of the new strategy over the old one, thereby constraining the policy update.

$$\mathrm{Loss}_{\mathrm{CLIP}} = \hat{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \mathrm{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon \right)\hat{A}_t \right) \right], \tag{13}$$

where $\hat{E}_t$ represents the expectation for time step t, $r_t(\theta)$ represents the ratio of the new strategy probability to the old strategy probability, ε is the hyperparameter, and $\hat{A}_t$ denotes the advantage estimate.

The physical guidance loss function, grounded in domain knowledge, is defined by the difference between the expected throughput and the actual throughput. The objective is for the agent's performance to progressively approach the theoretical maximum throughput.

In this context, n represents the number of nodes, and v_i denotes the actual rate of the i^{th} node, which is determined by the power and resource block allocation. The Shannon capacity formula is used as domain knowledge to establish the expected throughput.

$$\begin{aligned} E\left\{ v_{\mathrm{total}} \right\} &= E\left\{ B\log_2(1+\mathrm{SNR}) \right\} \\ &= E\left\{ B\log_2\left(1+\frac{P_{\mathrm{r}}}{N_0 B} \right) \right\}. \end{aligned} \tag{14}$$

Given that both N_0B and B are significantly greater than 1, and the signal-to-noise ratio (SNR) is appropriately amplified, the final equation can be formulated as follows:

$$\begin{aligned} E\left\{ B\log_2\left(1+\frac{P_{\mathrm{r}}}{N_0 B} \right) \right\} &\ll \log_2\left(1+E\left\{ P_{\mathrm{r}} \right\} \right) \\ &= \log_2\left(1+P_{\max} \right). \end{aligned} \tag{15}$$

The resultant loss function is then expressed as

$$\begin{aligned} \mathrm{Loss}_{\mathrm{new}} = \hat{E}_t\Big[&\min\left(r_t(\theta)\hat{A}_t, \mathrm{clip}\left(r_t(\theta), 1-\varepsilon, 1+\varepsilon \right)\hat{A}_t \right) \Big] \\ &+ \alpha \cdot \left[\log_2\left(1+P_{\max} \right) - \sum_{i=1}^{N} v_i \right]. \end{aligned} \tag{16}$$

Next, we analyze the properties of the physical guidance loss function rooted in domain knowledge.

It is known that $\mathrm{Loss}_{\mathrm{CLIP}}$ is smooth, bounded, and Lipschitz continuous (Schulman et al., 2017), and $\mathrm{Loss}_{\mathrm{knowledge}}$ can be abstracted as a function. Since the rate v, number of nodes n, and the expected throughput are all bounded, $\mathrm{Loss}_{\mathrm{new}}$ is also smooth, bounded, and Lipschitz continuous. Therefore, the physical guidance loss function $\mathrm{Loss}_{\mathrm{new}}$ satisfies the convergence conditions of the algorithm.

3.1.2 Distribution alignment module (DAM)

In scenarios where the agent simultaneously performs spectrum and power allocation actions,

mismatches in the action space may arise, leading to conflicts in the strategies. To address this issue, a versatile dynamic distribution mapping module is designed to optimize and adjust the multi-action output weights of agents in DRL. Depending on the degree of conflict between the action distributions, various methods are used to map and enhance the distribution, thereby avoiding unreasonable action combinations.

Specifically, considering the case where the agent simultaneously outputs two actions, A_{DAM} and B_{DAM}, if action A_{DAM} is selected as the benchmark, the distribution of action B_{DAM} is adjusted towards action A_{DAM}. When the output distributions of actions A_{DAM} and B_{DAM} do not completely conflict, a weighted mapping approach is used to adjust the weights, with the retention of action B_{DAM}'s information being controlled by parameters. Otherwise, when the output distributions of actions A_{DAM} and B_{DAM} are in complete conflict, the weights of the non-zero distribution are modified using a migration mapping method. The specific mapping formula is outlined below:

$$B'_{\mathrm{DAM}}[i]=\begin{cases}\dfrac{B_{\mathrm{DAM}}\cdot A_{\mathrm{DAM}}^{\beta'}}{\sum\left(B_{\mathrm{DAM}}\cdot A_{\mathrm{DAM}}^{\beta'}\right)}, & A_{\mathrm{DAM}}\cdot B_{\mathrm{DAM}}\neq 0,\\ \dfrac{\sum B_{\mathrm{DAM}}\cdot A_{\mathrm{DAM}}[i]}{\sum A_{\mathrm{DAM}}}, & A_{\mathrm{DAM}}[i]>0, A_{\mathrm{DAM}}\cdot B_{\mathrm{DAM}}=0,\\ 0, & A_{\mathrm{DAM}}[i]=0, A_{\mathrm{DAM}}\cdot B_{\mathrm{DAM}}=0,\end{cases} \tag{17}$$

where β' is the parameter that controls the degree of adjustment for the distributions of actions A_{DAM} and B_{DAM}.

3.2 Algorithm flow

The pseudo code for the KE-MAML method is given in Algorithm 1.

Algorithm 1 Knowledge-embedding model-agnostic meta-learning

Input: Step size hyperparameters——α, β, and γ

Output: Distribution over task——$p(T)$

// Step 1

Randomly initialize θ

while not done **do**

// Step 2

Sample batch of tasks $T_i\sim p(T)$

// Step 3

for all T_i **do**

Use DAM to select actions and interact with the environment according to θ

Calculate loss $\theta_i^{(k)}\leftarrow\theta_i^{(0)}-\beta\nabla_{\theta_i^{(0)}}L_{T_i}\left(\theta_i^{(k)};D_i\right)$

if experience buffer size > batch size

Compute adapted parameters with gradient descent: $\theta_i^{(k)}\leftarrow\theta_i^{(0)}-\beta\nabla_{\theta_i^{(0)}}L_{T_i}\left(\theta_i^{(k)};D_i\right)$

end if

end for

// Step 4

Update $\theta^{(0)}\leftarrow\theta^{(0)}-\beta\nabla_{\theta^{(0)}}\sum_{i=1}^{m}L_{T_i}\left(\theta_i^{(k)};D_i\right)$

end while

3.3 Algorithm complexity analysis

The optimization of the outer meta-parameters involves calculating the meta-gradients for multiple tasks. Let T denote the number of tasks. The gradients are computed, and a backpropagation step is performed for each task. Consequently, the time complexity for this step is $O(T\times k\times d_{\mathrm{w}})$, where k is the number of meta-learning task gradient updates and d_{w} is the dimensionality of the model parameters.

In the inner loop, the steps for updating the strategy include data collection, loss function calculation, and strategy updating. The time complexities for these steps are $O(N\times d_{\mathrm{w}})$, $O(N)$, and $O(m\times d_{\mathrm{w}})$, respectively. Here, N is the number of samples and m denotes the number of gradient updates during the inner loop updates. Therefore, the total time complexity for the inner loop is approximately $O(N\times d_{\mathrm{w}}+N+m\times d_{\mathrm{w}})$.

By combining the complexities of both the inner and outer loops, the overall time complexity of the algorithm is $O(T\times k\times d_{\mathrm{w}}+N\times d_{\mathrm{w}}+m\times d_{\mathrm{w}})$.

The model used in this method consists of three fully connected layers, with 300, 150, and 8 neurons in each layer. The number of parameters d_{w} is 49 820. The number of gradient updates m for the inner DRL algorithm is set to 60 during training and 5 during testing. For the meta-learning task, the number of gradient updates k is set to 30, 10, and 5 during training, and 0 during testing.

3.4 Algorithm convergence analysis

The convergence of the proposed method is discussed in two parts. First, the inner layer uses a reinforcement learning (RL) algorithm whose convergence is based on the theoretical framework of the strategy gradient method. By introducing a near-end constraint, the policy update range is limited, enabling the method to approach the optimal strategy while ensuring stability. Previous studies (Schulman et al., 2017) have shown that the near-end strategy optimization algorithm is convergent. The adaptation of this method to the near-end strategy optimization algorithm does not alter the algorithm's convergence condition, allowing us to consider that the inner layer algorithm of the proposed method is convergent.

The outer layer uses an improved MAML algorithm. MAML enables the model to quickly adapt to new tasks through multi-task training, with its convergence typically based on the assumption of task distribution. First, the meta-training tasks and test tasks originate from the same distribution, providing a geometric foundation for convergence. Second, as discussed in Section 3.1, the physical guidance loss function, designed for the MAML algorithm and based on domain knowledge, is Lipschitz continuous, which establishes the boundedness of the algorithm's parameter update direction. According to the MAML convergence theory (Fallah et al., 2020), this ensures that the outer algorithm of the proposed method is also convergent.

In summary, both the inner and outer layers of the proposed method are convergent, showing that the method achieves overall convergence.

4 Simulation results and analysis

4.1 Preparing the scenario

The simulation scenario is set within a sea surface region measuring $R \times R$, and includes a designated area that represents the communication environment. Within this region, an intelligent, low-speed base station node operates, while n high-speed user nodes (ships and aircraft) are randomly distributed and move continuously. The high-speed nodes are categorized into two groups: ship nodes and aircraft nodes, with their proportions randomly generated at the beginning of the simulation. All nodes exhibit random movement patterns, simulating realistic maritime conditions.

Fig. 4 shows a schematic of the simulation scenario. The base station node communicates with any mobile nodes that enter its range, while nodes outside this range are disconnected. Communication is established under the assumption that all nodes are within the base station's communicable range, with a communication frequency band selected from the satellite communication spectrum to account for maritime conditions.

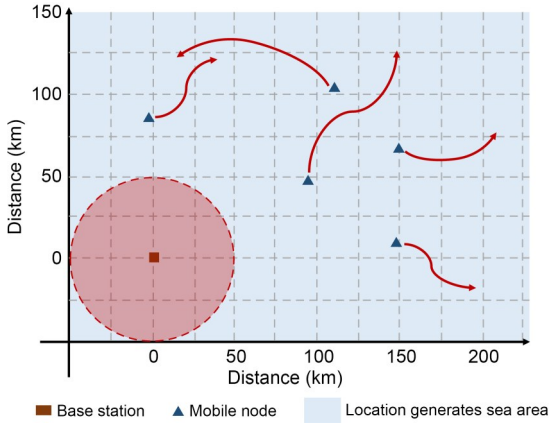

Fig. 4 Schematic of simulation scenario

The simulation consists of both training and testing phases. In the training phase, 600 tasks are randomly generated across three distinct interference channel environments, with each task lasting 60 s, totaling 36 000 samples. In the testing phase, 20 unique tasks are randomly selected from each of the three different channel environments, resulting in a total of 60 tasks. For performance evaluation, moving averages of every 10 tasks are computed.

(1) Channel environments are based on marine conditions;

(2) Node positions and velocities are initialized randomly;

(3) All nodes undergo continuous random movements with varying altitudes and communication link parameters.

Ocean channel parameters are referenced from ITU (2016), basic scene parameters are from Xia et al. (2020), and environmental communication factors such as weather and Doppler frequency shift are from

Bekkadal (2010). Other simulation parameters are provided in Table 1.

Table 1 Simulation parameter settings

Simulation parameter	Value
Number of agents	1
Agent speed (m/s)	[1,15]
Subscriber number	2–8
User speed (m/s)	[5,18]/[125,280]
Node turning angle	[−0.25π, 0.25π]
Task duration (s)	60
Agent transmit power (W)	100
Channel interference in training scenario (dB)	10±4、15±3、20±2
Test scenario channel interference (dB)	10, 25, 35
Optional range of node spectrum (MHz)	2010–2030
Resource block bandwidth (kHz)	15

To validate the performance of the proposed KE-MAML method, several leading meta-learning algorithms are selected for comparison:

(1) MAML (Finn et al., 2017);

(2) Model-agnostic meta-learning proximal policy optimization (MAML-PPO) (Jang et al., 2021);

(3) RL^2 (Duan et al., 2016).

Additionally, ablation experiments are conducted to validate the functionality and performance of the proposed components. The ablation comparisons include:

(1) Loss_MAML-PPO—A version of the MAML-PPO algorithm where the loss function is modified to incorporate domain-specific knowledge through a physical guidance loss function;

(2) DAM-MAML-PPO—The MAML-PPO algorithm enhanced solely by the inclusion of the DAM.

Each algorithm adopts a neural network of uniform size and training times, with hyperparameters shown in Table 2.

Table 2 Algorithm hyperparameters

Hyperparameter	Value
Learning rate	10^{-4}
Number of layers	3
Number of neurons	300/150/8
Sample time (s)	1
Discount factor	0.95
Sample batch size	600
Experience buffer length	1.2×10^{-5}
Outer loop update times (training/testing)	30/0
Inner loop update times (training/testing)	20/5

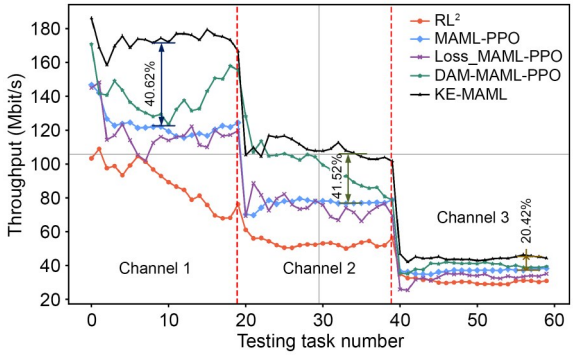

Fig. 5 Comparison of total throughput

4.2 Analysis of simulation results

Fig. 5 presents a comparison of the total throughput for nodes connected to the base station over the task duration, with the channel environments deteriorating from left to right. To begin, we focus on comparing performance within the ablation experiments. Examining the results under Channel 1 conditions, it is evident that, compared to MAML-PPO, the DAM-MAML-PPO in the ablation experimental group showed an improvement of 14.28%. This indicates that the DAM effectively reduces the output of non-standard actions, mitigating resource idleness or waste caused by policy conflicts. However, there remains a 23.04% gap relative to the proposed method. This discrepancy arises when the DAM converges the output distribution, as the information loss during the mapping process leads to inefficiencies or failures in the original strategy. As a result, this limits further performance enhancement and reduces the MAML-PPO algorithm's ability to optimize strategies.

On the other hand, Loss_MAML-PPO in the ablation experimental group shows a reduction in performance of 4.21%. This decline occurs when the agent's resource block allocation and power allocation actions conflict. The mere embedding of physical knowledge is insufficient to robustly guide the model, and it struggles to compensate for the performance drop induced by policy conflicts. When the physical knowledge loss function guides the model towards a better solution, the conflicting allocation of resource blocks and power slightly increases the difficulty for the agent to find the optimal solution.

However, when both modules are combined, the performance shows a significant improvement of 40.62% compared to the benchmark MAML-PPO algorithm. This enhancement is attributable to the DAM's ability to reduce unnecessary exploration by avoiding action combinations that deviate from physical reality, while the physical knowledge loss function guides the agent's exploration. Consequently, the physical knowledge loss function compensates for the information loss during the action distribution mapping process, thereby boosting the effectiveness of the combined modules.

In contrast, the RL^2 algorithm struggles to handle highly complex tasks and adapt to environments with strong dynamics and scarce data. Furthermore, the strategic conflicts among the agent's multiple output actions contribute to its overall inferior performance compared to MAML-PPO and its variants.

When comparing performance across different channel environments, the proposed method consistently shows a significant performance advantage, even as the channel conditions worsened. In Channel 2 and Channel 3 environments, performance improvements of 41.52% and 20.42%, respectively, are observed relative to the standard MAML-PPO algorithm. However, the margin of improvement gradually diminishes as the channel conditions deteriorate. This trend can be attributed to the severe channel impairments, which limit the maximum communication capacity, thereby reducing both the number of optimal solutions and the performance differences between various allocation schemes.

Fig. 6 presents the comparison of average throughput for number-independent nodes, while Fig. 7 shows the comparison of fairness coefficients for power allocation. In the Channel 1 environment, the proposed method outperforms the benchmark method (MAML-PPO) by 49.64% in average throughput and achieves an increase in average fairness coefficient of 0.04. This indicates that the proposed method not only improves throughput but also maintains a reasonable level of fairness. By comparing the performance of the algorithm across three different channel environments, we observe that the transmission rate decreases as channel interference increases. In the Channel 2 environment, although the average performance of the DAM-MAML-PPO method is comparable to that of

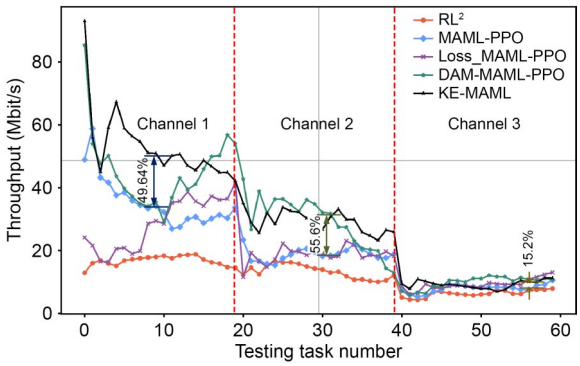

Fig. 6 Comparison of average throughput for number-independent nodes

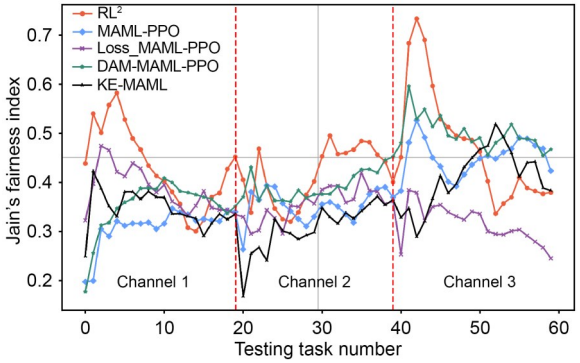

Fig. 7 Comparison of fairness coefficient

the proposed method, the transmission rate shows significant variability due to randomness. This variability further demonstrates the strong adaptability of the proposed method in dynamic environments. In the Channel 3 environment, the impact on the transmission rate is more pronounced, compressing the performance advantages among the methods. Nevertheless, the proposed method still achieves a 15.2% improvement.

A comparison between Figs. 5 and 6 reveals that the curves for total and average throughput of the proposed algorithm's nodes do not align, suggesting that the algorithm's strategy prioritizes strengthening higher-performing nodes over supporting weaker ones. This reflects the resource allocation strategy in scenarios where resources are constrained or unfamiliar.

Additionally, when compared to other control algorithms over time, although the proposed algorithm achieves higher average throughput in Channel 1, its fairness coefficient is lower. This implies that, in this scenario, throughput and fairness are inversely related. However, in Channels 2 and 3, the fairness

coefficient of the proposed algorithm gradually improves relative to other control algorithms. This improvement is attributable to the physical guidance loss function based on knowledge embedding, which helps the agent recognize that more challenging environments lead to smaller disparities in node allocation schemes, resulting in a fairer allocation with higher throughput. This highlights the proposed algorithm's ability to quickly adapt to environmental changes, identify optimal solutions for maximizing throughput, and exhibit strong generalization capabilities in unfamiliar environments.

Fig. 8 compares the continuous adaptability of the proposed algorithm across various $P_{\max}$ values. The results clearly show that the proposed method shows excellent continuous adaptability at all $P_{\max}$ values. As the model updates in response to changing channel environments, the overall throughput performance remains stable, with no significant drops. This highlights the superior adaptability of the proposed method.

Table 3 presents the average algorithm performance for tasks with the same channel condition intervals. The proposed method shows significant improvements in throughput performance across various channel environments, albeit with a trade-off in fairness. This effectively addresses the needs of resource-constrained environments, where maximizing throughput is often prioritized over fairness.

Notably, the superior performance of the meta-reinforcement learning algorithm is closely linked to the pre-trained global parameter model. Fig. 9 and Table 4 present the curve comparisons and the mean value and mean squared error comparisons of the model's maximum throughput performance across varying numbers of inner and outer loop training iterations.

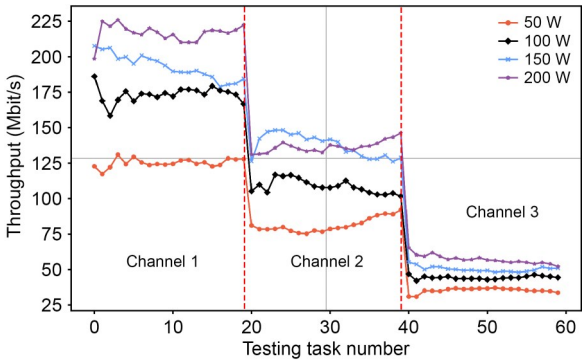

Fig. 8 Throughput comparison under different $P_{\max}$ values

Here, $x \times y$ represents the number of outer loops, while y denotes the number of tasks learned in each outer loop. A comparison between Figs. 9a and 9b reveals minimal performance disparities among models corresponding to different training quantities across the three channel environments. In contrast, the benchmark MAML-PPO algorithm shows significant performance variation, indicating a higher dependency on pre-training data compared to the proposed method. This is because the proposed algorithm leverages the physical guidance loss function, based on knowledge embedding, to guide model optimization even in the absence of extensive pre-training data. Additionally, the DAM reduces impractical action combinations, minimizing the data required for model exploration and learning, thereby improving the efficiency of pre-training data utilization. Consequently, the volume of pre-training data has a more significant impact on the benchmark algorithm than on the proposed method. Table 4 further shows that the mean squared error of the total throughput for the proposed method is smaller than that of the benchmark algorithm, highlighting the proposed algorithm's robustness in scenarios where

Table 3 Algorithm performance comparison

Algorithm	Total throughput (Mbit/s)			Average throughput for number-independent nodes (Mbit/s)			Fairness index		
	Channel 1	Channel 2	Channel 3	Channel 1	Channel 2	Channel 3	Channel 1	Channel 2	Channel 3
DAM-MAML-PPO	140.90	98.52	39.71	45.42	29.05	9.80 ↑	0.35	0.39	0.50 ↑
Loss_MAML-PPO	118.10	74.52	32.76	28.81	18.57	9.20	0.38	0.36	0.31
MAML-PPO	123.29	76.67	36.77	35.41	18.76	7.96	0.31	0.35	0.45
RL^2	88.77	53.25	30.59	16.73	13.26	6.30	0.43 ↑	0.42 ↑	0.48
KE-MAML	173.37 ↑	108.50 ↑	44.28 ↑	52.99 ↑	29.19 ↑	9.17	0.35	0.31	0.41

↑ indicates the optimal parameter

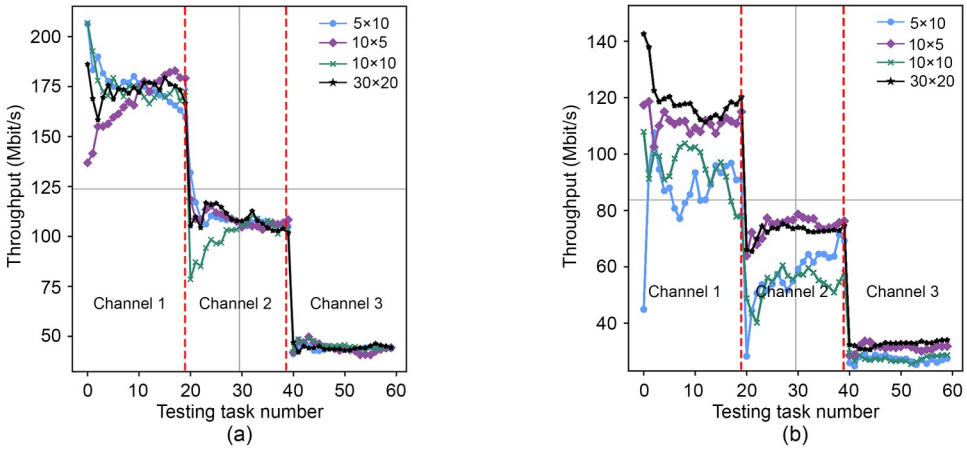

Fig. 9 Comparison of total throughput under different training data sizes for KE-MAML (a) and MAML-PPO (b)

Table 4 Comparison of total throughput under different training data sizes

Training data size	KE-MAML			MAML-PPO		
	Channel 1	Channel 2	Channel 3	Channel 1	Channel 2	Channel 3
5×10	175.71	108.94	43.79	92.00	61.40	31.27
10×5	167.37	108.55	44.04	115.41	78.57	35.49
10×10	175.19	100.36	45.16	99.10	58.23	31.60
30×20	173.37	108.50	44.28	123.29	76.67	36.77
Mean squared error	3.83 ↓	4.16 ↓	0.60 ↓	14.41	10.39	2.76

↓ indicates the optimal parameter

obtaining substantial pre-training data is challenging. This indicates a reduced reliance on pre-training data.

5 Conclusions

In maritime wireless transmission networks with resource-constrained OFDMA systems, power and spectrum allocation schemes play a crucial role in enhancing system throughput. To optimize this allocation, we propose a time–frequency resource management model coupled with an objective optimization function specifically designed for such systems. Additionally, we introduce KE-MAML, a joint resource allocation meta-reinforcement learning method based on knowledge embedding. The loss function of physical guidance, based on knowledge embedding, is tailored for the inner loop. The optimization and updating of this guidance model allow the agent to minimize ineffective experience in strategy learning within the current environment. Additionally, to address the issue of the low correlation between strategies when multiple strategies are output, an action DAM is designed at the action output stage. This module helps prevent unreasonable action-output combinations, reduces the difficulty of agent exploration, and minimizes resource waste caused by inappropriate action matching. Through these design enhancements, inefficient or ineffective action combinations and experience accumulation are curtailed, thereby reducing the amount of pre-training data required for the model.

Simulation results show that KE-MAML serves as an effective communication resource allocation strategy, sacrificing fairness in favor of maximizing system throughput in resource-limited and unfamiliar environments. Compared to various benchmark methods, KE-MAML significantly enhances system throughput across diverse channel conditions, further validating the effectiveness of our joint resource allocation meta-reinforcement learning method grounded in knowledge embedding.

This study not only broadens the application of existing knowledge embedding techniques but also offers novel solutions to the challenges of communication

resource allocation in unfamiliar environments. In the future, our work will concentrate on several key areas: the collection and verification of real-world maritime wireless communication network data, addressing the catastrophic forgetting problem in the model, and tackling the adaptation challenges associated with extending it to multi-agent systems. These efforts aim to enhance the robustness and applicability of our approach in practical, dynamic, and complex maritime environments.

Contributors

Zhongyang MAO designed the research. Zhilin ZHANG and Yang YOU processed the data. Jiafang KANG and Yaozong PAN drafted the paper. Xiguo LIU and Zhichao XU helped organize the paper. Zhilin ZHANG and Faping LU revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

Data are not available due to legal restrictions. Due to the nature of this research, participants of this study did not agree for their data to be shared publicly, so supporting data are not available.

References

Bekkadal F, 2010. Innovative maritime communications technologies. Proc 18th Int Conf on Microwaves, Radar and Wireless Communications, p.1-6.

Bossy B, Kryszkiewicz P, Bogucka H, 2022. Energy-efficient OFDM radio resource allocation optimization with computational awareness: a survey. *IEEE Access*, 10:94100-94132. https://doi.org/10.1109/ACCESS.2022.3203575

Chen SY, Rui LL, Gao ZP, et al., 2022. Cache-assisted collaborative task offloading and resource allocation strategy: a metareinforcement learning approach. *IEEE Internet Things J*, 9(20):19823-19842. https://doi.org/10.1109/JIOT.2022.3168885

Dhuheir M, Erbad A, Al-Fuqaha A, et al., 2024. Meta reinforcement learning for UAV-assisted energy harvesting IoT devices in disaster-affected areas. *IEEE Open J Commun Soc*, 5:2145-2163. https://doi.org/10.1109/OJCOMS.2024.3377706

Duan Y, Schulman J, Chen X, et al., 2016. RL^2: fast reinforcement learning via slow reinforcement learning. https://doi.org/10.48550/arXiv.1611.02779

Fallah A, Mokhtari A, Ozdaglar A, 2020. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. Proc 23rd Int Conf on Artificial Intelligence and Statistics, p.1082-1092.

Ferreira GO, Zanella AF, Bakirtzis S, et al., 2024. A joint optimization approach for power-efficient heterogeneous OFDMA radio access networks. *IEEE J Select Areas Commun*, 42(11):3232-3245. https://doi.org/10.1109/JSAC.2024.3431524

Finn C, Abbeel P, Levine S, 2017. Model-agnostic meta-learning for fast adaptation of deep networks. Proc 34th Int Conf on Machine Learning, p.1126-1135.

Gautam S, Lagunas E, Chatzinotas S, et al., 2019. Relay selection and resource allocation for SWIPT in multi-user OFDMA systems. *IEEE Trans Wirel Commun*, 18(5):2493-2508. https://doi.org/10.1109/TWC.2019.2904273

Han J, Lee GH, Park S, et al., 2022. Joint subcarrier and transmission power allocation in OFDMA-based WPT system for mobile-edge computing in IoT environment. *IEEE Internet Things J*, 9(16):15039-15052. https://doi.org/10.1109/JIOT.2021.3103768

Hou QS, Lee M, Yu GD, et al., 2023. Meta-gating framework for fast and continuous resource optimization in dynamic wireless environments. *IEEE Trans Commun*, 71(9):5259-5273. https://doi.org/10.1109/TCOMM.2023.3292257

Hu SY, Yuan X, Ni W, et al., 2024. OFDMA-F^2L: federated learning with flexible aggregation over an OFDMA air interface. *IEEE Trans Wirel Commun*, 23(7):6793-6807. https://doi.org/10.1109/TWC.2023.3334691

ITU, 2016. Recommendation ITU-R P.372-13. https://www.itu.int/rec/R-REC-P.372-13-201609-S

Jang D, Spangher L, Khattar M, et al., 2021. Using meta reinforcement learning to bridge the gap between simulation and experiment in energy demand response. Proc 12th ACM Int Conf on Future Energy Systems, p.483-487. https://doi.org/10.1145/3447555.3466589

Jha S, Ahmad S, Abdeljaber HAM, et al., 2024. Enabling resilient wireless networks: OFDMA-based algorithm for enhanced survivability and privacy in 6G IoT environments. *IEEE Trans Consum Electr*, 70(1):3810-3819. https://doi.org/10.1109/TCE.2024.3370414

Jin ZW, Ma ML, Wang Z, et al., 2025a. Optimal transmission schedule with privacy preservation for cyber-physical system against eavesdropping attack. *IEEE Signal Process Lett*, 32:436-440. https://doi.org/10.1109/LSP.2024.3514793

Jin ZW, Xu CH, Wang Z, et al., 2025b. Towards robust differential privacy in adaptive federated learning architectures. *IEEE Trans Consum Electr*, 71(2):4087-4099. https://doi.org/10.1109/TCE.2024.3525084

Kim Y, Choi Y, Yang HJ, 2023. Spectrum sensing for underwater cognitive radio with limited sensing time. *IEEE Commun Lett*, 27(8):2014-2018. https://doi.org/10.1109/LCOMM.2023.3291079

Le NT, Tran LN, Vu QD, et al., 2019. Energy-efficient resource allocation for OFDMA heterogeneous networks. *IEEE Trans Commun*, 67(10):7043-7057. https://doi.org/10.1109/TCOMM.2019.2936813

Letchford AN, Ni Q, Zhong ZY, 2020. A heuristic for fair dynamic resource allocation in overloaded OFDMA systems. *J Heuristics*, 26(1):21-32. https://doi.org/10.1007/s10732-019-09422-z

Li SC, Zhang N, Chen HB, et al., 2022. Joint subcarrier allocation, modulation mode selection, and trajectory design in a UAV-based OFDMA network. *IEEE Commun Lett*, 26(9):2111-2115. https://doi.org/10.1109/LCOMM.2022.3182016

Liu L, Cai L, Ma L, et al., 2021. Channel state information prediction for adaptive underwater acoustic downlink OFDMA system: deep neural networks based approach. *IEEE Trans Veh Technol*, 70(9):9063-9076. https://doi.org/10.1109/TVT.2021.3099797

Mao ZY, Zhang ZL, Lu FP, et al., 2024. Sea-based UAV network resource allocation method based on an attention mechanism. *Electronics*, 13(18):3686. https://doi.org/10.3390/electronics13183686

Meister G, Knuble JJ, Gliese U, et al., 2024. The ocean color instrument (OCI) on the plankton, aerosol, cloud, ocean ecosystem (PACE) mission: system design and prelaunch radiometric performance. *IEEE Trans Geosci Remote Sensing*, 62:5517418. https://doi.org/10.1109/TGRS.2024.3383812

Ning JH, Wang JL, Feng P, et al., 2023. A distributed framework for the ocean IoT network. Proc 34th Annual Int Symp on Personal, Indoor and Mobile Radio Communications, p.1-6. https://doi.org/10.1109/PIMRC56721.2023.10294049

Schulman J, Wolski F, Dhariwal P, et al., 2017. Proximal policy optimization algorithms. https://doi.org/10.48550/arXiv.1707.06347

Shi XH, Zhang S, Liu MZ, et al., 2025. Mystique: user-level adaptation for real-time video analytics in edge networks via meta-RL. *IEEE Trans Mob Comput*, 24(5):3615-3632. https://doi.org/10.1109/TMC.2024.3514088

Su YS, Liu X, Han GY, et al., 2021. A traffic load-aware OFDMA-based MAC protocol for distributed underwater acoustic sensor networks. *IEEE Trans Veh Technol*, 70(10): 10501-10513. https://doi.org/10.1109/TVT.2021.3109070

Sun GX, Wang XM, Jiang R, et al., 2022. Beamforming and resource allocation in multi-cell OFDMA systems based on deep transfer reinforcement learning. Proc 95th Vehicular Technology Conf, p.1-6. https://doi.org/10.1109/VTC2022-Spring54318.2022.9860615

Švedek V, Kurdija AS, Ilic Ž, 2023. Static and mobile relay selection with chunk-based subcarrier allocation in uplink OFDMA networks. Proc Int Symp on ELMAR, p.137-140.

Tan QY, He JJ, Gao YY, 2024. Deep reinforcement learning based OFDMA scheduling for WiFi networks with coexisting latency-sensitive and high-throughput services. Proc 5th Information Communication Technologies Conf, p.146-150. https://doi.org/10.1109/ICTC61510.2024.10601889

Tefera MK, Zhang SB, Jin ZW, 2023. Deep reinforcement learning-assisted optimization for resource allocation in downlink OFDMA cooperative systems. *Entropy*, 25(3): 413. https://doi.org/10.3390/e25030413

Tseng SM, Wang PH, Hsu YT, 2023. Modified loss function considering outage capacity for deep learning-based OFDMA NOMA video transmission resource management. Proc 8th Int Conf on Multimedia Communication Technologies, p.7-11. https://doi.org/10.1109/ICMCT60483.2023.00009

Wang J, Zhou HF, Li Y, et al., 2018. Wireless channel models for maritime communications. *IEEE Access*, 6:68070-68088. https://doi.org/10.1109/ACCESS.2018.2879902

Wang LY, Guo J, Zhu JQ, et al., 2024. Cross-layer wireless resource allocation method based on environment-awareness in high-speed mobile networks. *Electronics*, 13(3):499. https://doi.org/10.3390/electronics13030499

Wang T, You CC, 2024. Adaptive uplink scheduling and UAV association in UAV-assisted OFDMA cellular networks: a game-theoretical approach. *IEEE Access*, 12:63504-63514. https://doi.org/10.1109/ACCESS.2024.3396152

Wang T, You CC, He Z, et al., 2023. Distributed subcarrier assignment and discrete power allocation for multi-UAV millimeter-wave cooperative OFDMA networks with heterogeneous QoS consideration. *IEEE Access*, 11:123132-123148. https://doi.org/10.1109/ACCESS.2023.3328214

Wang XH, Su YS, Yang SD, et al., 2024. An OFDMA downlink acoustic communication scheme for AUV-based mobile underwater sensor network. *IEEE Sens J*, 24(7): 11527-11536. https://doi.org/10.1109/JSEN.2024.3361152

Wang XM, Sun GX, Xin YX, et al., 2022. Deep transfer reinforcement learning for beamforming and resource allocation in multi-cell MISO-OFDMA systems. *IEEE Trans Signal Inform Process Netw*, 8:815-829. https://doi.org/10.1109/TSIPN.2022.3208432

Xia TT, Wang MM, Zhang JJ, et al., 2020. Maritime Internet of Things: challenges and solutions. *IEEE Wirel Commun*, 27(2):188-196. https://doi.org/10.1109/MWC.001.1900322

Yan RW, Li Q, Xiong HG, 2024. Adaptive channel division and subchannel allocation for orthogonal frequency division multiple access-based airborne power line communication networks. *Sensors*, 24(23):7644. https://doi.org/10.3390/s24237644

Yang LW, Jia BY, Wang F, et al., 2022. Energy efficiency optimization of heterogeneous network resources based on OFDMA. Proc 20th Int Conf on Optical Communications and Networks, p.1-3. https://doi.org/10.1109/ICOCN55511.2022.9900961

Yang SD, Su YS, Wang XH, et al., 2024. Resource allocation for cognitive underwater acoustic downlink OFDMA system with a practical spectrum sensing scheme. *IEEE Internet Things J*, 11(5):8731-8745. https://doi.org/10.1109/JIOT.2023.3320391

Yin H, Huang YH, Han LC, et al., 2023. Thoughts on 6G integrated communication, sensing and computing networks. *Sci Sin Inform*, 53(9):1838-1842 (in Chinese). https://doi.org/10.1360/SSI-2023-0135

Yuan X, Hu SY, Ni W, et al., 2023. Joint user, channel, modulation-coding selection, and RIS configuration for jamming resistance in multiuser OFDMA systems. *IEEE Trans Commun*, 71(3):1631-1645. https://doi.org/10.1109/TCOMM.2023.3238062

Zhang L, Han SQ, Yang CY, 2023. Joint scheduling and power allocation with per-user rate constraints for uplink MU-MIMO OFDMA systems. Proc 97th Vehicular Technology Conf, p.1-5. https://doi.org/10.1109/VTC2023-Spring57618.2023.10200843