scientific reports

OPEN



Resource allocation of distributed MIMO radar based on the hybrid action space reinforcement learning

Jin Zhu^{1,2⊠}, Wenxu Liu², Feifei Lyu², Siwei Li¹ & Tianyang Zhang¹

The distributed multiple-input multiple-output (MIMO) radar system exhibits superior target localization capability by jointly processing target information from multiple radars under different observation angles. To improve the resource utilization of the distributed MIMO radar system, this paper proposes a hybrid action space reinforcement learning (HAS-RL) method, aiming to maximize the target localization performance under the radar resource constraints. Specifically, the Cramer–Rao Lower Bound (CRLB) incorporating the transmit radar power and receive radar selection is first derived and employed as the target localization performance metric of the distributed MIMO radar system. Subsequently, the radar resource allocation problem is modeled as a constrained optimization problem with continuous and discrete variables, and a hybrid action space reinforcement learning is proposed to solve the above optimization problem. Simulation results demonstrate that the proposed HAS-RL method can obtain better target localization performance under the given radar resource constraints.

Keywords Resource allocation, Radar signal processing, Reinforcement learning

The distributed multiple-input multiple-output (MIMO) radar system is a widely used radar system¹⁻³. Since receive radars in the system receive the echo signals from different observation angles, the distributed MIMO radar system can obtain more precise target information⁴. Compared with the traditional monostatic radar system, the distributed MIMO radar system has more accurate target localization and more substantial tracking capability.

With changes in mission requirements and application scenarios, the radar system needs to optimize resource utilization to meet different performance requirements. The distributed MIMO radar system contains various resources, such as transmitted power, bandwidth, radar position, etc. Theoretically, maximizing each resource in the radar system can achieve a higher target localization accuracy⁵. However, in practical applications, the radar system's resources are usually constrained. Thus, how to allocate constrained resources and improve resource utilization efficiency has become an important topic in the distributed MIMO radar system^{6–21}.

Godrich et al.⁶ first established the transmitted power allocation optimization model for target localization and solved the optimization problem with convex relaxation and local optimization algorithms. Feng et al.⁸ designed an alternating global search algorithm to improve the transmitted power utilization of radar systems. Shi et al.¹⁰ introduced semi-positive definite planning and Karush–Kuhn–Tucker (KKT) conditions to obtain better target localization performance under the radar system's power constraints. Guo et al.¹² proposed an improved depth-first search-based approach to solve the radar system's resource allocation for target localization.

Besides, some researchers¹⁴⁻²⁰ consider jointly optimizing the allocation of transmitted power, bandwidth, and radar position to improve the target localization performance of radar systems. Ma et al.¹⁴ considered a joint optimization scheme of radar selection and power allocation in a MIMO radar network to minimize the target localization error. The proposed scheme divides the optimization process into two steps, where each step transforms the optimization problem into Second-Order Cone Programming (SOCP) for a solution through convex relaxation. Sun et al.¹⁵ analyzed the effects of the transmit radar, transmitted power, and bandwidth on the target localization performance of the distributed MIMO radar system and used the cyclic minimization algorithm to decompse the joint resource optimization into two sub-optimization problems. However,

¹School of Artificial Intelligence, Xidian University, Xi'an 710071, China. ²CETC Key Laboratory of Aerospace Information Applications, The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China. ^{\Box}email: zhujin_cetc54@126.com decomposing the joint optimization of multiple resources into several sub-optimization problems requires more constraints or optimization processes.

Drawing on the above problem, this paper regards the joint transmitted power and transmit radar selection problem as a single optimization problem of transmitted power, in which the transmit radar is not selected when its transmitted power is set to 0. Specifically, we first employ the Cramer–Rao Lower Bound (CRLB) as the target localization performance metric of the distributed MIMO radar system. Then, we establish the resource allocation problem as a constrained optimization problem. Finally, we propose a hybrid action space reinforcement learning (HAS-RL) method to maximize the target localization performance under the radar resource constraints. The experimental results under two simulation scenarios demonstrate the effectiveness of our proposed HAS-RL method for radar resource allocation.

The remainder of this paper is organized as follows: Section II provides the derivation of the CRLB. Section III establishes the constrained optimization problem for resource allocation and describes the proposed HAS-RL method. The simulation experiments are reported in Section IV. Section V discusses the differences between our proposed method and other related methods. Finally, Section VI concludes this paper.

System model and Cramer–Rao lower bound System model

Given the 3D coordinate system, a distributed MIMO radar system consists of M transmit radars and N receive radars all placed at large distances⁴. The *m*-th transmit radar includes K_{T_m} individual nodes, and its phase center is located at $(\mathbf{x}_{m_t}, y_m, z_m), m = 1, 2, ...M$. Similarly, the *n*-th receive radar includes K_{R_n} individual nodes, and its phase center is located at $(\mathbf{x}_{n_t}, y_n, z_n), m = 1, 2, ...M$. Similarly, the *n*-th receive radar includes K_{R_n} individual nodes, and its phase center is located at $(\mathbf{x}_{n_t}, y_n, z_n), n = 1, 2, ...N$. Fig. 1 depicts the distributed MIMO radar system in the 3D coordinate system.

Assume that the waveform transmitted by the transmit radar T_m is $s_m(t)$ following the normalization $\int T |s_m(t)|^2 dt = 1$, and the effective bandwidth of waveform $s_m(t)$ is denoted by β_m . To simplify the analysis process, it is assumed that the waveforms of each transmit radar are orthogonal and satisfy the following equation:

$$\int {}_{T}\mathbf{s}_i(t) * s_j(t-\tau)dt = \begin{cases} 1 & i=j, \\ 0 & i\neq j, \end{cases}$$
(1)

Given the target Q located at (x_q, y_q, z_q) , the waveform transmission delay, transmitted by the transmit radar T_m , reflected by target Q then received by the receiver radar R_n , is defined as follows:

$$\tau_{mn} = \frac{\frac{\mathbb{R}(T_m,Q) + \mathbb{R}(R_n,Q)}{c}}{\sqrt{(x_m - x_q)^2 + (y_m - y_q)^2 + (z_m - z_q)^2}} + \sqrt{(x_n - x_q)^2 + (y_n - y_q)^2 + (z_n - z_q)^2},$$
(2)

where *c* stands for the light speed.

The baseband signal received by the receive radar R_n can be represented as

$$Z_{n}(t) = aR_{n} \sum_{m=1}^{M} h_{mn} \alpha_{mn} \sqrt{P_{m}} [w_{T_{m}}^{T} * a_{T_{m}}] s_{m}(t - \tau_{mn}) + \mathbf{AWGN}_{n}(t),$$
(3)



Fig. 1. Transmit and receive radars distribution in MIMO radar systems.

Scientific Reports | 2025 15:19593

where $\alpha_{mn,X_q} \propto \frac{1}{(T_m,X_q)^*(R_n,X_q)}$ is the path loss between the *m*-th transmit radar and the *n*-th receive radar. h_{mn} is the radar cross-section (RCS) of the *m*-*n* path with respect to target *Q*. P_{mn} , a_{T_m} , and w_{T_m} are the transmitted power, beamforming weight, and the steering vector of the transmit radar T_m . a_{R_n} is the steering vector of the receive radar R_n . **AWGN**_n(t) = [AWGN_{n,1}(t); \cdots ; AWGN_{n,KR_m}(t)] represents the Additive White Gaussian Noise (AWGN) with the characteristics of AWGN(t) ~ $N(0, \sigma_w^2)$.

Thus, the baseband signal received by the receive radar R_n after beamforming can be represented as:

$$\mathbf{r}_n(t) = \mathbf{w}_{R_n}^T * Z_n(t),\tag{4}$$

where w_{R_n} is the beamforming weight of the receive radar R_n .

Assumed that
$$G_{T_m} \stackrel{=}{=} w_{T_m}^T * a_{T_m}, G_{R_n} \stackrel{=}{=} w_{R_n}^T * a_{R_n}$$
 is the radar gain. When the regular transmit/
receive beamforming is used, and the target is illuminated by the main lobe, $G_{T_m} = K_{T_m}, G_{R_m} = K_{R_m}$.
Then the echo signal at time t of the n-th receive radar after receive beamforming is given by:

$$r_n(t) = G_{R_n} \sum_{m=1}^{M} h_{mn} \alpha_{mn} \sqrt{P_m} G_{T_m} s_m(t - \tau_{mn}) + N_n(t),$$
(5)

where $N_n(t) \sim N(0, K_{R_n} \sigma_w^2), n = 1, 2, \cdots, N$

Cramer-Rao lower bound

The Cramer–Rao Lower Bound (CRLB) denotes the lower bound of the mean square errors between each estimated result and the variant to be estimated, which measures the reliability of the estimation. If the mean square errors between each estimated result and the variant to be estimated can reach the CRLB, the estimated results can be approximated as substitutes for the variant to be estimated.

In the target localization task of the distributed MIMO radar system, the position information of target Q needs to be estimated and is defined as $\theta = [x_q, y_q, z_q]^T$. The conditional probability density function of the θ can be expressed as follows:

$$p(r_q|\theta) = \frac{1}{(2\pi K_R \sigma_w^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2K_R \sigma_w^2} \sum_{n=1}^N \int_T \left[r_{qn}(t) - G_{R_n} \sum_{m=1}^M \alpha_{mn} \sqrt{P_m} G_{T_m} h_{mn} s_m(t-\tau_{mn})\right]^2 dt\right\}, \quad (6)$$

where $r_{qn}(t)$ represents the echo signal of target Q received from receive radar R_n . Then, the Fisher Information Matrix (FIM) of θ can be expressed as follows.

$$\mathbf{J}(\theta) = E\{\left[\nabla_{\theta} \operatorname{In}p(r_q|\theta)\right] \left[\nabla_{\theta} \operatorname{In}p(r_q|\theta)\right]^T\},\tag{7}$$

where $\nabla[\cdot]$ denotes the gradient operator and $[\cdot]^T$ denotes the transpose operation of the matrix. According to the literature²², the CRLB of the target localization error can be expressed as:

$$CLRB = \operatorname{tr}\{\mathbf{C}_{CLRB}\} = tr\{[\mathbf{J}(\theta)]^{-1}\}.$$
(8)

Define ϕ is a function of θ , where $\phi = [\tau_{11}, \cdots, \tau_{mn}, \cdots, \tau_{MN}]$. According to the chain derivation rule, the $\mathbf{J}(\theta)$ in Eq. (5-8) can be rewritten as:

$$\mathbf{J}(\theta) = \mathbf{P}\mathbf{J}(\phi)\mathbf{P}^{T},
\mathbf{P} = \frac{\partial\phi}{\partial\theta},$$
(9)

where $\mathbf{J}(\phi) \in \mathbb{R}^{MN \times MN}$ is the FIM with respect to ϕ , which can be expressed as follows:

$$\mathbf{J}(\phi) = -E \begin{bmatrix} \frac{\partial^2 \log p(r|\phi)}{\partial \tau_{mn} \partial \tau_{m'n'}} \end{bmatrix} = \begin{cases} \frac{2}{K_R \sigma_w^2} 4\pi^2 G_{R_n}^2 \alpha_{mn}^2 P_m G_{T_m}^2 |h_{mn}|^2 \beta_m^2 & mn = m'n', \\ 0 & mn \neq m'n', \end{cases}$$
(10)

The matrix $\mathbf{P} \in R^{3 imes MN}$ can be expressed as:

$$\mathbf{P} = \frac{\partial \phi}{\partial \theta} = \left[\frac{\partial \tau}{\partial x}, \frac{\partial \tau}{\partial y}, \frac{\partial \tau}{\partial z}\right]^T,\tag{11}$$

where $\frac{\partial \tau}{\partial x}$ denotes the derivation of each element of τ with respect to x, and the derivation is shown below:

$$\frac{\partial \tau_{mn}}{\partial x} = \frac{x - x_m}{c\mathbb{R}(T_m, Q)} + \frac{x - x_n}{c\mathbb{R}(R_n, Q)}, \\ \frac{\partial \tau_{mn}}{\partial y} = \frac{y - y_m}{c\mathbb{R}(T_m, Q)} + \frac{y - y_n}{c\mathbb{R}(R_n, Q)}, \\ \frac{\partial \tau_{mn}}{\partial z} = \frac{z - z_m}{c\mathbb{R}(T_m, Q)} + \frac{z - z_n}{c\mathbb{R}(R_n, Q)}.$$
(12)

For simplicity, we define

$$g_{xm} \stackrel{\Delta}{=} \frac{x - x_m}{\mathbb{R}(T_m, Q)}, g_{xn} \stackrel{\Delta}{=} \frac{x - x_n}{\mathbb{R}(R_n, Q)}, g_{ym} \stackrel{\Delta}{=} \frac{y - y_m}{\mathbb{R}(T_m, Q)}, g_{yn} \stackrel{\Delta}{=} \frac{y - y_n}{\mathbb{R}(R_n, Q)}, g_{zm} \stackrel{\Delta}{=} \frac{z - z_m}{\mathbb{R}(T_m, Q)}, g_{zn} \stackrel{\Delta}{=} \frac{z - z_n}{\mathbb{R}(R_n, Q)}.$$
(13)

and Equation 11 can be transformed into:

$$\frac{\partial \tau_{mn}}{\partial x} = \frac{1}{c} \left(g_{xm} + g_{xn} \right), \frac{\partial \tau_{mn}}{\partial y} = \frac{1}{c} \left(g_{ym} + g_{yn} \right), \frac{\partial \tau_{mn}}{\partial z} = \frac{1}{c} \left(g_{zm} + g_{zn} \right). \tag{14}$$

Combining Eqs. 9, 10, and 11, Equation 8 can be simplified as:

$$\mathbf{C}_{CRLB} = \left(\mathbf{P}\mathbf{J}(\phi)\mathbf{P}^{T}\right)^{-1} = \xi \begin{bmatrix} g_{x} & g_{xy} & g_{xz} \\ g_{xy} & g_{y} & g_{yz} \\ g_{xz} & g_{yz} & g_{z} \end{bmatrix}^{-1},$$
(15)

where

$$\begin{aligned} \xi &= \left(\frac{8\pi^2}{K_R \sigma_w^2 c^2}\right)^{-1}, \\ g_x &= \sum_{m=1}^M \sum_{n=1}^N \alpha_{mn}^2 P_m G_{R_n}^2 G_{T_m}^2 |h_{mn}|^2 \beta^2 (g_{xm} + g_{xn})^2, \\ g_y &= \sum_{m=1}^M \sum_{n=1}^N \alpha_{mn}^2 P_m G_{R_n}^2 G_{T_m}^2 |h_{mn}|^2 \beta^2 (g_{ym} + g_{yn})^2, \\ g_z &= \sum_{m=1}^M \sum_{n=1}^N \alpha_{mn}^2 P_m G_{R_n}^2 G_{T_m}^2 |h_{mn}|^2 \beta^2 (g_{zm} + g_{zn})^2, \\ g_{xz} &= \sum_{m=1}^M \sum_{n=1}^N \alpha_{mn}^2 P_m G_{R_n}^2 G_{T_m}^2 |h_{mn}|^2 \beta^2 (g_{xm} + g_{xn}) (g_{zm} + g_{zn}), \\ g_{yy} &= \sum_{m=1}^M \sum_{n=1}^N \alpha_{mn}^2 P_m G_{R_n}^2 G_{T_m}^2 |h_{mn}|^2 \beta^2 (g_{xm} + g_{xn}) (g_{ym} + g_{yn}), \\ g_{yz} &= \sum_{m=1}^M \sum_{n=1}^N \alpha_{mn}^2 P_m G_{R_n}^2 G_{T_m}^2 |h_{mn}|^2 \beta^2 (g_{ym} + g_{yn}) (g_{zm} + g_{zn}). \end{aligned}$$

Thus, the CRLB of the target localization error can be expressed as:

$$CRLB = \sigma_{xCRB}^{2} + \sigma_{yCRB}^{2} + \sigma_{zCRB}^{2},$$

$$\sigma_{xCRB}^{2} = \frac{\xi}{|\mathbf{G}|} \left(g_{y} * g_{z} - g_{yz} g_{yz} \right),$$

$$\sigma_{yCRB}^{2} = \frac{\xi}{|\mathbf{G}|} \left(g_{x} * g_{z} - g_{xz} g_{xz} \right),$$

$$\sigma_{zCRB}^{2} = \frac{\xi}{|\mathbf{G}|} \left(g_{x} * g_{y} - g_{xy} g_{xy} \right).$$
(17)

where $|\mathbf{G}|$ is the determinant of the matrix G.

Methods

Optimiztion model of the resource allocation

In some practical scenarios, the radar system's total transmitted power is limited. Thus, the transmitted power allocation is necessary to improve the radar system's target localization performance. Previous studies^{11,12} demonstrate that CRLB can quantify the target localization performance of the MIMO radar system. Therefore, we establish the optimization model of the resource allocation to minimize the target localization error:

$$\min_{\mathbf{P}, \mathbf{w}, \mathbf{v}} CRLB + \|\mathbf{w}\|_{0},$$
s.t $P_{m} \leq P_{\max}, m = 1, 2, ..., M,$

$$\mathbf{w}^{T} \mathbf{P} \leq P_{\lim},$$

$$\|\mathbf{v}\|_{0} = \lambda,$$
(18)

where $\mathbf{P} = [P_1, P_2, ..., P_M]^T$ represents the assigned transmitted power of each transmit radar. $\mathbf{w} = [w_1, w_2, ..., w_M]^T$ represents the transmit radar selection vector and $w_m \in \{0, 1\}$ in which '1' indicates the transmit radar is selected while '0' indicates not selected. $\|\mathbf{w}\|_0$ represents the non-zero number of elements in the transmit radar selection vector. Similarly, $\mathbf{v} = [v_1, v_2, ..., v_N]^T$ represents the receive radar selection vector. λ denotes the constrained number of receive radar. P_{max} is the maximum transmitted power of each transmit radar, and P_{lim} is the total transmitted power in the radar system.

A hybrid action space reinforcement learning

The above resource allocation model is a typical combinatorial optimization problem with a relatively large solution space. In this paper, we transform the combinatorial optimization problem into a sequential decision problem and employ the reinforcement learning method to find the solution.

The core of reinforcement learning consists of states, actions, and rewards. In our radar resource allocation scenario, the state space consists of the transmit radars' power and the receive radars' state. The action space consists of allocating transmitted power to the transmit radars and selecting receive radars. The rewards are related to the CRLB. The action space is a hybrid action space containing both continuous attributes (i.e., transmitted power allocation to the transmit radars) and discrete attributes (i.e., receive radar selection). The typical reinforcement learning methods (e.g., DQN²³, DDPG²⁴, etc.) can only deal with discrete and continuous actions. Some researchers take a divide-and-conquer approach when extending them to the hybrid action space. Q-PAMDP²⁵ and Deep MAHHQN²⁶ first solve for optimal discrete actions and then solve the corresponding optimal continuous actions. PADDPG²⁷ and P-DQN²⁸ solve the optimal continuous actions corresponding to all discrete actions and then solve the optimal discrete actions. However, making discrete actions continuous leads to a more complex action space, while making continuous actions high-dimensional discrete introduces accuracy errors and increases the computational burden.

To address the above problems, we propose A Hybrid Action Space-based Reinforcement Learning (HAS-RL) method, which maps continuous and discrete actions to the policy space through two separate branches. Our proposed HAS-RL method is mainly based on the PPO algorithm²⁹, which also employs the clipping mechanism and advantage function estimation to stabilize the training process and improve sample efficiency.

Agent

The agent in the proposed HAS-RL method is the actor-critic structure³⁰, which consists of two sub-actor networks and a global critic network. The sub-actor network decides on transmit radar power allocation and receive radar selection according to the current radar state. The critic network evaluates the superiority of the policy. Fig. 2 depicts the structure of the agent. The sub-actor network decomposes the complex action space into discrete and continuous sub-spaces, each handled by a sub-actor network. The two sub-actors learn the policy that guides the action selection at its corresponding action space. These policies work together to determine the actions of the agent. The loss functions of these two branches are shown below:

$$L_d^{\text{CLIP}}(\theta_d) = \widehat{E}_t[\min(r_t^d(\theta_d)\hat{A}_t, \operatorname{clip}(r_t^d(\theta_d), 1-, 1+)\hat{A}_t)],$$

$$L_c^{\text{CLIP}}(\theta_c) = \widehat{E}_t[\min(r_t^c(\theta_c)\hat{A}_t, \operatorname{clip}(r_t^c(\theta_c), 1-, 1+)\hat{A}_t)],$$
(19)

where $r_t^d(\theta_d)$ stands for $\frac{\pi_{\theta_d}(a|s_t)}{\pi_{\theta_d}(\mathrm{old})(a|s_t)}$, $r_t^c(\theta_c)$ stands for $\frac{\pi_{\theta_c}(x_a|s_t)}{\pi_{\theta_c}(\mathrm{old})(x_a|s_t)}$, and \hat{A}_t is the advantage function.

State

The state is a combination of the transmitted power of transmit radars and the selected state of receive radars. For the *i*-th transmit radar, its state value is continuous $S_i \in [0, 3000]$, indicating its transmitted power. When the state value is 0, it means this transmit radar is off. For the *j*-th receive radar, its state value is discrete $S_j \in \{0, 1\}$, indicating whether to turn on this radar.

Action

The action contains the transmit radar power allocation and the receive radar selection, which is a hybrid action space. For the transmit radar, its action space is continuous $A_i \in [-\delta, \delta]$, and we set it as an additive or



Fig. 2. Schematic diagram of the proposed HAS-RL method.

subtractive change to the transmitted power. For the receive radar, its action space is discrete $A_j \in \{0, 1\}$, where 0 indicates keeping the current receive radar state and 1 indicates flipping the current receive radar state.

Reward

Traditional reinforcement learning methods tend to have a single reward. However, the single reward design can easily lead the agent to fall into a local optimum. Literature⁶ points out that the average power allocation is not an optimal solution for the radar system. Thus, we design two rewards by taking the CRLB under the average power allocation as the threshold, termed $CRLB_{avg}$. When the current CRLB is larger than the threshold, the agent deviates from the correct optimization direction. Thus, we set the reward to encourage the agent to explore other possible solutions. When the current CRLB is smaller than the threshold, the agent explores in the correct direction. Thus, the reward with constraint mechanism is designed to to gradually guide the agent toward the optimal strategy in that local area. This two-stage reward strategy provides a more precise learning guidance for the agent in different states.

Specifically, the reward in the first phase encourages the agent to explore as large a solution area as possible. Therefore, the reward in this phase does not employ penalties but uses the rescale strategy to legalize the state that breaks the total power constraint. The rescale strategy can avoid too many penalties that may mislead the learning direction of the agent and constrain its exploration. The reward function in the first stage is shown in Eq. 20:

$$r_1 = D_{scale} \times \alpha + \|\mathbf{w}\|_0 \quad \text{if} \quad CRLB_t \ge CRLB_{avg},\tag{20}$$

where $D_{scale} = \frac{CRLB_{t-1} - CRLB_t}{CRLB_t}$ denotes the superiority of state update at different time steps. α is set as 500 to control the exploration range of the agent, $CRLB_{avg}$ is the threshold, and $\|\mathbf{w}\|_0$ is the number of transmit radar.

In the second stage, the agent explores the desired optimization direction. Therefore, we introduce the soft constraint mechanism, which penalties the illegal power allocation to ensure the exploration of the agent will not deviate from the desired optimization direction.

$$r_2 = D + Penalty_{radar} + \|\mathbf{w}\|_0 \quad \text{if} \quad CRLB_t < CRLB_{avg}, \tag{21}$$

where $D = CRLB_{t-1} - CRLB_t$ denotes the superiority of state update at different time steps. In the second stage, the agent is already close to the desired optimization direction, and the CRLB changes slightly in different time steps. Thus, we adopt the difference in CRLB as the primary reward to guide the agent's exploration direction. *Penalty*_{radar} denotes the penalty for the illegal power allocation, which is defined as follows:

$$Penalty_{all_radar} = -|P_{total} - P_{lim}| \quad \text{if} \quad P \neq P_{lim},$$

$$Penalty_{per_radar} = -\sum_{i=1} min(|P_i - P_{max}|, |P_i|) \quad \text{if} \quad P_i \notin [0, P_{max}],$$
(22)

where P_{total} denotes the current total transmitted power, P_{lim} denotes the total transmitted power constraint, P_i denotes the power of the *i*-th transmit radar, and P_{max} denotes the maximum transmitted power of a single transmit radar.

Simulations and analysis Simulation setup

Radar system simulation scenario

Two different radar system simulation scenarios are introduced in this section. Fig. 3 illustrates the first simulation scenario, consisting of 8 transmit radars (red dots) and 10 receive radars (blue dots). Fig. 4 illustrates the second simulation scenario, which mainly consists of 5 transmit radars (red dots) and 7 receive radars (blue dots). The target in both two simulation scenarios is located in [350 km, 550 km, 9 km]. For the transmit radar, the power constraint is set as 3000W, the bandwidth is set as 30 MHZ, and the radar gain is set as 33 dB. As for the receive radar, the radar gain is also 33 dB. The target reflection cross section is set as the same value (i.e., $h_{mn} = 10$) for different paths.

Hyper-parameter setting in our proposed method

The proposed HAS-RL method is trained for 500,000 steps, and the learning rate is set to 0.0003. The clipping range is set to 0.2 to maintain the stability of the strategy update. The entropy coefficient is set to 0.01 to balance the exploration and utilization of the agent. The discount factor is taken as 0.99 to trade off the long-term and short-term rewards of the agent. The advantage estimation coefficient is s as 0.95 to encourage the agent to consider more future information when estimating the advantage function.

Comparison experiments settings

In this paper, we select the Average Power Allocation (APA) method, the Random Power Allocation(RPA) method and the Simulated Annealing (SA) algorithm for comparison. The SA algorithm is a heuristic search algorithm that solves the optimization problem by simulating the annealing process in physics. We set the initial temperature of the SA algorithm as $T_{\text{max}} = 20$, the annealing rate is taken as an exponential decrease with a decreased coefficient of 0.8, the termination temperature is set to $T_{\text{min}} = 0.1$, and the number of iterations at each temperature T is set to L=50.



Fig. 3. The distribution of target and radars in the first radar system simulation scenario.



Fig. 4. The distribution of target and radars in the second radar system simulation scenario.

Results and analysis

The first radar system simulation scenario

In the first simulation scenario, the distributed MIMO radar system contains 8 transmit radars and 10 receive radars, as shown in Fig. 3. We set the total transmitted power of the radar system to 16000W and constrain the number of receive radars to 8. Since the APA method applies uniform transmitted power for all transmit radars, we only need to consider the receive radar selection. In this simulation scenario, the solution space for the APA method is $C_{10}^8 = 45$, and we employ the exhaustive method to obtain the minimum CRLB. For the RPA method, we randomly conduct experiments 100 times to obtain the CRLBs under different radar states. For the SA method and the proposed HAS-RL method, we repeat the experiment 10 times.

Table 1 shows the comparison results between different allocation strategies. Compared to the APA method, the other three methods can obtain a smaller minimum CRLB, which verifies the importance of the power allocation for the radar system. However, the average CRLB of the RPA method is higher than that of other methods due to its randomness. This is because the RPA method may select irrational radar states, substantially increasing the target localization error. On the contrary, the SA and the proposed HSA-RL method achieve more stable results in the multiple repeat experiments and further improve the target localization performance. Compared with other power allocation strategies, our proposed HSA-RL method achieves the best localization performance, demonstrating its effectiveness.

In this simulation scenario, the total transmitted power of the radar system is 16000 W, and the transmitted power of each individual transmit radar is limited to 3000 W. Hence, the radar system at least contains 6 transmit radars. The 2nd to 4th rows of Table 1 report the comparision results of the RPA method under different numbers of transmit radars. We can see that selecting only 6 transmit radars for power allocation achieves better target localization performance than other settings for RPA. This is because selecting fewer transmit radars for power allocation make the transmitted power more concentrated on critical transmit radars (those closer to the target), leading to better target localization performance. However, selecting fewer transmit radars for power allocation may also ignore some important transmit radars, resulting in significant deviations among multiple repeat experiments.

	CRLB	
Method	Mean value	Minimum value
APA	2347.65 m	1197.26 m
RPA	4011.05 m	1137.44 m
(with 8 transimitter radars)	4011.05 111	
RPA	5777 02 ***	1052.28 m
(with 7 transimitter radars)	3777.93 111	
RPA	5810.04 m	1026 20 m
(with 6 transimitter radars)	3810.04 111	1030.30 III
SA	992.40 m	988.27 m
15 Ours	15 982.08 m	15 980.65 m

Table 1. Comparison performance between different power allocation methods in the first simulation scenario. The bold item indicates the optimal results.

Method	Transmitter radar power allocation	Reciver radar selection
APA	[2000 W, 2000 W]	[1, 1, 0, 1, 1, 1, 1, 1, 1, 0]
RPA	[2020 W 2260 W 1144 W 2259 W 2624 W 685 W 2702 W 2199 W]	[1, 1, 0, 1, 1, 1, 1, 1, 1, 0]
(with 8 transimitter radars)	[2030 W, 2209 W, 1144 W, 2238 W, 2034 W, 083 W, 2792 W, 2188 W]	
RPA	[2000 M 1701 M 0 M 2002 ··· 2000 ··· 277 ··· 277 ··· 2272 ···	[1, 1, 1, 1, 0, 1, 1, 1, 1, 0]
(with 7 transimitter radars)	[5000 W, 1/01 W, 0 W, 2992 W, 5000 W, 577 W, 2577 W, 2555 W]	
RPA	[2000 W 2118 W 2000 W 2000 W 0 W 0 W 0 W 1882 W 2000 W]	[1 1 1 1 0 0 1 1 1 1]
(with 6 transimitter radars)	[5000 W, 2118 W, 5000 W, 5000 W, 0 W, 0 W, 1882 W, 5000 W]	[1, 1, 1, 1, 0, 0, 1, 1, 1, 1]
SA	[3000 W, 3000 W, 0 W, 3000 W, 1252 W, 0 W, 2748 W, 3000 W]	[0, 1, 0, 1, 1, 1, 1, 1, 1, 1]
Ours	[3000 W, 3000 W, 0 W, 3000 W, 1000 W, 0 W, 3000 W, 3000 W]	[1, 0, 1, 1, 1, 1, 1, 1, 1, 0]

Table 2. Detailed allocation under different power allocation strategies in the first simulation scenario. The power allocation results are rounded to the nearest integer.

.....

Furthermore, Table 2 shows the detailed power allocation results corresponding to the minimum CRLB value under different allocation strategies. We can see that the SA method and the proposed HSA-RL method tend to allocate more transmitted power to T_0 , T_1 , T_3 , T_6 , and T_7 , since these transmit radars are closer to the target, while the T_2 , T_4 , and T_5 transmit radars are far away from the target.

The second radar system simulation scenario

In the second simulation scenario, the distributed MIMO radar system contains 5 transmit radars and 7 receive radars, as shown in Fig. 4. We set the total transmitted power of the radar system to 10000W and constrain the number of receive radars to 5. In this simulation scenario, the solution space for the APA method is $C_7^5 = 21$, and the exhaustive method is adopted to obtain the minimum CRLB. For the RPA method, SA method and the proposed HAS-RL method, we still follow the same setting as the first simulation scenario.

Table 3 demonstrates the comparison results between different power allocation methods in the second simulation scenario. We can see that the proposed HAS-RL method still achieves the smallest CRLB value, demonstrating the effectiveness of our proposed method again. In addition, compared with the first simulation scenario, the second simulation scenario contains fewer transmit and receive radars, resulting in poor target localization performance. Table 4 shows the detailed power allocation results corresponding to the minimum CRLB value under different allocation strategies.

Discussion

In this section, we discuss the differences and advantages of the proposed HAS-RL method over the decomposition optimization methods and SA methods.

Comparison with decomposition optimization methods

The typical paradigm of decomposition optimization methods is to transform the joint resource optimization problem into sub-optimization problems and solve them step by step. Ma et al.¹⁴ transformed the joint optimization problem of transmit radar selection and transmitted power allocation into two sub-optimization problems by pre-setting the number of selected transmit radars. Xie et al.¹⁷ performed the transmit power allocation by sequentially increasing the number of selected transmit radars and achieved the resource allocation results until the objective function no longer decreased. In contrast, we model the joint transmitted power and transmit radar selection problem as a single optimization problem of transmitted power, avoiding the additional hyperparameter settings and complex solution steps.

	CRLB	
Method	Mean value	Minimum value
APA	6252.80 m	2025.77 m
RPA	(200.91 m	1794.85 m
(with 5 transimitter radars)	0390.81 111	
RPA	4609 64 m	1757.62 m
(with 4 transimitter radars)	4009.04 111	
SA	1799.33 m	1756.82 m
Ours	1752.55 m	1751.14 m

Table 3. Comparison performance between different power allocation methods in the second simulation scenario. The bold item indicates the optimal results.

Method	Transmitter radar power allocation	Reciver radar selection
APA	[2000 W, 2000 W, 2000 W, 2000 W, 2000 W]	[1, 0, 1, 1, 0, 1, 1]
RPA	[2000 W 296 W 2000 W 2616 W 009 W]	[1, 1, 1, 0, 1, 1, 0]
(with 5 transimitter radars)	[3000 W, 380 W, 3000 W, 2010 W, 998 W]	
RPA	[2000 W 0 W 2000 W 2000 W 1000 W]	[1, 1, 1, 1, 0, 0, 1]
(with 4 transimitter radars)		
SA	[3000 W, 498 W, 3000 W, 3000 W, 502 W]	[1, 1, 1, 1, 0, 0, 1]
Ours	[3000 W, 1000 W, 3000 W, 3000 W, 0 W]	[1, 1, 0, 1, 0, 1, 1]

Table 4. Detailed allocation under different power allocation strategies in the second simulation scenario. The power allocation results are rounded to the nearest integer.

	Mean value of CRLB		
Method	First simulation scenario	Second simulation scenario	
SA	992.40 m ± 4.224 m	1799.33 m ± 5.508 m	
Ours	982.08 m ± 0.799 m	1752.55 m ± 0.918 m	

Table 5. Stability performance comparison between the proposed method and simulated annealing (SA)algorithm in the two simulation scenarios. The bold item indicates the optimal results.

Comparison with SA method

As shown in the Table 5, the proposed HAS-RL method exhibits stronger robustness and achieves more stable results in multiple repeat experiments. Specifically, in the first simulation scenario, our proposed HAS-RL method exhibits a standard deviation of 0.799m, while that of the SA method is 4.224m. Our proposed method also maintains the consistent stability advantage in the second scenario. In terms of convergence speed, the SA method exhibits a faster convergence speed than the proposed HAS-RL method. In the first simulation scenario, the SA requires about 22 seconds to reach the convergence condition. While the proposed HAS-RL method takes about 12 minutes to finish the training phase, yielding satisfactory power allocation results.

Conclusion

This paper considers the resource allocation problem in distributed MIMO radar systems. We establish a constrained optimization model to minimize the target localization error under the constraints of transmitted power and the number of receive radars and propose a hybrid action space reinforcement learning method to solve it. Experiments in different simulation scenarios show that our proposed method can effectively allocate resources for better target localization performance.

Our future work will mainly focus on the following two aspects: (1) introducing more metrics to comprehensively evaluate the target localization performance of the MIMO radar system, and (2) improving the convergence speed of the proposed HSA-RL method.

Data availability

The datasets used and/or analysed during the current study available at: "https://figshare.com/articles/dataset/R adar_system_simulation_scenario_zip/28358531".

Received: 6 February 2025; Accepted: 15 May 2025 Published online: 04 June 2025

References

- 1. Bekkerman, I. & Tabrikian, J. Target detection and localization using mimo radars and sonars. *IEEE Trans. Signal Process.* 54, 3873–3883 (2006).
- Zhu, J., Liu, W., Zhang, X., Lyu, F. & Guo, Z. Antenna placement optimization for distributed mimo radar based on a reinforcement learning algorithm. Sci. Rep. 13, 17487 (2023).
- 3. Wang, P. & Li, H. Target detection with imperfect waveform separation in distributed mimo radar. *IEEE Trans. Signal Process.* 68, 793–807 (2020).
- 4. Liggins, M. II., Hall, D. & Llinas, J. Handbook of multisensor data fusion: Theory and practice (CRC Press, 2017).
- 5. He, Q., Blum, R. S. & Haimovich, A. M. Noncoherent mimo radar for location and velocity estimation: More antennas means better performance. *IEEE Trans. Signal Process.* 58, 3661–3680 (2010).
- Godrich, H., Petropulu, A. P. & Poor, H. V. Power allocation strategies for target localization in distributed multiple-radar architectures. *IEEE Trans. Signal Process.* 59, 3226–3240 (2011).
- Godrich, H., Petropulu, A. & Poor, H. V. Optimal power allocation in distributed multiple-radar configurations. In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2492–2495 IEEE, (2011).
- Feng, H.-Z., Liu, H.-W., Yan, J.-K., Dai, F.-Z. & Fang, M. A fast efficient power allocation algorithm for target localization in cognitive distributed multiple radar systems. *Signal Process.* 127, 100–116 (2016).
- 9. Shen, Y., Dai, W. & Win, M. Z. Power optimization for network localization. IEEE/ACM Trans. Netw. 22, 1337-1350 (2013).
- Shi, C., Wang, Y., Wang, F., Salous, S. & Zhou, J. Power resource allocation scheme for distributed mimo dual-function radarcommunication system based on low probability of intercept. *Digital Signal Process.* 106, 102850 (2020).
- 11. Guo, J., Tao, H. & Shi, L. Sensor selection of distributed mimo radar for target localization. In 2021 CIE International Conference on Radar (Radar), 2160–2163 IEEE, (2021).
- 12. Guo, J. & Tao, H. Resource allocation scheme of netted radar system for target localisation. *IET Radar Sonar Navigation* 17, 1456–1468 (2023).
- 13. Zhang, H. et al. Joint power, bandwidth, and subchannel allocation in a uav-assisted dfrc network. *IEEE Internet of Things Journal* 1–17, https://doi.org/10.1109/JIOT.2024.3522181 (2025).
- 14. Ma, B., Chen, H., Sun, B. & Xiao, H. A joint scheme of antenna selection and power allocation for localization in mimo radar sensor networks. *IEEE Commun. Lett.* 18, 2225–2228 (2014).
- Yang, S., Zheng Naè, L. Y. & Xiukun, R. Resource allocation approach for target localization in distributed mlmoradar sensor networks. J. Syst. Eng. Electron. 39, 304–309 (2017).
- Yi, J., Wan, X., Leung, H. & Lü, M. Joint placement of transmitters and receivers for distributed mimo radars. *IEEE Trans. Aerosp. Electron. Syst.* 53, 122–134 (2017).
- 17. Xie, M., Yi, W., Kirubarajan, T. & Kong, L. Joint node selection and power allocation strategy for multitarget tracking in decentralized radar networks. *IEEE Trans. Signal Process.* 66, 729–743 (2017).
- Radmard, M., Chitgarha, M. M., Majd, M. N. & Nayebi, M. M. Antenna placement and power allocation optimization in mimo detection. *IEEE Trans. Aerosp. Electron. Syst.* 50, 1468–1478 (2014).
- Shi, C., Ding, L., Wang, F., Salous, S. & Zhou, J. Joint target assignment and resource optimization framework for multitarget tracking in phased array radar network. *IEEE Syst. J.* 15, 4379–4390 (2021).
- Li, Z., Xie, J., Liu, W., Zhang, H. & Xiang, H. Transmit antenna selection and power allocation for joint multi-target localization and discrimination in mimo radar with distributed antennas under deception jamming. *Remote Sens.* 14, 3904 (2022).
- 21. Zhang, H., Liu, W., Zhang, Q. & Liu, B. Joint customer assignment, power allocation, and subchannel allocation in a uav-based joint radar and communication network. *IEEE Internet Things J.* **11**, 29643–29660 (2024).
- 22. Guo, J. & Tao, H. Cramer-Rao lower bounds of target positioning estimate in netted radar system. *Digital Signal Process.* 118, 103222 (2021).
- 23. Mnih, V. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602 (2013).
- 24. Lillicrap, T. Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971 (2015).
- 25. Masson, W., Ranchod, P. & Konidaris, G. Reinforcement learning with parameterized actions. In *Proceedings of the AAAI conference* on artificial intelligence, vol. 30 (2016).
- 26. Fu, H. et al. Deep multi-agent reinforcement learning with discrete-continuous hybrid action spaces. arXiv preprint arXiv:1903.04959 (2019).
- 27. Hausknecht, M. & Stone, P. Deep reinforcement learning in parameterized action space. arXiv preprint arXiv:1511.04143 (2015).
- Xiong, J. et al. Parametrized deep q-networks learning: Reinforcement learning with discrete-continuous hybrid action space. arXiv preprint arXiv:1810.06394 (2018).
- 29. Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017).
- 30. Wang, Z. et al. Sample efficient actor-critic with experience replay. arXiv preprint arXiv:1611.01224 (2016).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62276197 and Grant 62171332.

Author contributions

J. Z. made significant contributions to the conception and analysis of the proposed method and wrote the main manuscript. W. L. planned the experiments and established simulation scenarios. F. L. and S. L. contributed to the data processing and experiment results analysis. T. Z provided suggestions on experiment analysis and manuscript writing.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025