Deep Reinforcement Learning-based Multi-Target Tracking and Association in Cooperative Millimeter-Wave ISAC Systems

Anonymous Authors[1]

Abstract

This paper investigates the multi-target tracking (MTT) problem in a cooperative millimeter-wave (mmWave) integrated sensing and communication (ISAC) system, where distributed base stations (BSs) emit directional beams to probe the states of associated targets. Owing to the complicated environment and target mobilities, the target-BS associations need to be dynamically adjusted to facilitate stable MTT performance during the tracking interval. In this work, we propose a unified deep reinforcement learning (DRL)-based framework to perform joint target state estimation, target-BS association adjustment, and beam prediction tasks. Specifically, a dynamic graph neural network (DGNN) is first developed to capture the spatio-temporal features among targets and perform target state estimation. In particular, an advantage actor-critic (A2C)-based controller is then proposed for flexible target-BS association adjustment and beam prediction under a novel auto-regressive policy, which ensures that the target-BS association constraints can be strictly satisfied. Numerical results demonstrate that the proposed scheme can adaptively update associations in response to target mobility, thereby significantly reducing the tracking error.

1. Introduction

Integrated sensing and communication (ISAC) has been identified as one of the six key use cases in the sixth-generation (6G) cellular network, which aims to seamlessly incorporate sensing functionalities into communication-centric cellular systems (Zheng et al., 2019; Liu et al., 2020; Zhang et al., 2021; Hu et al., 2026a). Due to the sufficiently large bandwidth, the millimeter-wave (mmWave) band is promising for 6G networks, as it boosts both high-throughput communication and high-resolution sensing performance. A fundamental task in ISAC is multi-target tracking (MTT), which aims to continuously monitor the kinematic states of multiple targets over a long time interval. Existing studies have investigated the MTT problem in single base station (BS) scenarios (Li et al., 2026a;b). However, relying on a single BS for sensing faces severe challenges including restricted sensing coverage and limited parameter estimation accuracy due to severe occlusion and path loss. To address these issues, recent research turns to focus on the networked ISAC system, where multiple BSs perform cooperative signal processing to improve the target sensing performance (Zhu et al., 2026; Hu et al., 2026b).

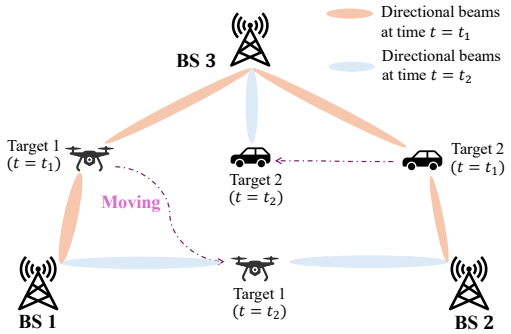

Figure 1. System model for multi-target tracking in the cooperative mmWave ISAC system, where the BSs are associated with appropriate targets in each time block and emit directional beams for tracking.

This work focuses on investigating the MTT problem in the cooperative mmWave ISAC system. Due to the large number of antennas but limited radio-frequency (RF) chains, each BS can only serve a limited number of targets simultaneously (Wang et al., 2022), meaning that the target-BS associations should be carefully adjusted during the tracking process, as shown in Fig. 1. In particular, there are two important tasks that should be addressed to achieve high-accuracy MTT in the considered system. **Task 1 - multi-target state estimation (MTSE):** Given measurements collected by distributed BSs, the objective is to accurately estimate the state of all targets. **Task 2 - target-BS association and beam prediction (TABP):** Given the his-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

torically estimated trajectories of targets, it is necessary to adaptively associate appropriate targets to each BS and predict the beam directions in the next time block such that we can maintain continuous tracking of all targets over a long time interval.

Recently, deep learning-based approaches have shown the powerful capability to solve the MTSE and TABP problems as two independent tasks. For instance, (Lian et al., 2026; Jiang et al., 2026) employ graph neural networks (GNNs) to perform cooperative positioning in static environments by extracting spatial correlations among BSs. Similarly, GNNs are also utilized in (Chan et al., 2026; Deng et al., 2023; Zhang et al., 2020) for the target-BS association based on instantaneous observations, while the association constraints cannot be guaranteed to be satisfied. In fact, the tasks of state estimation and target-BS association are highly coupled in the MTT problem. Specifically, the accuracy of the current state estimation determines the association strategy in the subsequent time block, which in turn affects the reliability of future measurements. However, existing works usually focus on decoupled approaches, which are inadequate to ensure reliable tracking performance due to the risk of error accumulation.

In this work, we propose a unified deep reinforcement learning framework to jointly solve the MTSE and TABP tasks of the MTT problem in the cooperative mmWave ISAC system. The proposed framework consists of a feature extractor and a decision controller, which jointly address the MTSE and TABP tasks in dynamic environments. Specifically, the feature extractor employs a dynamic graph neural network (DGNN) to capture the time-varying interactions among targets and BSs. Based on the extracted features, the controller then leverages the advantage actor-critic (A2C) scheme to determine strategies for target-BS association adjustment and beam prediction. In particular, an auto-regressive policy is developed to decompose the construction of the association matrix into a sequence of sub-actions, thereby handling the hard constraint imposed by the limited number of RF chains. Numerical results demonstrate that the proposed scheme achieves high-accuracy state estimation and adaptive association over a long interval under the target-BS association constraint.

2. System Model and Problem Formulation

2.1. System Model

This work considers a mmWave cell-free orthogonal frequency division multiplexing (OFDM) ISAC system in which M BSs communicate with U users and track I moving targets using the echo signals simultaneously. Each BS is equipped with a transmit uniform linear array (ULA) of N_{T} antennas and a receive ULA of N_{R} antennas. We assume that the transmit and receive antenna arrays are both equipped with N^{RF} RF chains, where $I \leq MN^{\mathrm{RF}}$. We consider a two-dimensional (2D) system, and the location of BS $m \in \mathcal{M} = \{1, \ldots, M\}$ is denoted as $\mathbf{p}_m = [p_{m,x}, p_{m,y}]^T$. Moreover, the mobility of each target $i \in \{1, \ldots, I\}$ is characterized by a "stop-and-go" model (Potter et al., 2010), where the location of each target is assumed to be fixed in one time block, but varying at different blocks. Let Q denote the number of time blocks to track the targets, and ΔT s denote the duration of each block. In the q-th block, the location of the target i is denoted as $\mathbf{u}_{i,q} = [x_{i,q}, y_{i,q}]^T$, $q = 1, \ldots, Q$, while the range and azimuth angle of the target i relative to the BS m are denoted by $d_{m,i,q} = \|\mathbf{u}_{i,q} - \mathbf{p}_m\|_2$ and $\theta_{m,i,q} = \arctan\left(\frac{y_{i,q}-p_{m,y}}{x_{i,q}-p_{m,x}}\right)$, respectively.

Let $\bar{K}$ denote the number of OFDM symbols within a block. We divide each block into two phases - tracking phase that consists of $K < \bar{K}$ OFDM symbols and communication phase that consists of $\bar{K} - K$ OFDM symbols. In this work, we focus on the multi-target tracking problem by leveraging the signals in the tracking phase. Due to the limited number of RF chains, each BS allocates at most KN^{RF} beams over the K OFDM symbols to track targets in each time block q. The set of targets that are assigned to BS m at block q is denoted as $\mathcal{S}_{m,q}$. Moreover, let L denote the number of sub-carriers. We assume that the BSs operate at non-overlapping frequency bands to avoid inter-cell interference, where the sub-carrier set for BS m at block q is defined as $\mathcal{L}_{m,q}$.

Let $\mathbf{a}_{\mathrm{T}}(\theta) = [1, \cdots, e^{j2\pi \frac{d_s}{\lambda}(N_{\mathrm{T}}-1)\cos\theta}]^T \in \mathbb{C}^{N_{\mathrm{T}}\times 1}$ and $\mathbf{a}_{\mathrm{R}}(\theta) = [1, \cdots, e^{j2\pi \frac{d_s}{\lambda}(N_{\mathrm{R}}-1)\cos\theta}]^T \in \mathbb{C}^{N_{\mathrm{R}}\times 1}$ denote the steering vectors of the transmit and receive ULAs of the BS towards the azimuth angle θ, respectively, with d_s being the antenna spacing and λ being the carrier wavelength. Therefore, the channel matrix between BS m and target i on the l-th sub-carrier during the k-th symbol in the q-th block, denoted by $\mathbf{H}_{m,i,q}^{k,l} \in \mathbb{C}^{N_{\mathrm{R}}\times N_{\mathrm{T}}}$, can be modeled as

$$\mathbf{H}_{m,i,q}^{k,l} = \alpha_{m,i,q} e^{j2\pi\nu_{m,i,q}kT_0} e^{-j2\pi l\Delta f \tau_{m,i,q}} \times \mathbf{a}_{\mathrm{R}}(\theta_{m,i,q})\mathbf{a}_{\mathrm{T}}^H(\theta_{m,i,q}), \forall m, q, k, l \in \mathcal{L}_{m,q}, \quad (1)$$

where $\alpha_{m,i,q}$ is the complex channel gain between BS m and target i, which depends on the radar cross-section (RCS) and the propagation path loss, $\tau_{m,i,q}$ and $\nu_{m,i,q}$ denote the time delay and Doppler frequency between BS m and target i, respectively, T_0 denotes the OFDM symbol duration, and Δf denotes the sub-carrier spacing.

During the k-th OFDM symbol in the tracking phase of the q-th block, the echo signal received by the BS m on the l-th sub-carrier is given as

$$\boldsymbol{\upsilon}_{m,q}^{k,l} = \sqrt{P}(\mathbf{F}_{m,q}^{\mathrm{RF},k})^H \sum_{i=1}^{I} \mathbf{H}_{m,i,q}^{k,l} \mathbf{W}_{m,q}^{\mathrm{RF},k} \mathbf{s}_{m,q}^{k,l} + (\mathbf{F}_{m,q}^{\mathrm{RF},k})^H \mathbf{n}_{m,q}^{k,l}, \quad \forall m, q, k, l \in \mathcal{L}_{m,q}, \quad (2)$$

where $\mathbf{s}_{m,q}^{k,l} \in \mathbb{C}^{N^{\mathrm{RF}}\times 1}$ denotes the m-th BS's transmit

symbol vector on the l-th sub-carrier of the k-th OFDM symbol in the q-th block, P denotes the transmit power, and $\mathbf{W}_{m,q}^{\mathrm{RF},k} \in \mathbb{C}^{N_{\mathrm{T}} \times N^{\mathrm{RF}}}$ and $\mathbf{F}_{m,q}^{\mathrm{RF},k} \in \mathbb{C}^{N_{\mathrm{R}} \times N^{\mathrm{RF}}}$ denote the frequency-flat phase shifter-based analog beamforming matrices of the k-th symbol in the q-th block. Here, $\mathbf{W}_{m,q}^{\mathrm{RF},k}$ and $\mathbf{F}_{m,q}^{\mathrm{RF},k}$ are designed by selecting N^{RF} beams from codebook $\mathcal{W}_{\mathrm{T}} = \left\{\mathbf{w}_{\mathrm{T}}^{(m)} \in \mathbb{C}^{N_{\mathrm{T}} \times 1}, m = 1, \ldots, M_{\mathrm{T}}\right\}$ and $\mathcal{W}_{\mathrm{R}} = \left\{\mathbf{w}_{\mathrm{R}}^{(m)} \in \mathbb{C}^{N_{\mathrm{R}} \times 1}, m = 1, \ldots, M_{\mathrm{R}}\right\}$, respectively, where each beam $\mathbf{w}_{\mathrm{T}}^{(m)}$ ($\mathbf{w}_{\mathrm{R}}^{(m)}$) is a pencil-like beam towards some pre-designed azimuth angle $\bar{\theta}_m$. In addition, $\mathbf{n}_{m,q}^{k,l} \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I}_{N_{\mathrm{R}}})$ denotes the additive white Gaussian noise (AWGN) vector at BS m with σ^2 being the power.

2.2. Problem Formulation

The MTT procedure in the considered system is performed as follows. In each time block q, each BS first locally estimates the sensing parameters of its assigned targets using the received signals in (2) by classical parameter estimation techniques (Stoica & Sharman, 1990; Schmidt, 1986). Specifically, the estimate at BS m for target $i \in \mathcal{S}_{m,q}$ in the q-th block is denoted by $\hat{\mathbf{z}}_{m,i,q} = [\hat{d}_{m,i,q}, \hat{\theta}_{m,i,q}]^T$ with $\hat{d}_{m,i,q}$ and $\hat{\theta}_{m,i,q}$ being the estimated range and angle of target i from BS m, respectively. Subsequently, the BSs forward their local estimates to the central processing unit (CPU) via fronthaul links. Then, the CPU performs tracking in two phases. **Phase 1 - multi-target state estimation:** The CPU jointly updates the kinematic state of each target i for the current block q, denoted by $\hat{\mathbf{u}}_{i,q} = [\hat{x}_{i,q}, \hat{y}_{i,q}]^T$ with $\hat{x}_{i,q}$ and $\hat{y}_{i,q}$ being the estimated coordinates of target i in the q-th block, and predicts the state of each target i for the next time block $q+1$, denoted by $\tilde{\mathbf{u}}_{i,q+1} = [\tilde{x}_{i,q+1}, \tilde{y}_{i,q+1}]^T$ with $\tilde{x}_{i,q+1}$ and $\tilde{y}_{i,q+1}$ being the predicted coordinates of target i in the $q+1$-th block. **Phase 2 - target-BS association and beam prediction:** Due to target mobility, the allocation of mmWave beams of each BS should be dynamically updated during the tracking process. Given the predicted target states for the next time block $q+1$ and the locations of the BSs, the CPU determines: i) the target-BS association matrix in the $q+1$-th block, denoted by $\mathbf{A}_{q+1} \in \{0,1\}^{M \times I}$, where its (m,i)-th element $\alpha_{m,i,q+1} = 1$ denotes that BS m is assigned to track target i in the $q+1$-th block, and $\alpha_{m,i,q+1} = 0$ otherwise. ii) the analog beamforming matrices for each BS used in the tracking phase of the $q+1$-th block, i.e., $\{\mathbf{W}_{m,q+1}^{\mathrm{RF},k}, \mathbf{F}_{m,q+1}^{\mathrm{RF},k}\}_{k=1,m=1}^{K,M}$.

In general, the tracking process in the CPU can be formulated as

$$\{\hat{\mathbf{u}}_{i,q}, \tilde{\mathbf{u}}_{i,q+1}\}_{i=1}^{I} = \mathcal{F}(\mathcal{Z}_q), \tag{3}$$

$$\begin{aligned}\{\mathbf{A}_{q+1}, &\{\mathbf{W}_{m,q+1}^{\mathrm{RF},k}, \mathbf{F}_{m,q+1}^{\mathrm{RF},k}\}_{k=1,m=1}^{K,M}\} \\ &= \mathcal{H}(\{\tilde{\mathbf{u}}_{i,q+1}\}_{i=1}^{I}, \{\mathbf{p}_m\}_{m=1}^{M}),\end{aligned} \tag{4}$$

where $\mathcal{Z}_q \triangleq \{\hat{\mathbf{z}}_{m,i,\tau} \mid i \in \mathcal{S}_{m,\tau}, m \in \mathcal{M}, \tau \in \{1, \ldots, q\}\}$ denotes the set of all historical measurements up to block q, $\mathcal{F}(\cdot)$ and $\mathcal{H}(\cdot)$ denote the mapping functions for Phase 1 and Phase 2, respectively. Finally, the CPU transmits the determined strategy to each BS for emitting signals in the tracking phase of the $q+1$-th block.

Therefore, the objective of the MTT problem is to jointly optimize the mapping functions $\mathcal{F}(\cdot)$ and $\mathcal{H}(\cdot)$ to minimize the overall tracking error of all targets during the tracking process, which can be formulated as

$$\min_{\mathcal{F}(\cdot),\mathcal{H}(\cdot)} \quad \frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\left[\sum_{i=1}^{I}\|\mathbf{u}_{i,q} - \hat{\mathbf{u}}_{i,q}\|^2\right] \tag{5a}$$

$$\text{s.t.} \quad (3), (4), \tag{5b}$$

$$\sum_{i=1}^{I}\alpha_{m,i,q} \le N^{\mathrm{RF}}, \quad \forall m, q, \tag{5c}$$

$$\sum_{m=1}^{M}\alpha_{m,i,q} \ge 1, \quad \forall i, q, \tag{5d}$$

$$\alpha_{m,i,q} \in \{0,1\}, \quad \forall m, i, q, \tag{5e}$$

$$\mathbf{W}_{m,q}^{\mathrm{RF},k}[:,n] \in \mathcal{W}_T, \mathbf{F}_{m,q}^{\mathrm{RF},k}[:,n] \in \mathcal{W}_R, \quad \forall m, k, q, n \in \{1, \ldots, N^{\mathrm{RF}}\}. \tag{5f}$$

Here, Constraint (5c) denotes the target-BS association constraint that the number of targets tracked by each BS should not exceed its available RF chains. Constraint (5d) guarantees that each target can be tracked by at least one BS. Constraint (5e) enforces the binary nature of the assignment variables, and (5f) enforces the analog beamforming vectors to predefined codebooks.

Due to the non-convex and combinatorial characteristics, it is challenging to derive the globally optimal solution for (5) via conventional optimization-based schemes. Moreover, (5) induces a Markov decision process (MDP) for MTT, wherein the existing GNN-based schemes are myopic and fail to capture the long-term impact of instantaneous decisions on future tracking accuracy. To solve this challenging sequential decision-making problem in complicated environments, we propose a tailored deep reinforcement learning (DRL) scheme to jointly learn the mapping functions $\mathcal{F}(\cdot)$ and $\mathcal{H}(\cdot)$ for optimizing long-term performance in the MTT problem.

3. DRL-based Method

In this section, we introduce a DRL framework for MTT in cooperative mmWave ISAC systems. The proposed approach consists of a DGNN-based feature extractor and an A2C-based decision controller, as shown in Fig. 2. Specifically, we first construct a dynamic bipartite graph to model the time-varying topology between targets and BSs, where nodes correspond to BSs and targets, while the edges represent the dynamic target-BS associations. Based on the

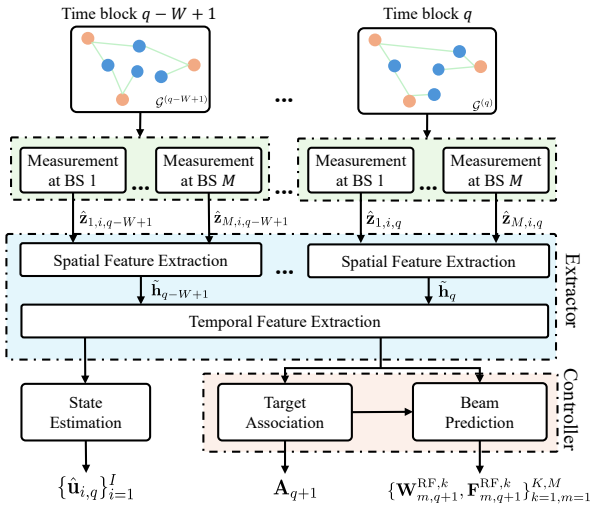

Figure 2. Diagram of the proposed DRL-based MTT framework.

graph, the extractor employs a spatio-temporal attention architecture to capture the spatial correlations within the network topology and the temporal dynamics of target mobility. Then, the controller leverages the A2C framework to proactively determine the target-BS association matrix and the selected beam indices for the upcoming time block. Within the controller, we design an auto-regressive policy that iteratively allocates each available RF chain to one selected target-BS pair to satisfy the target-BS association constraint. In the following, we introduce the proposed extractor and controller in detail, as well as the policy for training the network.

3.1. Feature Extractor

The extractor aims to capture the spatio-temporal features within the network topology and target dynamics. Specifically, we define the cell-free network in the q-th block as a bipartite graph $\mathcal{G}^{(q)} = (\mathcal{V}_{\text{BS}}, \mathcal{V}_{\text{TG}}, \mathcal{E}^{(q)})$, where $\mathcal{V}_{\text{BS}}$ and $\mathcal{V}_{\text{TG}}$ represents the set of BS nodes and target nodes, respectively, which remain consistent during the tracking process, and $\mathcal{E}^{(q)}$ is the set of edges defined by the target-BS association in the q-th block. Specifically, the node feature of BS $m \in \mathcal{V}_{\text{BS}}$, denoted by $\mathbf{x}_m$, is defined as its physical coordinates $\mathbf{p}_m$, i.e., $\mathbf{x}_m = \mathbf{p}_m$. The node feature of target $i \in \mathcal{V}_{\text{TG}}$ is initialized as a zero vector, which will be updated in the DGNN. The edge feature from BS m to target i in bock q, denoted by $\mathbf{e}_{m,i,q}$, is defined as the local distance and angular measurement to target i from BS m, i.e., $\mathbf{e}_{m,i,q} = \hat{\mathbf{z}}_{m,i,q}$. The proposed DGNN consists of a Spatial Attention Module (SAM) and a Temporal Attention Module (TAM), which will be introduced as follows.

Spatial Attention Module: The SAM consists of L message passing layers that aim to fuse the position-related measurements from each assigned BS to achieve networked localization. We define the message $\mathbf{b}_{m,i,q}$ from BS node m to target node i in block q as

$$\mathbf{b}_{m,i,q} \triangleq (\text{MLP}(\mathbf{x}_m) \parallel \text{MLP}(\mathbf{e}_{m,i,q})), \tag{6}$$

where $\parallel$ denotes the concatenation operation, $\text{MLP}(\cdot)$ denotes the multi-layer perceptron (MLP). In the l-th layer ($l = 1, \ldots, L$), the hidden feature of each target node i, denoted as $\mathbf{h}_{i,q}^{(l)} \in \mathbb{R}^{D\times 1}$ with D being the feature dimension, is iteratively updated by message passing and aggregation. Specifically, we first compute the attention weights between the current hidden state of the target and the incoming messages from neighboring BS nodes, which quantify the relative contribution of each BS measurement. These incoming messages are then aggregated via a weighted combination and added to the previous hidden state via a residual connection. The complete message passing process can be expressed as

$$\mathbf{h}_{i,q}^{(l)} = \text{LayerNorm}\left(\text{GELU}\left(\mathbf{h}_{i,q}^{(l-1)} + \sum_{m\in\mathcal{N}(i)} \alpha_{m,i,q}^{(l)} \mathbf{b}_{m,i,q}\right)\right), \tag{7}$$

where

$$\alpha_{m,i,q}^{(l)} \triangleq \frac{\exp\left(\text{ATTN}^{(l)}\left(\mathbf{h}_{i,q}^{(l-1)} \parallel \mathbf{b}_{m,i,q}\right)\right)}{\sum_{k\in\mathcal{N}(i)} \exp\left(\text{ATTN}^{(l)}\left(\mathbf{h}_{i,q}^{(l-1)} \parallel \mathbf{b}_{k,i,q}\right)\right)} \tag{8}$$

with $\mathcal{N}(i)$ denoting the set of neighboring BS nodes connected to the i-th target node, and $\text{ATTN}^{(l)}(\cdot)$ denoting the attention module in the l-th layer.

After L message passing updates, the final hidden features of targets are concatenated to form $\tilde{\mathbf{h}}_q = [(\mathbf{h}_{1,q}^{(L)})^T, \ldots, (\mathbf{h}_{I,q}^{(L)})^T]^T \in \mathbb{R}^{ID\times 1}$. This vector serves as the latent representation of the spatial topology in the q-th time block. Subsequently, the TAM is employed to capture the temporal dynamics of the graph across successive blocks.

Temporal Attention Module: To model temporal dependencies, we construct an input sequence utilizing a sliding window of historical spatial representations over the past W time blocks, coupled with a positional encoding matrix $\mathbf{P} \in \mathbb{R}^{ID\times W}$ to preserve the sequential order. The initial temporal embedding sequence $\mathbf{Z} \in \mathbb{R}^{ID\times W}$ is given by

$$\mathbf{Z} = [\tilde{\mathbf{h}}_{q-W+1}, \ldots, \tilde{\mathbf{h}}_q] + \mathbf{P}. \tag{9}$$

The temporal feature extraction is then performed by a Transformer encoder composed of two stacked multi-head Self-Attention (MHSA) layers. Specifically, the first layer is used to estimate the current states for all targets for the q-th block, i.e., $\{\hat{\mathbf{u}}_{i,q}\}_{i=1}^{I}$. The second layer then predicts the target states for the upcoming $q+1$-th block, i.e., $\{\tilde{\mathbf{u}}_{i,q+1}\}_{i=1}^{I}$.

3.2. Decision Controller

Following the extraction of spatio-temporal dependencies, the controller in the CPU adaptively determines the target-BS associations and selected emitting beams in the subsequent time block to maintain high-accuracy tracking. Within

the MDP framework, the CPU acts as the agent, interacting with the dynamic environment through the distributed BSs according to a learned policy and receiving rewards based on the resultant tracking performance.

However, due to the target-BS association constraint in (5c), generating a feasible association matrix in a single step is intractable. To overcome this challenge, we propose an auto-regressive policy to sequentially construct the target-BS associations. Specifically, given that each of the M BSs is equipped with N^{RF} RF chains, the overall association strategy is decomposed into $J = M \times N^{\mathrm{RF}}$ iterations. At each iteration, all valid target-BS pairs are scored based on the extracted spatio-temporal features and the real-time available resources. The target-BS pair yielding the highest selection probability is then selected as the action for that step. In the following, we first introduce the formulated MDP problem, and then introduce the proposed policy.

MDP Formulation: In the $q+1$-th time block, we define the state $s_{q+1} \in \mathcal{S}$ by the extracted latent feature of all targets from the extractor, i.e., $s_{q+1} = [\tilde{\mathbf{u}}_{1,q+1}^T, \tilde{\mathbf{u}}_{2,q+1}^T, \ldots, \tilde{\mathbf{u}}_{I,q+1}^T]^T$. The complete association action is defined as the set of sub-actions, denoted by $a_{q+1} = \{a_{q+1}^{(1)}, \ldots, a_{q+1}^{(J)}\}$, where each sub-action $a_{q+1}^{(j)} = (m^{*(j)}, i^{*(j)}) \in \{\mathcal{M} \times \mathcal{I}\}$ represents the association between BS $m^{*(j)}$ and target $i^{*(j)}$ at the j-th iteration. Accordingly, the relationship between the association variable $\alpha_{m,i,q+1}$ and the action set a_{q+1} can be expressed as

$$\alpha_{m^*,i^*,q+1} = \begin{cases} 1, & \text{if } (m^*, i^*) \in a_{q+1}, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The per-step reward r_{q+1} is evaluated after the action selection of the $q+1$-th block. In our networked tracking task, the immediate reward in the $q+1$-th block is defined as the negative summation of the localization errors across all targets, i.e., $r_{q+1} = -\sum_{i=1}^{I} \|\mathbf{u}_{i,q+1} - \hat{\mathbf{u}}_{i,q+1}\|^2$. Therefore, the cumulative discounted reward can be defined as $R = \sum_{q=0}^{Q-1} \gamma^q r_{q+1}$, where $\gamma \in (0, 1]$ is the discount factor. Note that a larger γ emphasizes long-term rewards and thus promotes policies that optimize long-term tracking performance, whereas a smaller γ leads to more myopic decision-making that prioritizes immediate performance.

Auto-Regressive Policy Design: To solve the formulated MDP, we develop the controller based on the Advantage Actor-Critic (A2C) architecture (Mnih et al., 2016), where an actor network learns a policy $\pi(a_{q+1} \mid s_{q+1})$ that can select the action a_{q+1} based on s_{q+1} to maximize the reward R, and a critic network evaluates the state value of the action for updating the policy.

Specifically, we can decouple the policy according to the chain rule (Bello et al., 2017):

$$\pi(a_{q+1} \mid s_{q+1}) = \prod_{j=1}^{J} \pi^{(j)}(a_{q+1}^{(j)} \mid s_{q+1}, a_{q+1}^{(1:j-1)}), \tag{11}$$

where $a_{q+1}^{(1:j-1)}$ denotes the previously selected BS-target pairs up to iteration $j-1$. Therefore, the actor network is designed by a MLP that operates J times in a block. During the j-th iteration ($j \in \{1, \ldots, J\}$), the input to the network consists of two components: i) the state s_{q+1}, which remains static in each iteration, and ii) a dynamic context vector $\mathbf{v}_{q+1}^{(j)} = [v_{q+1,1}^{(j)}, \ldots, v_{q+1,M}^{(j)}]^T \in \{0, \ldots, N^{\mathrm{RF}}\}^M$, where the m-th element $v_{q+1,m}^{(j)} = N^{\mathrm{RF}} - \sum_{k=1}^{j-1} \mathbb{I}\big(m^{*(k)} = m\big)$. This vector encodes the remaining resources for each BS at the j-th iteration and is updated at each iteration to make the network aware of the prior selections. Accordingly, the MLP outputs a probability vector $\mathbf{c}_{q+1}^{(j)} \in \mathbb{R}^{MI \times 1}$, where the k-th element $c_{k,q+1}^{(j)}$ with index $k = (m-1)I + i$ denotes the probability for assigning BS m to target i. Moreover, a binary mask $\mathbf{m}_{q+1}^{(j)} \in \{0, 1\}^{MI \times 1}$ is introduced to mask out the invalid BS-target pairs based on the remaining RF chains of each BS and the existing connections. As a result, the final output of the actor network in the j-th iteration is a stochastic policy given by $\pi^{(j)}(\cdot \mid s_{q+1}) = \mathbf{c}_{q+1}^{(j)} \odot \mathbf{m}_{q+1}^{(j)}$, followed by normalization. The sub-action $a_{q+1}^{(j)}$ is then sampled from this masked distribution, i.e., $a_{q+1}^{(j)} \sim \pi^{(j)}(\cdot \mid s_{q+1})$. After completing J-th iteration, the critic network is utilized to evaluate the current policy by estimating the state value function $V(s_{q+1})$ and update the model parameters accordingly.

After the complete $\mathbf{A}_{q+1}$ is determined, the beam prediction decision is executed by jointly considering $\mathbf{A}_{q+1}$ and the predicted state $\tilde{\mathbf{u}}_{q+1}$. Specifically, for each assigned BS-target pair, the state representations are concatenated and fed into an auxiliary beam prediction head to output the selected beam indices for the transmitter and receiver of each BS, denoted by $\mathcal{B}_{m,q+1}^{(\mathrm{tx})} \in \{1, \ldots, M_T\}$ and $\mathcal{B}_{m,q+1}^{(\mathrm{rx})} \in \{1, \ldots, M_R\}$, respectively. The beam prediction head for BS m can be expressed as

$$\{\mathcal{B}_{m,q+1}^{(\mathrm{tx})}, \mathcal{B}_{m,q+1}^{(\mathrm{rx})}\} = \text{Top-K}(\text{Softmax}\big(\text{MLP}([\mathbf{p}_m; \tilde{\mathbf{u}}_{i,q+1}])\big)). \tag{12}$$

Accordingly, in the $q+1$-th block, the analog beamforming matrix $\{\mathbf{W}_{m,q+1}^{\mathrm{RF},k}\}_{k=1}^{K}$ ($\{\mathbf{F}_{m,q+1}^{\mathrm{RF},k}\}_{k=1}^{K}$) can be constructed by grouping the selected beams in $\mathcal{B}_{m,q+1}^{(\mathrm{tx})}$ ($\mathcal{B}_{m,q+1}^{(\mathrm{rx})}$) into K groups, each consisting of N^{RF} beams.

3.3. Training Policy

The training process of the proposed deep reinforcement learning architecture is executed in two sequential phases: a supervised pre-training phase and a Proximal Policy Optimization (PPO) training phase. In Phase 1, we employ supervised learning to pre-train the extractor and the beam prediction head of the controller. The training is guided by a joint loss function that combines the mean squared error

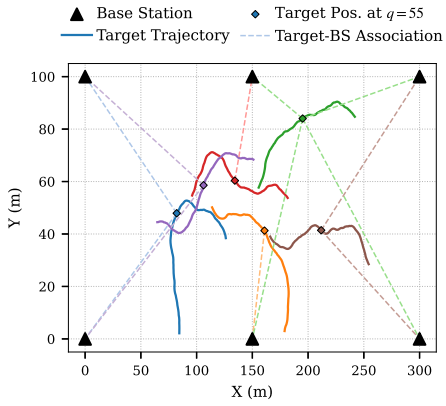

Figure 3. Illustration of target-BS associations.

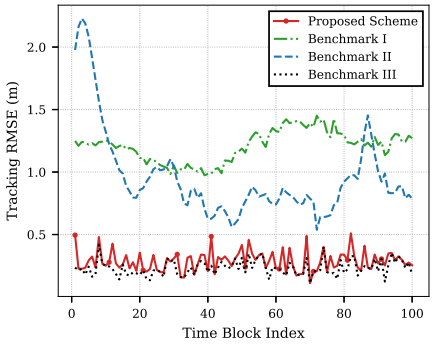

Figure 4. The Tracking RMSE of each time block.

(MSE) loss (Zhu et al., 2021) for the state estimation task and the cross-entropy (CE) loss for the beam prediction task. In Phase 2, we fix the parameters of the extractor, focusing solely on optimizing the actor and critic networks in the controller. At each block q, the actor network samples the action a_q based on the current policy $\pi(\cdot|s_q)$ and observes the reward r_q. A transition tuple $\{s_q, a_q, r_q, s_{q+1}\}$ is then stored in the replay buffer. After collecting a trajectory of Q_Δ blocks, the stored data is utilized to update the networks. Due to space limitations, we refer readers to (Schulman et al., 2017) for the details of the PPO training.

4. Numerical Experiments

We consider an area of $300 \times 100\ m^2$ with $M = 6$ BSs and $I \in \{4, 6, 8\}$ targets. The BSs are located at $\mathbf{p}_1 = [0, 0]^T$, $\mathbf{p}_2 = [150, 0]^T$, $\mathbf{p}_3 = [300, 0]^T$, $\mathbf{p}_4 = [0, 100]^T$, $\mathbf{p}_5 = [150, 100]^T$, and $\mathbf{p}_6 = [300, 100]^T$, respectively. All BSs are equipped with ULAs of $N_T = N_R = 32$ antennas and $N^{RF} = 2$ RF chains. Each analog beam has to be selected from a discrete Fourier transform (DFT) codebook (He et al., 2018) of size 32. The mobility of each target i is modeled as $\mathbf{u}_{i,q+1} = \mathbf{u}_{i,q} + \mathbf{v}_{i,q}\Delta T$, where $\mathbf{v}_{i,q} = [v^x_{i,q}, v^y_{i,q}]^T$ denotes the velocity vector of target i in the q-th block, and $\Delta T = 0.1$ s. We assume that $v^x_{i,q} = v\cos(\beta_{i,q})$ and $v^y_{i,q} = v\sin(\beta_{i,q})$, where the target speed is fixed as $v = 10$ m/s. The moving direction in the horizontal plane $\beta_{i,q}$ follows a dynamic model over time, given by $\beta_{i,q+1} = \beta_{i,q} + \Delta\beta$, where $\Delta\beta \sim \mathcal{U}(-15°, 15°)$.

Table 1. Tracking RMSE (m) over number of targets

METHOD	$I = 4$	$I = 6$	$I = 8$
PROPOSED	**0.2639**	**0.2984**	**0.4994**
BENCHMARK I	0.9747	1.1764	1.5975
BENCHMARK II	1.1138	1.1346	1.1942
BENCHMARK III	0.2352	0.2502	OOM[1]

We provide three benchmark schemes. In **Benchmark I**, we adopt a traditional optimization-based approach that employs the Kalman Filter for state estimation and the standard convex optimization techniques for target-BS association. In **Benchmark II**, we adopt a "learn-to-optimize" approach, where the DGNN is trained using the optimal solutions as ground-truth labels to predict target-BS associations and beam selections. In **Benchmark III**, we adopt the DGNN as the extractor, coupled with the exhaustive search over all possible combinations to determine the optimal association and beam prediction results.

In Fig. 3, we illustrate the simulated cooperative ISAC system with $I = 6$ and $Q = 100$. A snapshot of the target-BS association at block $q = 55$ demonstrates the flexible resource allocation capability of the proposed scheme. In Fig. 4, we evaluate the average tracking root mean square error (RMSE) of all targets during the tracking procedure. It is observed that the proposed scheme significantly outperforms the optimization-based approach (Benchmark I) and the "learn-to-optimize" approach (Benchmark II), while achieving comparable performance to the exhaustive search-based scheme (Benchmark III). This result validates the capability of the proposed scheme to find near-optimal strategy for target-BS associations and beam predictions. In Table 1, we evaluate the tracking error versus the number of targets. It is observed that the tracking performance degrades as the number of targets increases, as the RF-chain constraint reduce the average number of beams available per target. Nevertheless, the proposed scheme achieves a low tracking error compared to the benchmarks.

5. Conclusion

This work investigated the MTT problem in the cooperative mmWave ISAC system. We designed a DRL-based framework that comprises a DGNN-based extractor and an A2C-based controller to jointly perform target state estimation, target-BS association adjustment and beam prediction tasks, in which a novel auto-regressive policy was employed to address the target-BS association constraint. Simulation results demonstrated that the proposed scheme achieves high-accuracy tracking performance over a long time interval.

[1]This scheme suffers from out-of-memory (OOM) errors due to the extremely large combinatorial search space.

References

Bello, I., Pham, H., Le, Q. V., Norouzi, M., and Bengio, S. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*, 2017.

Chan, K.-L., Chang, R. Y., Chien, F.-T., and Poor, H. V. Beamforming and load-balanced user association in RIS-Aided mmWave systems via adaptive attention graph neural networks. *IEEE Trans. Wireless Commun.*, 25: 5781–5796, 2026. Early Access.

Deng, W., Liu, Y., Li, M., and Lei, M. GNN-aided user association and beam selection for mmWave-integrated heterogeneous networks. *IEEE Wireless Commun. Lett.*, 12(11):1836–1840, Nov. 2023.

He, S. et al. Codebook based hybrid precoding for millimeter wave MIMO systems. *IEEE Trans. Signal Process.*, 2018.

Hu, Y., Zhang, S., Murch, R., and Liu, L. SAR/ISAR imaging in 6G network, 2026a. [Online]. Available: https://arxiv.org/abs/2604.00583.

Hu, Y., Zhu, W., Wu, C., Zhang, S., Zhang, J. A., and Liu, L. Networked tracking of multiple moving targets in 6G network, 2026b. [Online]. Available: arXiv preprint arXiv:2604.19709.

Jiang, P., Li, M., Liu, R., and Liu, Q. Graph learning for cooperative cell-free ISAC systems: From optimization to estimation. *IEEE Trans. Wireless Commun.*, 25:13992–14008, 2026.

Li, T., Zhu, W., Zhang, S., Cao, J., Cui, S., and Liu, L. m^{3}TrackFormer: Transformer-based mmWave multi-target tracking with lost target re-acquisition capability, 2026a. [Online]. Available: https://arxiv.org/abs/2602.18254.

Li, T., Zhu, W., Zhang, S., Cao, J., Cui, S., and Liu, L. M^{2}TrackFormer: Transformer-based mmWave tracking with lost target re-acquisition capability. In *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 21096–21100, 2026b.

Lian, L., Bai, C., Xu, Y., Dong, H., Cheng, R., and Zhang, S. Learning to beamform for cooperative localization and communication: A link heterogeneous GNN-based approach. *IEEE Trans. Wireless Commun.*, 25:6177–6190, 2026.

Liu, F., Masouros, C., Petropulu, A. P., Griffiths, H., and Hanzo, L. Joint radar and communication design: Applications, state-of-the-art, and the road ahead. *IEEE Trans. Commun.*, 68(6):3834–3862, Jun. 2020.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., and Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1928–1937, 2016.

Potter, L. C., Ertin, E., Parker, J. T., and Cetin, M. Sparsity and compressed sensing in radar imaging. *Proceedings of the IEEE*, 98(6):1006–1020, Jun. 2010.

Schmidt, R. O. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.*, 34(3): 276–280, Mar. 1986.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Stoica, P. and Sharman, K. C. Maximum likelihood methods for direction-of-arrival estimation. *IEEE Trans. Acoust., Speech, Signal Process.*, 38(7):1132–1143, Jul. 1990.

Wang, Z., Li, M., Liu, R., and Liu, Q. Joint user association and hybrid beamforming designs for Cell-Free mmWave MIMO communications. *IEEE Trans. Commun.*, 70(11): 7307–7321, Nov. 2022.

Zhang, J. A. et al. Enabling joint communication and radar sensing in mobile networks—a survey. *IEEE Commun. Surveys Tuts.*, 24(1):306–345, Nov. 2021.

Zhang, Q., Liang, Y.-C., and Poor, H. V. Intelligent user association for symbiotic radio networks using deep reinforcement learning. *IEEE Trans. Wireless Commun.*, 19 (7):4535–4548, Jul. 2020.

Zheng, L., Lops, M., Eldar, Y. C., and Wang, X. Radar and communication coexistence: An overview: A review of recent methods. *IEEE Signal Process. Mag.*, 36(5):85–99, Sep. 2019.

Zhu, W., Tao, M., Yuan, X., and Guan, Y. Deep-learned approximate message passing for asynchronous massive connectivity. *IEEE Trans. Wireless Commun.*, 20(8):5434–5448, Aug. 2021.

Zhu, W., Gao, J., Zhang, S., Tao, M., and Liu, L. Joint multi-user tracking and signal detection in reconfigurable intelligent surface-assisted cell-free ISAC systems, 2026. [Online]. Available: https://arxiv.org/abs/2602.18018.