基于双层深度 PPO 的自适应锚点选择 UWB 定位方法

周 元，吴仕勋†，蓝章礼，徐 凯，张 淼，靳 双

(重庆交通大学 信息科学与工程学院，重庆 400074)

摘 要: 复杂室内环境下的非视距传播、多径以及动态干扰等问题严重制约 UWB 定位精度的提升, 现有的锚点选择策略难以在信号质量与几何分布之间实现自适应平衡. 为此, 提出一种基于双层深度近端策略优化 (DDPPO) 的自适应锚点选择方法, 将锚点选择建模为序贯决策问题, 并利用深度强化学习实现智能筛选. 首先, 构建融合信道冲激响应、几何分布、轨迹时序与信号质量的多源状态空间, 实现对动态环境的全面感知. 其次, 设计层次化双层 PPO 架构将锚点选择解耦为数量决策与组合决策两个层次, 结合课程学习策略引导模型快速收敛. 最后, 策略网络输出的锚点子集经加权最小二乘解算位置, 以定位误差为主导构建奖励函数, 该函数所包含的样本难度自适应机制可根据实时误差动态调整对锚点数量的偏好, 生成的奖励信号反馈至决策网络学习形成闭环. 在四类真实室内场景数据集上的实验表明, DDPPO 方法平均定位误差为 0.219 米, 较现有六种方法降幅达 45.9% 至 72.5%, 在定位精度与计算效率间取得良好平衡.

关键词: UWB 定位；锚点选择；深度强化学习；近端策略优化；自适应决策；几何精度因子

中图分类号: TN92；TP18 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2026.0066

引用格式: 周元, 吴仕勋, 蓝章礼, 等. 基于双层深度 PPO 的自适应锚点选择 UWB 定位方法 [J]. 控制与决策.

An adaptive anchor selection UWB localization method based on dual-layer deep PPO

ZHOU Yuan，*WU Shi-xun*†，*LAN Zhang-li*，*XU Kai*，*ZHANG Miao*，*JIN Shuang*

(School of Information Science and Engineering，Chongqing Jiaotong University，Chongqing 400074，China)

Abstract: In complex indoor environments, UWB positioning accuracy is severely limited by non-line-of-sight (NLoS), multipath, and dynamic interference. Moreover, conventional anchor selection strategies fail to adaptively balance signal quality and favorable geometric distribution. To address these challenges, this paper proposes an adaptive anchor selection method based on a double-layer deep proximal policy optimization (DDPPO) framework, modeling anchor selection as a sequential decision problem with deep reinforcement learning. First, a multi-source state space integrating channel impulse response (CIR) signals, geometric layout, temporal trajectory information, and signal quality metrics is constructed to comprehensively perceive the dynamic environment. Second, a hierarchical dual-layer PPO architecture is designed to decouple anchor selection into quantity and combination decisions, and combines a curriculum learning strategy for rapid convergence. Finally, the policy network's anchor subset is directly used for position estimation via weighted least squares (WLS) model. The positioning error is taken as the dominant term to construct the reward function, which incorporates a sample-difficulty adaptive mechanism to dynamically adjust anchor count preference based on real-time error. The resulting reward signal is fed back to the decision network, forming a closed-loop learning system. Experimental on four real-world indoor datasets show that the DDPPO method achieves an average error of 0.219 m, representing an improvement of 45.9% to 72.5% over six existing methods, while balancing accuracy and efficiency.

Keywords: UWB positioning； anchor selection； deep reinforcement learning； proximal policy optimization； adaptive decision-making；geometric dilution of precision

收稿日期: 2026-01-20；**录用日期:** 2026-03-13.

基金项目: 重庆市自然科学基金项目 (CSTB2024NSCQ-MSX0275)；研究生教育“课程思政”示范项目 (KCSZ2025009).

责任编委: 魏秀琨.

†通信作者. E-mail: wushixun333@163.com.

0 引 言

随着物联网、智能制造和基于位置服务(Location-Based Services, LBS)需求的快速增长, 室内定位已成为智慧城市、智慧医疗、智能工厂等领域的核心技术支撑[1]. 超宽带 (Ultra-Wideband, UWB)技术凭借大带宽、高时间分辨率和抗多径干扰能力, 在视距 (Line-of-sight, LoS) 条件下可实现厘米级定位精度, 已经成为室内定位的有力解决方案[2]. 然而, 实际室内环境的复杂性严重制约了 UWB 技术潜力的发挥[3]. 现有的文献大多聚焦于非视距 (Non-Line-of-sight, NLoS) 传播引发的测距正偏差、多径效应导致的信号传播路径复杂化, 以及动态障碍物与人员遮挡对定位可靠性的削弱等误差问题[4-6]. 但定位系统的整体性能优化, 除了需攻克上述测距层面的关键挑战外, 还在很大程度上依赖于锚点的几何布局设计与实时选择策略[7]. Wang 等人[8] 基于几何精度衰减因子 (Geometric Dilution of Precision, GDOP) 最小化准则来优化锚点配置, 扩展了低 GDOP 锥形配置的应用, 但该方法仅针对静态锚点布局优化, 未结合实时信号质量. Yi 等人[9] 针对高 GDOP 条件下传统三边定位精度下降的问题, 提出了约束最小二乘三边定位 (Constrained Least Squares Trilateration, CLST) 算法, 在保持计算效率的同时在严重 GDOP 条件下也能实现高定位性能, 仿真和实验验证了其有效性. 以上方法在静态条件下可以提高定位精度, 却默认锚点位置与数量一成不变, 且对实时信道冲激响应 (Channel Impulse Response, CIR) 信息质量未做过多的关注. 与之互补的, Fontaine 等人[10] 利用 CIR 指纹的锚点子集选择方法, 通过神经网络从单个 UWB 数据包的 CIR 中提取特征实现高质量锚点的筛选. 该方法虽然可以挑出高质量锚点, 但完全忽略几何分布, 导致在走廊、角落等极端几何下精度大幅度下降.

近年来, 深度学习 (Deep Learning, DL) 与强化学习 (Reinforcement Learning, RL) 为解决复杂环境下的定位问题提供了新思路[11]. 在 DL 领域, Li 等人[12] 探索了一种基于变分贝叶斯的半监督方法, 利用少量真值与大量无标签数据有效缓解 NLoS 误差. Kim 等人[13] 首次将深度 Q 网络 (Deep Q-Network, DQN) 用于 UWB 锚点和定位策略选择, 该方法可以有效识别 LoS 链路, 初步验证了深度强化学习 (Deep Reinforcement Learning, DRL) 在锚点选择决策中的有效性, 但仅针对单锚点选择, 并未解决多锚点组合选择的核心问题. Coppens 等人[14] 通过自监督 DRL 修正测距误差, 将 CIR 作为输入状态, 通过智能体来预测测距修正值以最小化校正距离与估计距离间的误差. Wang 等人[15] 提出了一种基于 RL 补偿滤波的多智能体协同定位算法来探索多机器人协同定位, 将其用于扩展卡尔曼滤波 (Extended Kalman Filter, EKF) 残差补偿, 该算法结合了先验位置信息、EKF 及多智能体间的信息共享, 显著提升了定位精度. 但该方法未涉及锚点选择, 仅聚焦误差补偿.

综上所述, 现有锚点选择方法无法根据实时变化的信道状态与空间几何关系进行动态调整, 割裂了几何分布与信号质量的协同优化, 往往导致在复杂场景下选择了受干扰严重或几何分布不佳的锚点子集. 此外, RL 方法多采用单层架构, 面临动作空间爆炸难题, 且未充分挖掘多源信号特征, 难以平衡定位精度与实时性. 针对上述问题, 本文提出了一种基于双层深度近端策略优化 (Dual-layer Deep Proximal Policy Optimization, DDPPO) 的自适应锚点选择方法, 用于 UWB 室内定位系统, 将锚点选择问题建模为一个序贯决策过程, 采用 DRL 实现锚点自适应选择策略. 本文的主要贡献如下:

1) 提出融合 CIR 信号、几何分布、轨迹时序信息和信号质量的多维度特征, 显著提升了特征表达的丰富性和决策的准确性, 有效解决传统方法特征利用不充分的问题.

2) 提出层次化双层 PPO 架构, 将复杂的锚点组合选择问题解耦为数量决策与组合决策两个层次, 大幅降低动作空间复杂度; 并结合课程学习策略, 在训练初期设计基于鲁棒统计的规则引导与衰减机制来确定K值, 实现训练过程快速稳定收敛.

3) 设计奖励函数包括 GDOP、信号质量以及样本难度自适应奖励机制, 根据定位误差动态调整锚点数量偏好, 并将 PPO 决策与加权最小二乘 (Weighted Least Squares, WLS) 物理模型融合实现了高精度的定位.

4) 在公开数据集的四个真实室内场景中进行验证, 本方法实现了 0.219 米的平均定位精度, 相比所对比的六种方法平均提升 64.4%. 在环境 0、1、2、3 中, 相较于各场景下的最优对比方法, 定位误差分别降低了 39.6%、54.5%、46.1% 和 37.2%, 展现出良好的泛化能力. 消融实验进一步验证了各模块的有效性, 且模型在不同环境中表现稳定, 表明了该方法具备较强的鲁棒性与实用价值.

1 基于 DDPPO 的自适应锚点选择方法

传统锚点选择方法多依赖固定策略, 难以适应

动态变化的环境干扰. 为此, 本文将 UWB 定位过程中每个定位时刻的动态锚点选择问题建模为一个马尔可夫决策过程 (Markov decision process, MDP), 通过 DRL 方法实现自适应的锚点选择, 以最小化实时定位误差. MDP 由五元组$< S, A, P, R, \gamma >$来定义它的核心要素. 将当前环境中所有的锚点集合定义为$\mathcal{A} = A_1, \ldots, A_N$, 其中$N$表示当前环境中所部署的锚点个数. 目标是在每个定位时刻t从$\mathcal{A}$中动态地选出大小为$K(3 \leqslant K \leqslant N)$的锚点子集, 使得在当前时间步的位置估计误差最小. 状态空间S包含当前定位场景的关键特征, 动作空间A定义所有可能的锚点选择动作, 状态转移概率P描述执行动作后转移到下一状态的概率分布, 奖励函数R评估所选锚点对应的定位误差反馈, 折扣因子γ用于权衡即时与长期奖励[16].

本文提出的 DDPPO 方法整体流程如图 1 所示, 主要包括两个模块. 多源异构状态空间构建模块将基于当前时刻的 CIR 信号、信号质量、几何分布、轨迹时序信息、信号质量等构建综合状态表示. 层次化双层 PPO 决策模块通过两级决策实现锚点选择, 第一层K值选择网络根据状态与K值特征动态确定子集规模K, 再经第二层组合选择网络调用对应规格的决策网络输出锚点子集, 并结合 WLS 定位模型求解坐标. 为提升训练效率引入课程学习策略, 将规则引导与 RL 渐进融合以加速收敛并改善策略质量.

2　多源异构状态空间构建

状态空间S需综合反映环境特征、历史轨迹及锚点属性等来支撑锚点选择决策. 本文基于当前时刻所有锚点数据中提取 4 类特征, 构建了如图 2 所示的多源状态向量$s_t = [\mathbf{f}_{\text{CIR}}, \mathbf{f}_{\text{singal}}, \mathbf{f}_{\text{geo}}, \mathbf{f}_{\text{temporal}}]$.

2.1　CIR 特征提取

CIR 能完整反映 UWB 信号多径传播特性, 是区分 LoS/NLoS 的关键[17]. 对每个锚点$i \in 0, 1, \ldots, N-1$, 从其 CIR 信号中以首达路径 (First Arrival Path, FP) 峰值位置t_{FP}为基准, 截取前后长度为L_{FPW}的观测窗口覆盖首径信号的主能量区, 提取其幅值mag_i、功率pow_i、归一化相位信息Φ_i. 令$h_{i,l}$为第i个锚点 CIR 信号的第l个采样点, 计算公式如下:

$$\text{mag}_i = \frac{1}{L_{\text{FPW}}} \sum_{l=1}^{L_{\text{FPW}}} |h_{i,l}|, \tag{1}$$

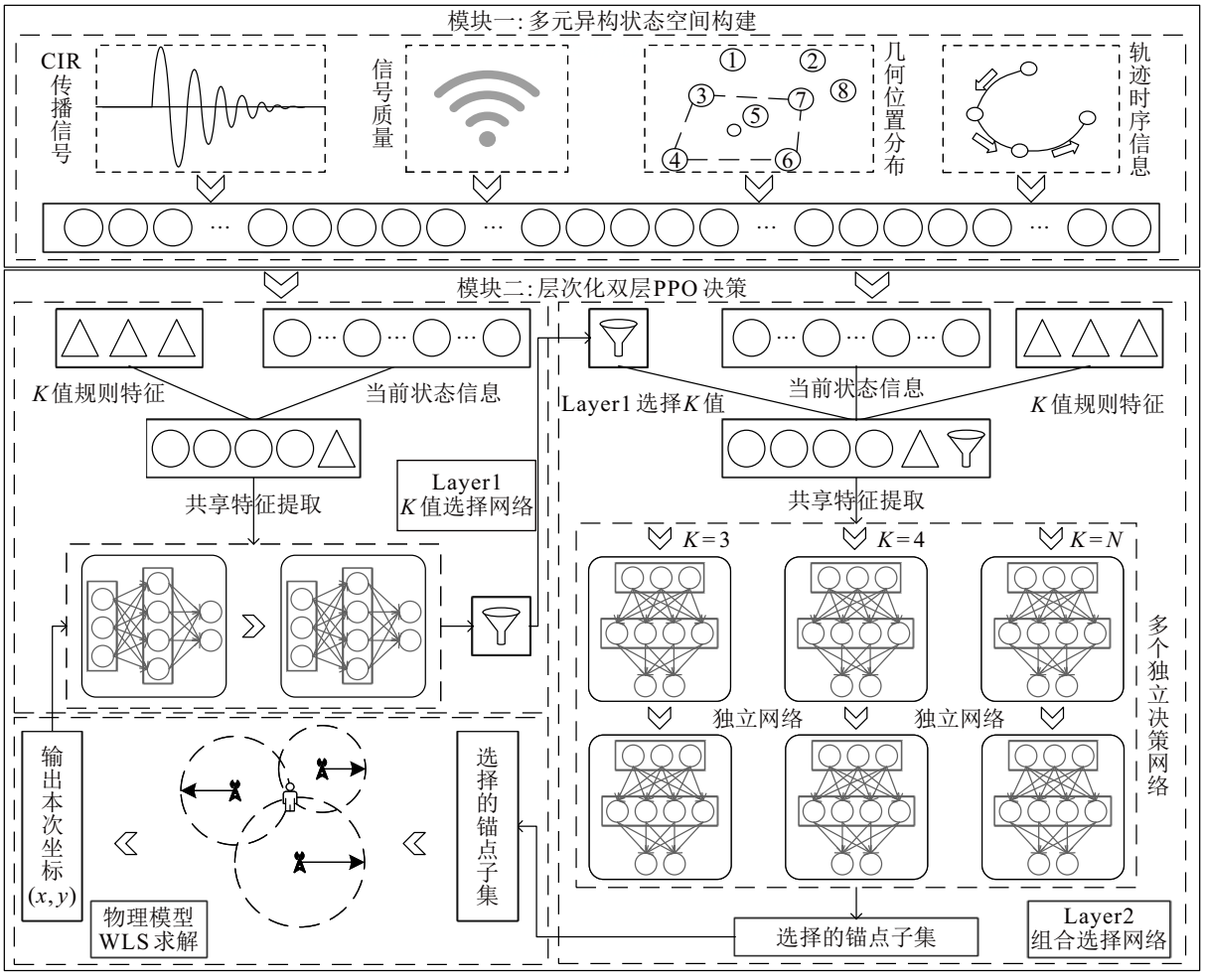

图1　DDPPO 方法整体流程

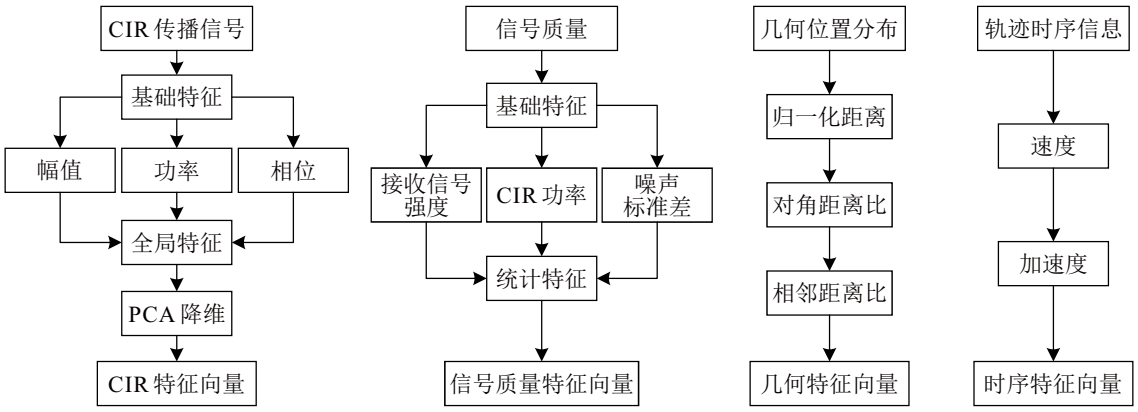

图2 多源异构状态空间特征构成

$$\mathrm{pow}_i = \frac{1}{L_{\mathrm{FPW}}}\sum_{l=1}^{L_{\mathrm{FPW}}}|h_{i,l}|^2, \tag{2}$$

$$\Phi_i = \frac{1}{\pi}\arg\frac{1}{L_{\mathrm{FPW}}}\sum_{l=1}^{L_{\mathrm{FPW}}} e^{\mathrm{j}\cdot\arg(h_{i,l})} \in [-1,1]. \tag{3}$$

其中, j为虚数单位, $\arg(\cdot)$表示复数的幅角函数, 返回值范围$[-\pi,\pi]$.

为进一步挖掘全局信号规律计算四类全局统计特征. 幅值均值μ_{mag}反映整体信号强度, 避免选择单一强信号锚点而忽略全局信号薄弱的情况; 幅值方差σ_{mag}反映锚点信号的稳定性; 全局功率均值μ_{pow}与功率比值R_{pow}反映信号分布均匀性.

综上所述, 原始构建的 CIR 特征向量维度为$3N+4$, 其高维特性不仅会引入冗余信息, 还可能影响后续特征的有效提取. 为此, 本文采用主成分分析 (Principal Component Analysis, PCA) 对特征向量进行降维, 得到压缩后的特征向量$\mathbf{f}_{\mathrm{CIR}}$, 保留核心信息并提升训练效率.

2.2 信号质量特征提取

本文基于所有锚点的原始 FP 等信息, 从整体信号水平、锚点信号一致性及环境噪声水平三个维度提取信号质量特征. 使用平均接收信号强度$\mathrm{RSS}_{\mathrm{mean}}$和平均 CIR 功率$\mathrm{CIRP}_{\mathrm{mean}}$反映整体信号水平, 通过 RSS 标准差$\mathrm{RSS}_{\mathrm{std}}$和 CIR 功率标准差$\mathrm{CIRP}_{\mathrm{std}}$体现锚点间信号状态的一致性, 并借助平均噪声标准差$\mathrm{Noise}_{\mathrm{mean}}$刻画环境噪声水平. 上述特征共同构成信号质量特征向量$\mathbf{f}_{\mathrm{singal}}$.

2.3 几何特征提取

几何特征可以刻画用户位置与锚点间的空间关系, 本文采用基于测距值的锚点对比方案, 通过比较不同锚点间的距离关系来推断用户的空间位置特征. 令在时刻t所有测距值集合为$\mathbf{r}=[r_0,r_1,\ldots,r_{N-1}]$, 其中$r_i$表示用户到第$i$个锚点之间的实测距离, 构建归一化距离、对角距离比和相邻距离比这三类特征.

归一化距离特征可以消除绝对距离尺度差异对特征有效性的影响, 反映了各锚点之间的相对远近信息, 则第i个锚点的归一化距离d_i^{norm}定义为:

$$d_i^{\mathrm{norm}} = \frac{r_i}{\max(r_0,\ldots,r_{N-1})}. \tag{4}$$

对角距离比特征可以刻画用户相对于对角锚点的空间位置偏向性. 假设锚点按空间位置配对为$\lfloor N/2\rfloor$对对角锚点, 第j对对角锚点的距离比ρ_j^{diag}定义为:

$$\rho_j^{\mathrm{diag}} = \frac{r_j}{r_{j+\lfloor N/2\rfloor}},\ j=0,1,\ldots,\lfloor N/2\rfloor-1. \tag{5}$$

若$\rho_j^{\mathrm{diag}}>1$则说明用户更靠近第j个锚点所在象限, 反之则更靠近其对角的锚点.

相邻距离比特征能够挖掘用户在局部区域的相对位置信息, 消除锚点部署顺序对局部位置感知的干扰, 第k对相邻锚点的距离比ρ_k^{adj}定义为:

$$\rho_k^{\mathrm{adj}} = \frac{r_{2k}}{r_{2k+1}},\ k=0,1,\ldots,\lfloor N/2\rfloor-1. \tag{6}$$

综合得到几何特征向量$\mathbf{f}_{\mathrm{geo}}$, 从相对距离尺度、对角空间分布及局部线性关系三个层面完整描述了用户与锚点的空间几何关系, 引导模型选择空间覆盖均匀的锚点子集.

2.4 时序特征提取

时序特征可以捕捉用户的运动状态[18], 通过分析位置时间序列的连续性约束锚点选择决策, 提高动态场景定位稳定性. 设用户在时刻t的位置为(x_t,y_t), 本文基于用户位置的归一化历史序列$\mathbf{P}_t=(x_t,y_t)_{t=0}^{T-1}$提取速度与加速度特征. 速度特征包含方向分量与速率大小两个维度. 其中, 速度方向分量$\mathbf{v}_t=[v_{x,t},v_{y,t}]^{\mathrm{T}}$可拆解为$x$方向和$y$方向, 分别反映用户在对应方向上的瞬时运动快慢与方向, 速率大小$v_{\mathrm{norm},t}$即合速度是速度方向分量构成的向量模值. 为保证特征的完整性, 定义初始时刻的用户速度为零$\mathbf{v}_0=\mathbf{0}$. 加速度特征同样包含方向分量与变化率

大小两个维度. 加速度方向分量$\mathbf{a}_t=[a_{x,t},a_{y,t}]^{\mathrm{T}}$对应$x$方向和$y$方向的速度变化, 反映用户在各方向上运动速度的变化情况, 加速度变化率大小$a_{\mathrm{norm},t}$即合加速度是加速度方向分量构成的向量模值. 考虑到加速度计算需基于前后两个时刻的速度, 定义前两个时刻 (t<2) 用户加速度为零$\mathbf{a}_t=\mathbf{0}$.

综上可以得到时序特征向量$\mathbf{f}_{\mathrm{temporal}}$, 从瞬时运动快慢和运动状态变化率两个维度完整刻画了用户的动态运动特性, 有助于突破单时刻静态信息的局限性, 帮助策略适应运动趋势.

3　层次化双层 PPO 决策

3.1　动作空间设计

动作空间定义为锚点选择的所有可能组合. 若基于锚点数量N直接枚举所有组合会导致动作空间指数级增长. 为缓解这一问题, 本文采用层次化双层动作空间设计, 将锚点选择分解为锚点数量K值选择和具体组合选择这两个阶段, 先由第一层网络确定锚点子集大小$K(3\leqslant K\leqslant N)$, 再由第二层网络从$N$个锚点中选出具体的$K$元子集, 对应$(N,K)^{\mathrm{T}}$种可能. 该设计将最大动作规模限制为某一$K$值对应的组合数, 有效避免组合爆炸问题, 显著提升训练效率与收敛稳定性.

3.2　课程学习策略引导

为提升模型泛化能力与收敛效率, 针对K值选择层引入课程学习 (Curriculum Learning, CL) 策略, 通过规则方法引导初期训练并逐步过渡至自主决策. 规则引导比例随训练衰减至零, 以缓解初期探索困难的问题. 令d_i表示用户与锚点i之间的真实距离, 规则方法首先使用所有有效锚点进行 WLS 解算获得初始位置估计$\hat{\boldsymbol{P}}$和测距残差$e_i=d_i-|\hat{\boldsymbol{p}}-\boldsymbol{p}_i|_2$. 采用基于中位数绝对偏差 (Median Absolute Deviation, MAD) 的鲁棒尺度估计s对残差标准化以更好地消除残差量纲影响同时抑制 NLoS 异常值:

$$s=1.4826\cdot\underset{i}{\mathrm{median}}\{|e_i-\tilde{e}|\},\ z_i=\frac{|e_i|}{s}. \tag{7}$$

其中, 系数 1.4826 使得s在正态分布下近似等于标准差, $\tilde{e}$为残差向量$\mathbf{e}$的中位数[19]. z_i表示第i个锚点的标准化测距残差. 锚点置信度定义为$\mathrm{score}_i=1/(1+z_i^2)$, 标准化残差越小, 锚点就越可靠, 置信度越高. 然后根据置信度对降序排序后的锚点逐次选择前K个锚点组合, 计算以下指标: GDOP G_W[9]、修剪残差分位数z_p、目标函数J_K. 为避免单个异常锚点影响评估, 对选中的K个锚点的标准化残差进行修剪, 当$K\geqslant 4$时, 剔除最大值后计算其 75% 分位数z_p来反映测距质量. K值选择的目标函数J_K定义为:

$$J_K=G_W+\lambda_K\cdot z_p+\alpha_K\cdot K. \tag{8}$$

其中, λ_K为残差权重系数, α_K为锚点数量惩罚系数, 避免过度使用锚点. 规则方法通过最小化J_K确定综合最优的K_{best}, 在区间$[K_{\min},K_{\mathrm{best}}]$内寻找满足硬阈值条件的最小可行解$K_{\mathrm{feasible}}$, 此处$K_{\min}=3$:

$$K_{\mathrm{feasible}}=\min\{K:\,G_W\leqslant\tau_g\cap z_p\leqslant\tau_z\}. \tag{9}$$

其中τ_g为几何质量阈值, τ_z为残差质量阈值, 符号$\cap$要求几何质量与标准化残差必须同时低于各自阈值. 区间上限设为K_{best}是为了确保所寻解在综合代价上不差于理论最优解最终规则输出的K值:

$$K_{\mathrm{rule}}=\begin{cases}K_{\mathrm{feasible}}\ \text{若可行解存在}\\ K_{\mathrm{best}}\ \text{否则}\end{cases} \tag{10}$$

K_{rule}为基于先验规则计算的推荐K值. 该策略优先选择满足质量阈值的最小K值, 若无可行解则选择目标函数最优值, 在几何质量、测距可靠性和计算成本间取得平衡, 为 PPO 训练提供高质量的初始引导. 上述规则方法依赖 WLS 初始解计算残差, 在强 NLoS 场景下初始解偏差可能影响K_{rule}的可靠性. 为此, 本文在课程学习中引入引导率衰减机制, 规则使用概率从初始值线性衰减至 0. 该设计使策略在训练早期借助先验知识快速收敛, 同时逐步过渡至完全自主决策. 当规则推荐值出现偏差时探索机制仍允许策略以概率选择其他K值, 避免被错误规则持续误导.

3.3　奖励函数设计

为评估锚点选择动作a_t的优劣并引导智能体学习, 本文设计了一个双层统一的多目标奖励函数$R(s_t,a_t)$, 同时优化K值选择与锚点组合选择两层策略. 该函数为两层网络提供同一奖励信号, 确保决策目标的一致性. 设基于动作a_t选择的锚点子集$\mathcal{S}_t$得到的位置估计为$\hat{p}_t=(\hat{x}_t,\hat{y}_t)$. 奖励函数定义为:

$$R(s_t,a_t)=-w_1\varepsilon_t+w_2r_{G_t}+w_3r_{Q_t}+r_{\mathrm{adapt}}. \tag{11}$$

其中, $\varepsilon_t=||\hat{p}_t-p_t||_2$为奖励函数主导项, 表示$t$时刻的定位误差直接反映定位精度, 取负值以将最小化误差转化为最大化奖励. w_1, w_2, w_3为权重系数.

$$\begin{cases}r_{G_t}=-\max(0,G_t-G_{\mathrm{th}})\\ r_{Q_t}=-\max(0,Q_{\mathrm{th}}-Q_t)\end{cases} \tag{12}$$

r_{G_t}作为几何约束项, 其中G_t是t时刻所选锚点组合的 GDOP, 值越小表示锚点几何构型对定位越有利. 当它超过阈值G_{th}时施加惩罚, 引导智能体选择几何分布更优的锚点组合. r_{Q_t}作为信号质量辅助

项, 以$Q_t=\dfrac{1}{K}\sum\limits_{A_i\in\mathcal{S}_t}\mathrm{pow}_i$为度量来代表$t$时刻所选锚点的平均信号质量, 低于阈值$Q_{\mathrm{th}}$时施加惩罚, 鼓励选择信号质量好的锚点. r_{adapt}为自适应奖励项, 由K值选择引导r_K与样本难度自适应引导$r_{\mathrm{difficulty}}$组成:

$$r_{\mathrm{adapt}}=r_K+r_{\mathrm{difficulty}}. \tag{13}$$

令K_t为智能体在时刻t选择的锚点数量, 定义偏差$K_{\mathrm{diff}}=|K_t-K_{\mathrm{rule}}|$, K_t根据K_{diff}设置分段奖励, 完全匹配时给予最大奖励, 轻微偏差时给予适度奖励, 严重偏离时施加惩罚. 该设计作为一种软约束引导 PPO 的学习方向, 在定位精度相近时引导策略倾向于先验知识推荐的K值, 同时保留自主探索更优策略的空间.

$$r_{\mathrm{difficulty}}=\begin{cases}r_{\mathrm{easy}}(\varepsilon_t,K_t) & \varepsilon_t\leqslant 0.3\\ r_{\mathrm{med}}(\varepsilon_t,K_t) & 0.3<\varepsilon_t<0.7\\ r_{\mathrm{hard}}(\varepsilon_t,K_t) & \varepsilon_t\geqslant 0.7\end{cases} \tag{14}$$

其中, $r_{\mathrm{difficulty}}$为锚点根据定位误差自动判断样本难度的自适应奖励值. $r_{\mathrm{easy}}(\varepsilon_t,K_t)$为简单区域保留精度的同时节省计算资源, $r_{\mathrm{med}}(\varepsilon_t,K_t)$为中等难度区域综合考虑, $r_{\mathrm{hard}}(\varepsilon_t,K_t)$为困难区域提供充足冗余信息, 引导不同锚点数以实现精度与效率的智能平衡.

3.4 基于双层 PPO 的决策算法框架

本节构建完整的双层 PPO 决策算法框架, 首先策略网络根据状态生成决策, WLS 执行定位并产生结果, 该结果经奖励函数计算后, 驱动 PPO 算法更新策略网络, 形成自主学习闭环, 最终得到最佳策略网络π_K^*, π_C^*. 具体流程图如图 3 所示.

3.4.1 层次化策略网络架构

双层 PPO 决策网络中的两层网络均采用 3 层全连接结构, 共享统一奖励函数实现协同优化, 输出锚点选择决策a_t. 在第一层K值选择网络中, 输入当前状态s_t和归一化规则推荐K值K_{rule}. 基于共享特征提取层, Actor 分支输出选择不同K值的概率分布$\pi_K(K|s)$, Critic 分支输出当前状态的价值估计$V_K(s_t)$, 为策略更新提供优势函数计算依据. 训练初期采用课程学习策略, 以指定概率使用规则方法K_{rule}作为决策, 以相对概率使用 PPO 策略确定K值, 实现从先验知识到自主决策的平滑过渡.

在确定好K值后, 在第二层组合选择网络中, 基础的网络结构与K值选择网络一致, 输入融合K值信息的状态特征, Actor 分支输出该K值下对应得所有锚点组合的概率分布, 独立选择锚点组合C, 输出组合概率$\pi_C(C|s,K)$, Critic 分支输出当前融合状态的价值估计$V_C(s_t)$, 辅助策略梯度更新. 针对不同K值配置独立网络, 从训练开始即完全由 PPO 自主决策.

3.4.2 加权最小二乘定位

在完成双层策略决策后, 对于选定的锚点子集$\mathcal{S}_t$, 采用 WLS 方法计算位置估计$\hat{p}_t$, WLS 的目标函数为:

$$\min_{\hat{x},\hat{y}}\sum_{A_i\in\mathcal{S}_t}w_i\left(\sqrt{(\hat{x}-x_i)^2+(\hat{y}-y_i)^2}-d_i\right)^2. \tag{15}$$

其中, $w_i=\dfrac{\mathrm{pow}_i}{\sum\limits_{j\in\mathcal{S}_t}\mathrm{pow}_j}$为动态权重, 使功率强的锚点

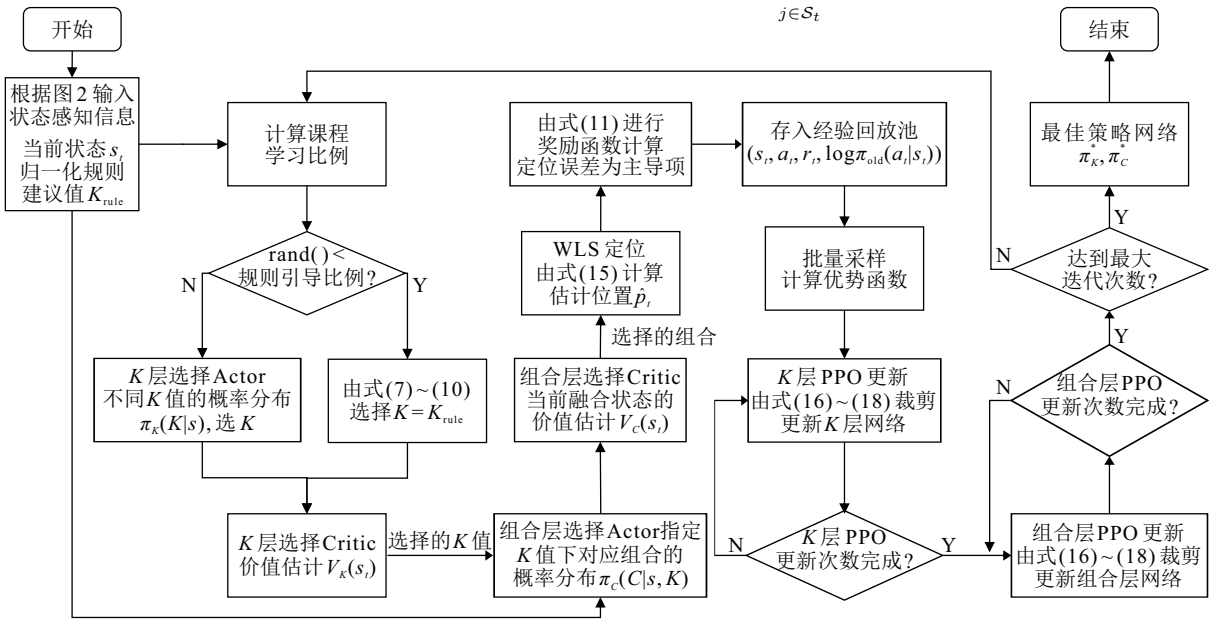

图3 双层 PPO 决策网络流程

获得更大权重, 有效抑制 NLoS 误差. 通过迭代优化求解位置估计$\hat{p}_t$, 该结果将为奖励函数提供主导信号, 形成闭环反馈.

3.4.3　基于 PPO 的策略优化算法

PPO 算法通过其独特的裁剪机制限制策略更新的幅度, 使每次优化均处于相对安全的范围内, 能有效维持训练稳定性, 避免过大梯度导致的策略崩溃问题, 同时支持经验数据复用, 显著提升样本效率[20]. 通过学习的反向更新过程, 依据奖励函数$R(s_t, a_t)$来改进策略网络. 两层网络均采用相同的优化机制, 分别独立更新. 定义t时刻的优势函数$\hat{A}_t = R(s_t, a_t) - V(s_t)$, 引导策略向获得更高奖励的方向更新策略. $V(s_t)$为当前 Critic 网络输出的状态价值估计. 由于K值选择层与组合选择层分别对应$V_K(s_t)$和$V_C(s_t)$, 优化机制相同. PPO 通过最大化如下裁剪目标函数来稳定更新策略参数θ:

$$L^{\mathrm{CLIP}}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, \mathrm{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)]. \quad (16)$$

其中, $r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)}$为重要性采样比率, 表示新旧策略的概率比值, ϵ为裁剪系数.

Critic 网络通过最小化价值损失来提升其预测未来累积奖励的准确性, 价值函数误差定义为:

$$L_{\mathrm{VF}} = \mathbb{E}_t[(V(s_t) - R(s_t, a_t))^2]. \quad (17)$$

由此可以得到本文 PPO 的总损失函数定义为:

$$L = L_{\mathrm{CLIP}} + c_1 \cdot L_{\mathrm{VF}} - c_2 \cdot S_p. \quad (18)$$

其中L_{CLIP}是公式 (16) 中的裁剪目标函数. S_p为策略熵, c_1和c_2分别为价值函数误差与熵正则化项的权重系数.

在训练过程中, 为高效优化上述损失函数并提升策略的泛化能力, 本文采用经验回放机制, 构建经验池存储转移样本$(s_t, a_t, r_t, \log \pi_{\mathrm{old}}(a_t|s_t))$, 每次训练批量随机采样, 进行梯度更新以降低样本间相关性, 提升学习稳定性.

4　实验数据处理

4.1　实验数据集

本文实验采用 Klemen Bregar 等人发布的公开 UWB 定位数据集[21] 该数据集在 4 个真实室内环境中采集, 使用 8 个锚点和 1 个移动标签同步记录测量信息. 四类环境均实现 LoS 与 NLoS 样本均衡分布, 符合真实场景特性. 环境 0 与环境 1 为居住空间, 验证模型对墙体、家具引发干扰的适应能力; 环境 2 为工业场景, 评估金属结构与复杂干扰下的性能; 环境 3 为办公空间, 测试对温和遮挡与弱多径效应的处理效果. 实验中, 四类环境样本严格分离, 分别构建独立的训练与测试集, 确保性能评估互不干扰.

4.2　评价指标

为全面评估本文所提 DDPPO 模型的定位性能, 同时验证锚点选择策略的有效性, 本文的模型评价指标选取平均绝对误差 (Mean Absolute Error, MAE)、均方根误差 (Root Mean Square Error, RMSE) 和定位误差中位数 (Median Position Error, MdPE). 令M为测试集样本总数量, $(x_{i_{\mathrm{est}}}, y_{i_{\mathrm{est}}})$为第$i$个样本的估计位置, $(x_{i_{\mathrm{true}}}, y_{i_{\mathrm{true}}})$为第$i$个样本的真实位置. MAE 反映整体定位精度, 值越小精度越高:

$$\mathrm{MAE} = \frac{1}{M}\sum_{i=1}^{M}\sqrt{(x_{i_{\mathrm{est}}} - x_{i_{\mathrm{true}}})^2 + (y_{i_{\mathrm{est}}} - y_{i_{\mathrm{true}}})^2}. \quad (19)$$

RMSE 可以放大较大误差影响, 衡量高精度场景适配性, 数值越小, 模型的高精度定位能力越强:

$$\mathrm{RMSE} = \sqrt{\frac{1}{M}\sum_{i=1}^{M}[(x_{i_{\mathrm{est}}} - x_{i_{\mathrm{true}}})^2 + (y_{i_{\mathrm{est}}} - y_{i_{\mathrm{true}}})^2]}. \quad (20)$$

MdPE 可以有效规避极端误差干扰, 将所有测试样本的定位误差e_i按升序排列后取中位数:

$$\mathrm{MdPE} = \begin{cases} e_{(\frac{M+1}{2})} & M\text{为奇数} \\ \dfrac{e_{(\frac{M}{2})} + e_{(\frac{M}{2}+1)}}{2} & M\text{为偶数} \end{cases} \quad (21)$$

4.3　实验参数配置

本实验采用 Min-Max 归一化将各类特征统一映射至 [0,1] 区间, 以消除量纲差异并提升模型收敛稳定性. 按照 7:3 的比例划分为训练集与测试集, 采用分层抽样策略各区域样本分布上保持一致. 关键参数配置如表 1 所示.

5　实验结果与分析

5.1　多环境性能评估

5.1.1　特征提取相关性分析

特征提取的质量直接影响定位精度和系统鲁棒性[22]. 为评估不同特征组之间的独立性, 本研究对所有特征进行标准化后计算 Pearson 相关系数[23]. 如图 4 所示, 四类特征组间的相关系数绝对值均低于 0.52. 其中, 信号质量特征与几何特征之间存在中等程度的负相关性, 而其余特征组之间仅呈现弱相关. 这表明四类特征分别从不同维度捕捉定位信息, 彼此之间具有良好的互补性.

CIR 特征包含丰富的多径传播信息, 但其高维特性会导致维度灾难的问题. 为此, 本文采用 PCA

表1 DDPPO 方法关键参数配置

参数类型	参数名称	参数值
网络架构	Actor网络	256→128→动作数
	Critic网络	256→128→1
	共享层	256→256
	状态维度-Layer1	70
	状态维度-Layer2	69
	动作空间-Layer1	6
	动作空间-Layer2	$C(8,K)$
	Dropout	0.1
训练超参数	学习率	1×10^{-4}
	折扣因子	0.99
	裁剪系数	0.2
	价值损失系数	0.5
	熵系数	0.05
	批次大小	64
	经验池容量	10000
	课程学习衰减	0.7→0
	训练总轮数	300
	训练测试比例	7:3
奖励函数超参数	定位误差ε_t权重	1.0
	几何约束r_{G_t}权重	0.01
	信号质量r_{Q_t}权重	0.05
	GDOP阈值	1.5
	信号质量阈值	0.01
K值选择规则超参数	残差权重系数	0.4
	锚点惩罚系数	0.02
	几何质量阈值	3.5
	残差质量阈值	3.0
特征提取	CIR降维	17
	CIR窗口长度	70
	轨迹历史窗口	2

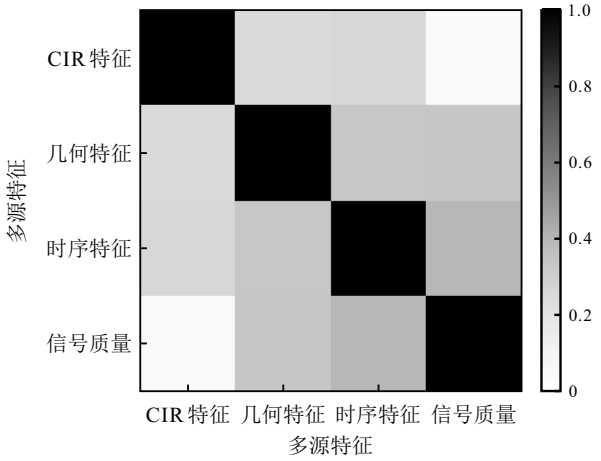

图4 四类特征组之间的 Pearson 相关系数热图

进行降维, 通过对特征协方差矩阵进行特征分解, 得到按方差贡献率从大到小排列的主成分. 如图 5 所示, 第一主成分的方差贡献率为 30.4%. 本文最终选择保留前 17 个主成分, 其累计方差贡献率为 98.16%, 在大幅降维的同时几乎完整保留原始特征的信息内容.

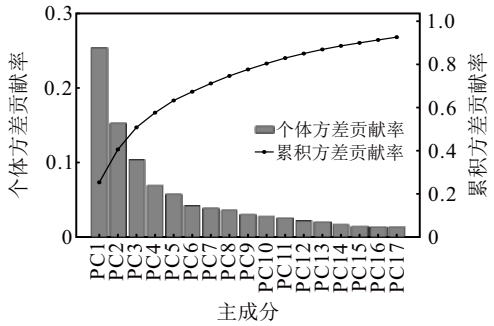

图5 CIR PCA 主成分方差贡献度分析

5.1.2 DDPPO 定位性能

为全面评估所提出的 DDPPO 方法的定位性能, 本文在四个复杂程度不同的室内环境中进行了系统性测试. 每个环境均采用空间均匀采样策略保证测试集的代表性.

从图 6 可以观察到, 在环境相对简单的 Env1 中, 智能体的选择高度集中在$K=5$, 能够学习到在此环境下中等数量的锚点能够在定位精度、几何构型与信号质量之间达成最佳平衡. 在最复杂的 Env2 中, 智能体以$K=5$为主导选择, 同时保留相当比例的$K=7$选择, 在保证几何足够鲁棒的前提下优先筛选信号质量更优的锚点子集, 而非简单地最大化锚点数量. 在复杂度适中的 Env0 与 Env3 中, 智能体绝大多数情况下分别采用$K=7$与$K=8$的选择策略, 通过增加锚点数量来提升几何多样性与信号冗余保障定位的稳定性, 同时在局部特别困难或简单的区域会根据智能体感知到实时状态的几何构型与信号质量进行锚点的取舍. 这些差异化的策略分布表明了智能体可以通过模型的学习, 根据实时感知信息动态地自适应调整锚点数量.

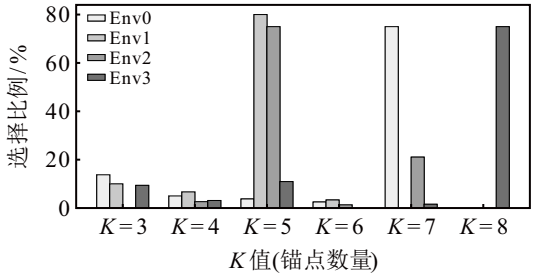

图6 K值分布示意图

由表 2 可知, DDPPO 在四种环境下均取得了较高的定位精度, 平均 MAE 为 0.219 米, 平均 RMSE 为 0.242 米, 平均 MdPE 为 0.211 米. 其中在 Env1 中表现最佳, MAE 为 0.106 米; 即使在最具挑战性的 Env2 中, 仍能保持 0.354 米的 MAE, 表明该方法在

复杂场景下具有良好的定位鲁棒性.

表2　DDPPO 在多环境下的定位性能统计 (m)

环境	MAE	RMSE	MdPE
Env0	0.183	0.195	0.185
Env1	0.106	0.132	0.085
Env2	0.354	0.392	0.333
Env3	0.233	0.250	0.243
平均	0.219	0.242	0.211

5.2　消融实验

为验证本文提出的 DDPPO 模型有效性, 设计锚点选择策略消融和多源特征消融两类消融实验, 以检验各核心模块的必要性.

5.2.1　锚点选择策略消融

实验共设置三组对照包括: 组别 1 全锚点选择[24]、组别 2 MLP 视距选择[25] 以及组别 3 单层 PPO 策略. 所有方法均采用相同的特征提取流程、数据标准化处理及 WLS 定位算法, 并固定随机种子以保证测距噪声一致以准确反映不同锚点选择策略对定位精度的影响. 组别 1 采用全锚点选择策略, 直接使用全部锚点参与 WLS 定位; 组别 2 使用与双层 PPO 结构相同的 MLP 网络, 输出 LoS 锚点组合; 组别 3 则基于单层 PPO 进行锚点选择, 该设置移除了 K 值选择网络与课程学习策略, 其动作空间覆盖所有可能的锚点组合.

从表 3 可以看出, DDPPO 在所有四个环境中表现均优于其他锚点选择策略. 与全锚点选择策略相比 MAE 降低 79.3%. 在环境最复杂的 Env2 中 MAE 降幅达 87.7%, 因为该方法对 NLoS 干扰与多径效应具有较强的适应能力, 在恶劣环境下具较好鲁棒性. 此外, 相较于基于 MLP 的视距锚点选择方法, DDPPO 将 MAE 降低了 67.2%. 与单层 PPO 锚点选择策略相比, MAE 降幅也达到 75.1%.

表3　消融实验结果对比 (m)

环境	消融组别	MAE(m)	RMSE	MdPE
Env0	组别1	0.396	0.464	0.305
	组别2	0.391	0.541	0.270
	组别3	0.453	0.567	0.356
	Proposed	0.183	0.195	0.185
Env1	组别1	0.279	0.329	0.249
	组别2	0.298	0.406	0.197
	组别3	0.548	0.666	0.491
	Proposed	0.106	0.132	0.085
Env2	组别1	2.881	3.122 1	2.763
	组别2	1.489	4.196	0.668
	组别3	1.936	3.187	1.344
	Proposed	0.354	0.392	0.333
Env3	组别1	0.663	0.822	0.524
	组别2	0.490	0.573	0.404
	组别3	0.582	0.724	0.509
	Proposed	0.233	0.250	0.243

5.2.2　多源特征消融

在保持其它训练条件不变的前提下, 本文分别移除 CIR 特征、时序特征、几何特征或信号质量特征, 重新训练 DDPPO 模型并在相同的测试集上评估定位性能. 从表 4 可以看出, 本文所采用多源特征的任一缺失都会导致定位精度下降, 各个特征都贡献了有效信息. 其中, 无几何特征平均误差增幅 170.3% 对定位精度影响最大, CIR 特征的缺失在最复杂的 Env2 中影响最大, 时序特征和信号质量特征的缺失分别造成定位精度 116.9% 和 102.7% 的损失. 这表明多源状态空间中各类特征的相对尺度通过归一化处理得到有效平衡, 特征间信息冗余度低.

表4　多源状态空间特征消融实验结果 (m)

消融配置	Env0	Env1	Env2	Env3	平均
无CIR	0.292	0.169	1.484	0.353	0.575
无时序	0.682	0.186	0.672	0.361	0.475
无几何	1.049	0.166	0.793	0.361	0.592
无信号质量	0.406	0.166	0.849	0.355	0.444
Proposed	**0.183**	**0.106**	**0.354**	**0.233**	**0.219**

5.3　奖励函数阈值敏感性分析

本文对奖励函数中 GDOP 阈值和信号质量阈值这两个关键超参数进行了敏感性分析. 在四个环境中分别对两阈值进行线性扫描, 观察定位误差 MAE 值的变化情况, 实验过程中保持其它参数不变. 图 7 的实验结果表明, 在最复杂的 Env2 场景中, 阈值变

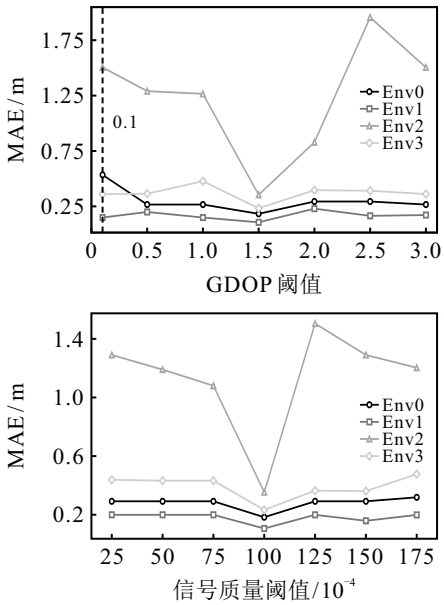

图7　奖励函数阈值敏感性分析

化对定位误差影响较大, 本文所选择阈值能够获得最低误差. 在另外三个场景中, 定位误差随阈值变化的波动较小, 在本文所选阈值附近多数环境均表现出较好的性能. 这些结果说明本文所选阈值能够有效平衡信号质量过滤与可用锚点数量.

5.4 对比实验

为全面评估所提方法的性能, 本文选取了近年来具有代表性的 UWB 室内定位方法进行对比实验. 具体包括以下几类方法: (1) 强化学习方法: 包括采用 DQN 进行锚点选择的开创性工作 DQN-PedLoc[13] 及自监督改进 SS-DDPG[14]; (2) 主流深度学习方法: 基于 Transformer 架构的 F-BERT[26]、基于特征解耦生成的 IIns-VAE[27] 以及结合时序卷积与对比学习的 SimCLR-SC[28]; (3) 混合模型方法: AI-EKF[29]. 所有方法均在相同的四个环境和相同的测试集上进行评估, 以确保对比的公平性.

从表 5 可以看出, 本文所提方法在所有环境中均取得最优性能, 平均 MAE 为 0.219m. 相比强化学习方法 DQN-PedLoc、SS-DDPG, 平均 MAE 分别降低 63.1% 和 66.8%. 这说明双层架构有效解决了大动作空间的收敛难题. 相比深度学习方法 F-BERT、IIns-VAE、SimCLR-SC, 平均 MAE 分别降低 45.9%、63.3% 和 66.2%. , 可以发现纯数据驱动方法在 Env2 中误差激增, 而 DDPPO 仅为 0.354m 且相对稳定, 证明了几何规则约束的重要性. 相比混合方法 AI-EKF, 平均 MAE 降低 72.5%, 避免了 LoS 误判导致的误差累积. 此外, 从不同环境角度看, 该方法在 Env1 中提升最为显著, 改进幅度达 54.5%; 在最为复杂的 Env2 中则展现出最强的鲁棒性, 能有效应对 NLoS 干扰. 整体上, DDPPO 在四个环境中的 MAE 标准差仅为 0.090 m, 明显低于所有对比方法.

表5 四个环境 MAE 对比实验结果 (m)

方法	Env0	Env1	Env2	Env3	平均
DQN-PedLoc[13]	0.597	0.349	0.819	0.612	0.594
SS-DDPG[14]	0.530	0.298	1.265	0.547	0.660
F-BERT[26]	0.304	0.233	0.657	0.426	0.405
IIns-VAE[27]	0.446	0.277	1.135	0.527	0.596
SimCLR-SC[28]	0.303	0.310	1.605	0.371	0.647
AI-EKF[29]	0.579	0.417	1.342	0.841	0.795
Proposed	**0.183**	**0.106**	**0.354**	**0.233**	**0.219**

图 8 直观对比了七种方法在四个环境中的定位轨迹. 可以看出, 本文所提方法在所有场景下均能生成与真实轨迹高度贴合、平滑连续的预测路径. 相比之下, 其余方法的预测轨迹则普遍存在明显偏移与波动. 轨迹对比图进一步验证了表 5 中的定量结果, 直观展示了 DDPPO 方法在复杂室内环境下的优越性能.

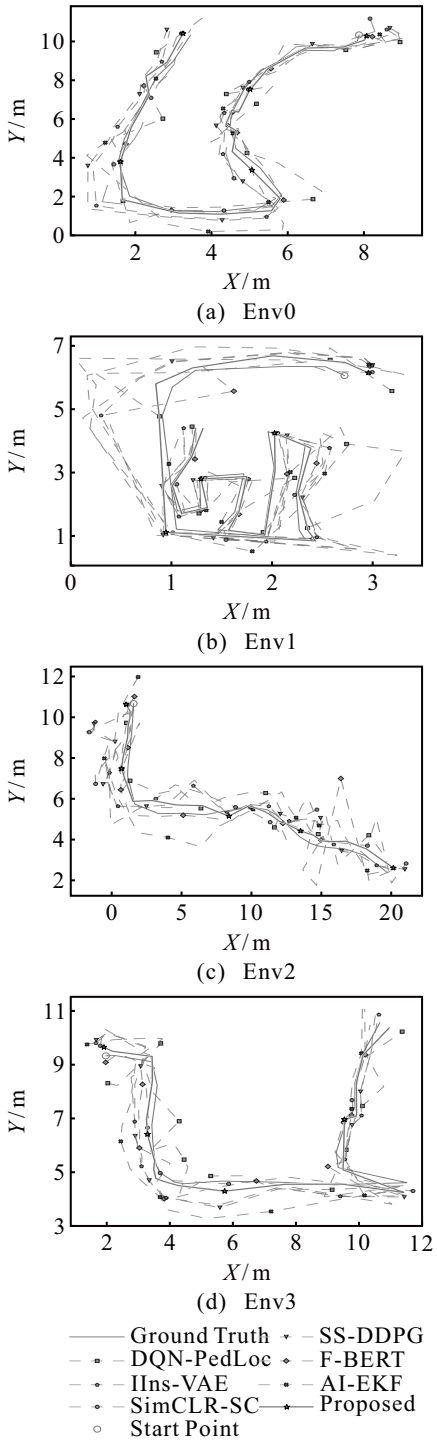

图8 四个环境在不同方法下的轨迹对比图展示

如表 6 所示, 本文统计了各对比方法的训练总时间与单样本测试时间. 结果显示, 本文方法在训练时间上略高于与其余方法, 这是因为双层决策机制所带来的额外训练开销, 但其单次定位时间仅为 4.280 毫秒, 明显低于大多数对比方案. 这表明所提方法在保持高定位精度的同时, 能够较好地满足实时性需求, 从而实现了定位精度与计算效率之间的有效平衡.

表6　各方法训练与测试时间对比

方法	训练时间(min)	测试时间(ms)
DQN-PedLoc[13]	3.49	1.287
SS-DDPG[14]	14.74	4.562
F-BERT[26]	0.58	6.280
IIns-VAE[27]	13.76	4.958
SimCLR-SC[28]	2.43	14.063
AI-EKF[29]	3.11	50.791
Proposed	15.88	4.280

6　总结与展望

本文针对复杂室内环境下 UWB 定位系统因静态锚点选择策略导致的性能瓶颈, 提出了一种基于 DDPPO 的自适应锚点选择定位方法. 该方法创新性地将锚点选择建模为序贯决策过程, 通过设计层次化的 PPO 决策架构, 将锚点数量与具体组合选择解耦, 有效克服动作空间组合爆炸的难题. 同时构建融合多源状态信息与自适应多目标奖励的优化机制, 实现信号质量与几何分布的实时动态平衡. 决策输出的锚点组合会直接使用 WLS 物理模型进行解算, 定位误差会以奖励函数主导项的形式反馈给决策网络, 确保了系统在提升环境自适应能力的同时保持可解释性. 在公开数据集的多个复杂环境上的实验表明, 本方法在多种室内场景下均能实现稳定、精准的定位. 为支持研究的可重复性, DDPPO 框架的完整代码已公开发布在 GitHub 代码库中, 链接访问: https://github.com/idoneknow/ddppo_uwb.

未来研究将在本文提出的 DDPPO 框架基础上进一步探索元学习或域对抗机制, 使模型能够适应不同建筑布局与锚点配置的新环境, 并解决新环境信息采集工作量大且困难的问题, 提升方法的泛化能力与实际部署价值.

参考文献 (References)

[1] Liu Y H, Yang Z, Wang X P, et al. Location, localization, and localizability[J]. Journal of Computer Science and Technology, 2010, 25(2): 274-297.

[2] Nkrow R E, Silva B, Boshoff D, et al. NLOS identification and mitigation for time-based indoor localization systems: Survey and future research directions[J]. ACM Computing Surveys, 2024, 56(12): 1-41.

[3] 蒋浩然, 谷丰, 滕天启, 等. 融合距离梯度的单目视觉-惯性-UWB 紧耦合导航定位方法[J]. 控制与决策, 2025, 40(8): 2566-2578.
(Jiang H R, Gu F, Teng T Q, et al. Distance gradient integrated monocular visual-inertial-UWB tightly coupled localization approach[J]. Control and Decision, 2025, 40(8): 2566-2578.)

[4] Xiao Z, Hei Y Q, Yu Q, et al. A survey on impulse-radio UWB localization[J]. Science China Information Sciences, 2010, 53(7): 1322-1335.

[5] Wu Y F, He X, Mo L F, et al. Self-attention-assisted TinyML with effective representation for UWB NLOS identification[J]. IEEE Internet of Things Journal, 2024, 11(15): 25471-25480.

[6] Wang T Y, Li Y X, Liu J C, et al. Multipath-assisted single-anchor localization *via* deep variational learning[J]. IEEE Transactions on Wireless Communications, 2024, 23(8): 9113-9128.

[7] Bach S H, Khoi P B, Yi S Y. Global UWB system: A high-accuracy mobile robot localization system with tightly coupled integration[J]. IEEE Internet of Things Journal, 2024, 11(9): 16618-16626.

[8] Wang C Y, Ning Y P, Wang J, et al. Optimized deployment of anchors based on GDOP minimization for ultra-wideband positioning[J]. Journal of Spatial Science, 2022, 67(3): 455-472.

[9] Bach S H, Yi S Y. Constrained least-squares trilateration for indoor positioning system under high GDOP condition[J]. IEEE Transactions on Industrial Informatics, 2024, 20(3): 4550-4558.

[10] Fontaine J, Van Herbruggen B, Shahid A, et al. Ultra wideband (UWB) localization using active CIR-based fingerprinting[J]. IEEE Communications Letters, 2023, 27(5): 1322-1326.

[11] Niu Z A, Yang H Z, Zhou L, et al. Deep learning-based ranging error mitigation method for UWB localization system in greenhouse[J]. Computers and Electronics in Agriculture, 2023, 205: 107573.

[12] Li Y X, Mazuelas S, Shen Y. A variational learning approach for concurrent distance estimation and environmental identification[J]. IEEE Transactions on Wireless Communications, 2023, 22(9): 6252-6266.

[13] Kim D H, Park J, Ko Y B, et al. Deep Q-network based UWB anchor and strategy selection for accurate pedestrian localization in vehicular environments[C]. IEEE International Conference on Machine Learning for Communication and Networking. Barcelona, 2025: 1-6.

[14] Coppens D, van Herbruggen B, Shahid A, et al. Removing the need for ground truth UWB data collection: Self-supervised ranging error correction using deep reinforcement learning[J]. IEEE Transactions on Machine Learning in Communications and Networking, 2024, 2: 1615-1627.

[15] Wang R, Xu C, Sun J, et al. Cooperative localization for multi-agents based on reinforcement learning compensated filter[J]. IEEE Journal on Selected Areas in Communications, 2024, 42(10): 2820-2831.

[16] Hajiakhondi-Meybodi Z, Hou M, Mohammadi A. JUNO: Jump-start reinforcement learning-based node selection for UWB indoor localization[C]. GLOBECOM 2022 - 2022 IEEE Global Communications Conference. Rio de Janeiro, 2022: 6194-6199.

[17] 刘林, 宋雨昊. 基于可信度的非视距识别与定位算法[J]. 中国惯性技术学报, 2025, 33(10): 972-978.

(Liu L, Song Y H. NLoS recognition and localization algorithm based on credibility[J]. Journal of Chinese Inertial Technology, 2025, 33(10): 972-978.)

[18] 夏秋, 吴仕勋, 徐凯, 等. 融合时空相关性的元强化学习 GNSS 定位方法[J]. 导航定位学报, 2025, 13(6): 48-56.
(Xia Q, Wu S X, Xu K, et al. Meta-reinforcement learning method for GNSS positioning with integrating spatiotemporal correlation[J]. Journal of Navigation and Positioning, 2025, 13(6): 48-56.)

[19] Rousseeuw P J, Croux C. Alternatives to the median absolute deviation[J]. Journal of the American Statistical Association, 1993, 88(424): 1273-1283.

[20] 王晴, 王浩然, 辛斌, 等. 基于近端策略优化的动态武器目标分配[J]. 控制与决策, DOI: 10.13195/j.kzyjc.2025.0910.
(Wang Q, Wang H R, Xin B, et al. Dynamic weapon-target assignment based on proximal policy optimization[J]. Control and Decision, DOI: 10.13195/j.kzyjc.2025.0910.)

[21] Bregar K. Indoor UWB positioning and position tracking data set[J]. Scientific Data, 2023, 10(1): 744.

[22] 薛晶, 姜苏英, 周强, 等. 结合特征融合与域对抗学习的超宽带非视距识别算法[J]. 计算机工程与应用, DOI: 10.3778/j.issn.1002-8331.2506-0246.
(Xue J, Jiang S Y, Zhou Q, et al. Algorithm for ultra-wideband non-line-of-sight recognition combining feature fusion and domain adversarial learning[J]. Computer Engineering and Applications, DOI: 10.3778/j.issn.1002-8331.2506-0246.)

[23] 郑恩让, 孟鑫, 姜苏英, 等. 基于 1DCNN 和 LSTM 融合的超宽带 NLoS/LoS 识别方法研究[J]. 通信学报, 2025, 46(6): 285-302.
(Zheng E R, Meng X, Jiang S Y, et al. Research on ultra wide band NLoS/LoS recognition method based on the fusion of 1DCNN and LSTM[J]. Journal on Communications, 2025, 46(6): 285-302.)

[24] Liu A, Lin S W, Wang J G, et al. A succinct method for non-line-of-sight mitigation for ultra-wideband indoor positioning system[J]. Sensors, 2022, 22(21): 8247.

[25] 尹振东, 吴芝路, 任广辉, 等. 基于 MLP 神经网络的超宽带信号自适应识别算法的研究[J]. 重庆邮电大学学报: 自然科学版, 2008(2): 156-159.
(Yin Z D, Wu Z L, Ren G H, et al. MLP neural network based adaptive UWB modulation scheme recognition algorithm[J]. Journal of Chongqing University of Posts and Telecommunications: Natural Science Edition), 2008(2): 156-159.)

[26] Yang H C, Wang Y J, Seow C K, et al. Fuzzy transformer machine learning for UWB NLOS identification and ranging mitigation[J]. IEEE Transactions on Instrumentation and Measurement, 2025, 74: 1-17.

[27] Lv H C, Feng J L, Shou H J, et al. UWB localization based on dual-channel neural network and total least square method[J]. IEEE Sensors Journal, 2024, 24(3): 3477-3487.

[28] Wu S X, Wang X, Zhang M, et al. Temporal convolutional neural network UWB positioning method based on SimCLR-CIR-SC autonomous classification[J]. IEEE Sensors Journal, 2025, 25(15): 30161-30174.

[29] Kim D H, Farhad A, Pyun J Y. UWB positioning system based on LSTM classification with mitigated NLOS effects[J]. IEEE Internet of Things Journal, 2023, 10(2): 1822-1835.

作者简介

周元 (1998−), 女, 硕士生, 主要研究方向为室内 UWB 定位, E-mail: zhou_yy2025@163.com;

吴仕勋 (1983−), 男, 副教授, 博士, 主要研究方向为无线通信、无线定位以及人工智能, E-mail: wushixun333@163.com;

蓝章礼 (1973−), 男, 教授, 博士, 主要研究方向为数字图像处理、模式识别及交通信息处理, E-mail: 385551137@qq.com;

徐凯 (1970−), 男, 教授, 硕士, 主要研究方向为模糊控制、自适应控制及智能算法, E-mail: xkxjxwx@hotmail.com;

张淼 (1988−), 男, 副教授, 博士, 主要研究方向为无线网络资源分配、凸优化、人工智能驱动资源优化、智能反射面, E-mail: miao.zhang@cqjtu.edu.cn;

靳双 (1992−), 男, 讲师, 博士, 主要研究方向为智能交通系统、网联自动驾驶及协同控制, E-mail: 884580930@qq.com.