arXiv:2303.02486v2 [cs.RO] 7 Jul 2023

基于注意力的深度强化学习在多人人机团队初始任务分配中的应用

Ruiqi Wang[1], Dezhong Zhao[1,2],和 Byung-Cheol Min[1]

摘要——多人人机团队能够通过具备不同能力和专业知识的人类与机器人协作，高效完成复杂的大规模任务。为了高效运行这类高度异构的团队并及时提升团队绩效，需要制定复杂的初始任务分配策略，充分考虑团队成员和任务之间的个体差异。尽管现有研究在根据智能体状态和绩效重新分配任务方面取得了令人鼓舞的成果，但忽视团队固有的异质性，限制了这些方法在现实场景中的有效性。本文将多人人机团队中的初始任务分配问题新颖地建模为一个上下文多属性决策过程，并提出一种基于注意力的深度强化学习方法。我们引入跨属性注意力模块，以编码状态表示中多个属性之间潜在且复杂的依赖关系。我们在大规模威胁监视场景中开展案例研究，展示了该模型的优势。

I. 引言

人机团队通过将自主系统的一致性和精确性等优势，与人类所具备的创造力和适应性相结合，在各个领域都能带来显著益处 [1]。对更高效地解决复杂且大规模任务（如灾害响应、搜索与救援以及环境监测）的需求日益增长，推动了人们对多人类-多机器人（MH-MR）团队的关注 [2]。MH-MR团队中具有不同能力、专业知识和特征的多个人类与机器人开展协作，为提升团队互补性、生产力和多功能性带来了巨大潜力 [3], [4]。然而，这类团队的高度异质性也导致在高效协调智能体方面面临更多运行挑战。因此，要充分发挥MH-MR团队的潜力，必须解决异构智能体之间的最优任务分配这一关键挑战。

尽管在多机器人系统的上下文中，任务分配已得到充分研究 [5]–[7] ，以及人类-自主性
涉及MH-MR团队的协作 [8]–[10], 场景受到的关注相对较少。在有限的相关文献中，大多数研究仅关注过程内任务分配，并将人类和机器人的状态以及任务绩效指标作为判断依据。然而，这些

[1]美国印第安纳州西拉斐特，普渡大学计算机与信息技术系SMART实验室。[wang5357,minb]@purdue.edu.[2]中国北京，北京化工大学机电工程学院。DZ_Zhao@buct.edu.cn.

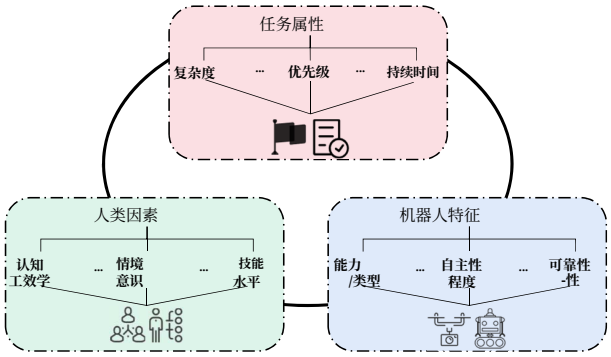


图 1. 用于构建 MH-MR团队初始任务分配问题的上下文多属性决策过程示意图。

研究往往忽视了初始任务调度[11]–[14]，而这种调度会考虑人类、机器人和任务之间的个体差异。初始任务安排会显著影响团队绩效。恰当的初始分配能够发挥MH-MR团队固有的异质性，确保一开始就将任务分配给能力和技能相匹配的个体，从而在更短时间内取得最佳表现。反之，不合理的设置会妨碍团队通过过程内任务再分配来优化表现，即使经过较长时间也难以改善。

为弥补文献中的空白，我们旨在探索多人人机团队中的初始任务分配（ITA）问题。具体而言，我们希望确定如何在任务开始时，将一项由一组任务组成的任务（其中各任务具有不同属性）最优地分配给一个由能力各异的人类和特征各异的机器人组成的团队。如图 1 所示，我们通过将该问题表述为上下文多属性决策过程（CMADP） [15]来进行研究。此过程需要考虑一个包含三大类属性的上下文并据此作出决策：人类因素、机器人特征和任务属性；每一类属性都包含多个子属性，而这些子属性由各个变量构成 [16]。然而，确定初始任务分配的最优决策面临两大挑战。首先，决策并非仅由多种属性和子属性决定，还受到它们之间潜在依赖关系与相互作用的影响 [17]。其次，不同关系的重要性是动态的，会随不同任务场景而变化。

为应对这些挑战，我们提出了一种基于**A**ttention的深度**R**einforcement**L**earning方法（AtRL）。AtRL的架构与流程如图2所示。我们模型的核心是跨属性注意力状态表示学习，旨在学习一个经过自适应编码相关信息的多属性上下文全局状态

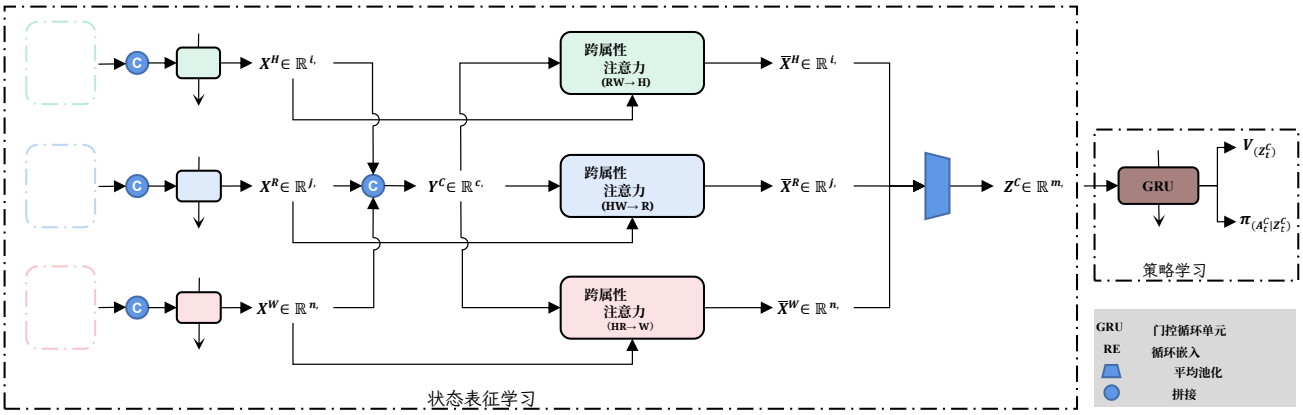


图 2. 所提出的 AtRL 框架示意图。由 $c^{(\cdot)(*)} \in \mathbb{R}^{L,D}$ 组成的多属性数据输入首先分别馈入循环嵌入层 $RE_{(\cdot)}$，生成三个维度相同的属性序列 $X^{(\cdot)} \in \mathbb{R}^{L_{(\cdot)},d}$ d（第 IV-B 节）；其中，$c^{(\cdot)(*)}$ 表示 $.^{th}$ 属性类别中的 $*^{th}$ 子属性序列，而 L 和 D 分别表示该序列的长度和维度。随后，将这些序列拼接起来，生成低层状态表示 $Y^C \in \mathbb{R}^{c,d}$，并分别将其与各属性序列 $X^{(\cdot)}$ 一同输入跨属性注意力层。在每个跨属性注意力层中，通过计算当前属性特征与在 Y^C 中编码的特征之间的自适应依赖关系，利用其他两个属性中揭示的相关信息来调整每个属性序列（第 IV-C 节）。最后，将增强属性序列 $X^{(\cdot)}$ 输入均值池化层，生成多属性上下文的高层状态表示 $Z^C \in \mathbb{R}^{m,d}$，并将其传递给策略网络，以学习价值函数 V 和策略 π（第 IV-D 节）。

表示。具体而言，跨属性注意力层自适应地捕捉不同属性类别之间的潜在依赖关系，持续利用另外两个属性中高度相关的信息来增强某一特定属性内的特征。随后，增强后的属性特征序列经过均值池化层，生成全局状态表示。为评估AtRL及其中的跨属性注意力的性能，我们在大规模威胁监视与识别任务场景中开展了一项案例研究。

主要贡献可总结如下：1）我们率先研究了MH-MR团队中的初始任务分配问题，并对上下文多属性决策过程进行了问题建模；2）为应对上下文多属性决策过程带来的挑战，我们提出了一种新颖的基于学习的框架，其中包含跨属性注意力表示学习模块，可自适应地捕捉多属性上下文中的潜在依赖关系；3）通过涉及大规模威胁监视任务场景的案例研究，我们证明了AtRL 及其内部跨属性注意力机制的有效性。

II. 背景与预备知识

相关工作。 多个人类与多台机器人（MH-MR）团队由多个人类智能体和多台机器人组成，它们协同工作，以实现由不同任务构成的共同目标 [2]。这类团队具有很强的异质性，因为人类智能体与机器人在能力、技能和特征方面存在显著差异。例如，人类智能体的认知抗疲劳能力和工作负荷各不相同，操作技能和决策能力也各有差异。此外，MH-MR团队中的机器人通常类型多样，例如在野外搜救行动中同时使用无人机（UAV）和无人地面车辆（UGV），因此呈现出不同的机器人特征。要充分利用这种高度异质性并最大限度地提升团队绩效，就需要采用复杂的初始任务分配策略，确保从一开始就为个体分配与其能力和技能相匹配的任务。

遗憾的是，现有文献有限，且大多数针对MH-MR团队的工作负载分配方法都忽视了这种内在异质性，仅考虑根据每个人类智能体和机器人的状态与绩效进行过程中的工作负载再分配 [11]–[14]。例如，Mina 等人 [11]开发了一种自适应工作负载分配系统，该系统会监测人类认知工作负载、机器人状态和个体绩效，并据此调整工作负载，确保每个智能体都处于最佳工作状态。尽管这类过程中的任务分配方法取得了令人鼓舞的成果，但未能考虑 MH-MR团队的根本异质性，可能会降低其在实际场景中的有效性。尤其是，如果初始工作负载分配不合理，那么即使经过较长时间，也很难通过过程中的任务再分配来优化团队绩效。另一方面，鲜有研究考虑 MH-MR团队在任务开始时的任务分配。例如，Humann 等人 [18] 构建了一个多无人机多操作员监视系统模型，模拟人类操作员数量和两种类型的无人机对团队绩效的影响。然而，该模型未能充分刻画人类个体差异，更重要的是，没有提供优化初始工作负载分配的方法。为弥补这些不足，我们探索了兼顾MH-MR团队内在异质性的最优初始任务分配问题。我们对情境多属性决策过程进行问题建模，并提出一种基于注意力的强化学习方案来解决这一问题。

状态表征学习。状态表征学习是强化学习中的一个关键问题，旨在自动提取有用且紧凑的环境状态表示，以促进策略学习过程，尤其是在状态空间结构庞大而复杂时 [19]。近期研究表明，基于注意力的神经网络（如图注意力网络和 Transformer层） [20], 能够通过将原始感官输入映射为低维状态表示，有效推理状态空间，并捕捉环境中最相关的特征。随后，可将学得的状态表示作为策略网络的输入，以学习最优策略。

近端策略优化。在本研究中， [21] 采用近端策略优化（PPO）作为训练算法。PPO中的策略优化目标可表示为以下三个损失项的加权和：用于策略改进的裁剪代理目标（$\mathcal{L}^{CP}$）、价值函数损失（$\mathcal{L}^{V}$）以及用于正则化的熵损失（$\mathcal{L}^{O}$），形式化表示为：

$$\mathcal{L}(\theta,\phi) = w_p\mathcal{L}^{CP}(\theta) - w_v\mathcal{L}^{V}(\phi) + w_e\mathcal{L}^{O}(\theta) \quad (1)$$

with

$$\begin{aligned}\mathcal{L}^{CP}(\theta) &= \mathbb{E}_t[\sum_{0}^{t}\min\left(r_t(\theta)\hat{A}_t,\right.\\ &\qquad \left.\text{clip}\left(r_t(\theta),1-\epsilon,1+\epsilon\right)\hat{A}_t\right)]\\ \mathcal{L}^{V}(\phi) &= \mathbb{E}_t[\sum_{0}^{t}\left(V_\phi\left(s_t\right)-\hat{A}_t\right)^2]\\ \mathcal{L}^{O}(\theta) &= \mathbb{E}_t[\sum_{0}^{t}\mathcal{O}\left(\pi_\theta\left(a_t\mid s_t\right)\right)]\end{aligned} \quad (2)$$

其中， w_p、 w_v和 w_e 分别表示裁剪策略损失、价值函数损失和熵损失的损失系数。 A 表示优势函数的估计值， ϵ 为裁剪比率， $\mathcal{O}(\cdot)$ 表示熵， $r_t(\theta)$ 为重要性采样比率，定义如下：

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \quad (3)$$

其中， $\pi_{\theta_{\text{old}}}$ 为先前策略。

III. 问题建模

我们将MH-MR团队中的初始任务分配（ITA）问题建模为上下文多属性决策过程（CMADP），其定义为一个元组（W， α， C， A， $\mathcal{T}$， R），其中：

- W:={$w_1,\ldots,w_n$} 是一组属性各异的有限 n 任务
- $\overline{\alpha}$:={$\alpha_1,\ldots,\alpha_i,\alpha_1,\ldots,\alpha_j$} 是一组能力各异的有限 i人类智能体和特征各异的 j 机器人
- C := {$C^H \times C^R \times C^W$} 是观测到的联合多属性上下文，包括以 h 个维度测量的 i 名人类智能体的联合因素， C^H:={{$c_1^{H_1},\ldots,c_i^{H_1}$} $\times\ldots\times$ {$c_1^{H_h},\ldots,c_i^{H_h}$}}，以及以 r 个维度表示的 j 个机器人， C^R:={{$c_1^{R_1},\ldots,c_j^{R_1}$} $\times\ldots\times$ {$c_1^{R_r},\ldots,c_j^{R_r}$}}，以及以 w 个维度评估的 n 项任务的联合属性， C^W:={{$c_1^{W_1},\ldots,c_n^{W_1}$} $\times\ldots\times$ {$c_1^{W_w},\ldots,c_n^{W_w}$}}
- A:=(a_1 $\times$) 是通过将 w_n 分配给 α_i 和/或 α_j而形成的联合分配决策动作
- $\mathcal{T}$:= $Pr(C' \mid C)$ 是一个随机的上下文观测转移概率
- R:= $f_R(C,A)$ = $\mathbb{E}[R_t(C_t,A_t)]$ = $\mathbb{E}[\sum_{k=1}^{n} r_k(c_t,a_t)]$ 是奖励函数，当在时间步 t观测到联合上下文 C_t 并采取联合动作 A_t 后，该函数会给出即时奖励R_t 。它是每个已分配任务 n的个体奖励 r_n 之和，反映了任务 n的个体任务表现。

与上下文多臂老虎机问题 [22],类似，上下文多属性决策过程可以视为玩家（ITA策略）与对手（上下文选择器）之间的重复博弈。在每个时间步 t，对手根据博弈历史 H_{t-1} =选择一个多属性上下文 $C_t \in C$ [(C_1,A_1,R_1) $\ldots,(C_{t-1},A_t,R_{t-1})$],，并且同一上下文可以被多次选中。玩家根据其策略 $\pi(A_t|C_t)$选择一个ITA动作 $A_t \in A$ ，并获得即时奖励 R_t，该奖励仅取决于(C_t,A_t)。玩家的目标是找到使期望奖励$\mathbb{E}[\sum_{s=1}^{t} f_R(C_s,A_s)]$最大化的最优策略 π^*： $\mathbf{C}_t \mapsto \mathbf{A}_t$。

该公式能够对 MH-MR 团队中 ITA问题 的上下文动态进行基础建模，并明确界定了影响决策过程的各类属性，包括人类因素、机器人特征和任务特定因素。值得注意的是，我们的上下文多属性决策过程不同于传统的马尔可夫决策过程（MDP）：初始工作负载决策动作不会影响上下文（状态），而上下文（状态）体现了团队和任务固有的异质性。然而，由对手概念引入的上下文观测转移概率 $\mathcal{T}$，发挥着与状态转移函数相似的作用。

IV. 方法

本节将介绍我们提出的 AtRL，用于多人人机团队的初始任务分配。

A. 概述

我们的目标是解决上一节中建模的上下文多属性决策问题，并训练一种分配策略，使其能够根据人类智能体、机器人和任务固有的异质性，确定最优的初始任务分布；这种异质性被定义为多属性上下文。然而，由于多属性上下文的状态空间十分复杂，且隐含地包含不同属性和子属性之间的依赖关系，因此，学习一种能够高效且自适应地建模此类潜在交互的状态表示至关重要。为此，我们引入跨属性注意力来进行状态表征学习，从而更好地理解多属性上下文，并加速策略学习过程。

B. 循环嵌入

考虑一个由 i 个人类智能体和 j 台机器人组成、负责完成 n 项任务的MH-MR团队。如第III节的问题建模所述，联合多属性上下文包含三个主要属性：以 h 个维度衡量的 i 个人类智能体的因素、以 r 个维度描述的 j 台机器人的特征，以及以 w 个维度评估的 n 项任务的属性。

形式上，我们可以得到三个原始属性序列：$C^H \in \mathbb{R}^{i,\ h}$、$C^R \in \mathbb{R}^{j,\ r}$ 和 $C^W \in \mathbb{R}^{n,\ w}$，如下所示：

$$\begin{aligned} C^H &= \circ\left(C^{H_1}, C^{H_2}, \ldots, C^{H_h}\right) \\ C^R &= \circ\left(C^{R_1}, C^{R_2}, \ldots, C^{R_r}\right) \\ C^W &= \circ\left(C^{W_1}, C^{W_2}, \ldots, C^{W_w}\right) \end{aligned} \tag{4}$$

其中，$\circ$ 表示拼接操作；$C^{H_h} \in \mathbb{R}^{i,\ 1}$、$C^{R_r} \in \mathbb{R}^{j,\ 1}$和 $C^{W_w} \in \mathbb{R}^{n,\ 1}$ 分别指三个属性类别中子属性的序列数据。

然后，对每个原始属性序列进行非线性变换操作，例如通过全连接层应用Tanh激活函数，再将其输入含有 d 个隐藏单元的RNN单元：

$$\begin{aligned} \hat{C}_i^H &= \mathrm{RNN}\left(\hat{C}_{i-1}^H, Tr\left(C_i^H\right)\right) \\ \hat{C}_j^R &= \mathrm{RNN}\left(\hat{C}_{j-1}^R, Tr\left(C_j^R\right)\right) \\ \hat{C}_n^W &= \mathrm{RNN}\left(\hat{C}_{n-1}^W, Tr\left(C_n^W\right)\right) \end{aligned} \tag{5}$$

其中，$\hat{C}_i^H$、$\hat{C}_j^R$和$\hat{C}_n^W$分别表示最终时间步的隐藏状态，Tr 表示非线性变换操作。需要注意的是，这三个RNN的输入维度各不相同。

我们将这些最终隐藏状态视为低层属性序列 $X^H = C_i^H \in \mathbb{R}^{i\ d}$ $X^R = C_j^R \in \mathbb{R}^{j\ d}$ $X^W = \hat{C}_n^W \in \mathbb{R}^{n,\ d}$。每个序列旨在获取各属性的低层特征，同时考虑空间信息；鉴于任务分配决策依赖于顺序，这一点至关重要。还需注意的是，经过循环嵌入后，三个属性序列具有相同的维度 d，因此在后续的跨属性注意力机制中，点积操作在数学上是可行的。

随后，将这三个属性序列拼接起来，生成低层多属性状态表示 $Y^C \in \mathbb{R}^{c,\ d}$ 如下：

$$Y^C = \circ\left(X^H, X^R, X^W\right). \tag{6}$$

C. 跨属性注意力增强状态表示

跨属性注意力旨在全面、自适应地将人类因素、机器人特征和任务属性之间的潜在相关信息编码到多属性上下文的状态表示中。具体而言，在跨属性注意力层中，会计算属性序列X 与低层多属性状态表示之间的注意力分数

Y^C，用于指导单属性特征的调整

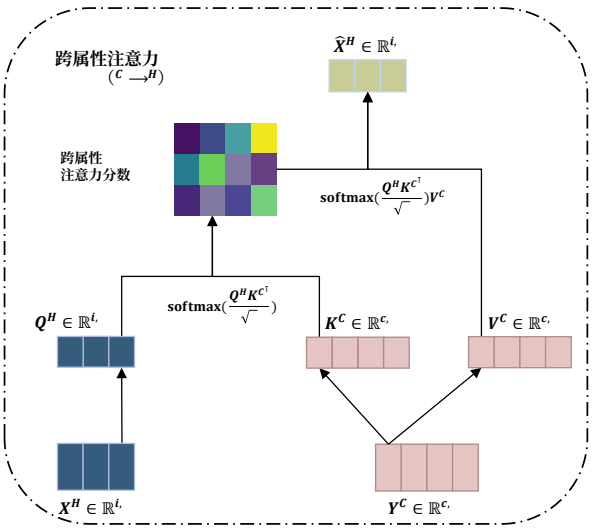


图3. 跨属性注意力机制示例。

方法是将状态表示中嵌入的另外两种属性的相关信息纳入其中。

下面以人类因素的属性序列 $X^H \in \mathbb{R}^{i,\ d}$作为单属性输入示例，说明跨属性机制的工作原理。形式化地，查询 Q_H、键 K_C 和值 V_C 定义为

as:

$$\begin{aligned} Q^H &= X^H \cdot W^{Q^H} \\ K^C &= Y^C \cdot W^{K^C} \\ V^C &= Y^C \cdot W^{V^C} \end{aligned} \tag{7}$$

其中，$W^{Q^H} \in \mathbb{R}^{d,\ q}$、$W^{K^C} \in \mathbb{R}^{d,\ k}$以及 $W^{V^C} \in \mathbb{R}^{d,\ v}$表示三组可训练权重。

与 [23], 中使用的自注意力机制类似，我们的跨属性注意力允许相关信息从多属性表示流向人类因素的属性序列，从而揭示属性之间的潜在依赖关系，如图3所示：

$$\begin{aligned} \hat{X}^H &= CA(Y^C \rightarrow X^H) \\ &= \mathrm{Attention}(Q^H, K^C, V^C) \\ &= \mathrm{softmax}\left(\frac{Q^H \cdot K^{C\top}}{\sqrt{k}}\right) V^C. \end{aligned} \tag{8}$$

该过程可以并行执行多次，以实现多头跨属性注意力。随后，可以得到增强的人类因素属性序列$\overline{X}^H$，如下所示：

$$\overline{X}^H = \mathrm{LN}\left(\mathrm{FFN}\left(\hat{X}^H\right) + X^{(H)}\right) \tag{9}$$

其中，LN 表示层归一化操作，FFN 表示前馈层。

同样，通过应用相同的步骤，我们可以获得另外两个属性序列 X^R 和 X^W（机器人特征和任务属性）的增强特征表示。最后，通过平均池化层聚合这三个序列，生成高层状态

表示，用于表征多属性上下文 Z^C：

$$f_\vartheta\left(Z^C\right)=\mathrm{MP}\left(\overline{X}^H,\overline{X}^R,\overline{X}^W\right) \tag{10}$$

其中，MP表示平均池化操作。为简洁起见，我们将学习得到的多属性状态表示记为$f_\vartheta\left(Z^C\right)$，其中 ϑ 表示跨属性状态表征学习中所有可训练参数的集合。

该均值池化层十分有用，因为全局状态表示可能具有较高维度，尤其是在处理大规模或高度异构的多人人机团队和任务时。该层可将由增强属性序列拼接而成的高维全局状态表示映射到低维空间，从而使后续策略学习在计算上可行。

D. 策略学习

在获得经跨属性注意力增强的状态表示 Z^C 后，我们将其输入策略网络。由于状态表示具有序列性和复杂性，我们采用门控循环单元（GRU）而非传统全连接层作为策略网络。这使智能体能够更好地捕捉状态表示中的动态模式，并学习最优ITA策略。

按照第 III 节中的表述，在策略学习的每个步骤中，智能体都会获得通过跨属性学习得到的多属性状态表示 Z_t^C。基于当前策略 π_θ，它会选择联合分配策略 A_t，将任务分配给团队中的每个人类智能体和机器人，并从奖励函数 $f_R(C,\ A)$中获得奖励 R_t；该奖励函数用于评估所选任务分配下团队的表现。智能体使用 PPO 优化其策略和价值函数的参数集 θ、ϕ，以最大化式 1 中描述的目标函数。此外，状态表征学习中的参数集 ϑ 也作为价值函数的一部分进行优化。

V. 威胁监视场景中的案例研究

本节介绍我们的案例研究，旨在将所提出的 AtRL 应用于大规模威胁监视任务场景[1]并对其进行验证。

A. 任务场景

我们在案例研究中考虑的应用领域是使用 MH-MR 团队对一个区域内的未知威胁进行大规模监视与分类。这种场景在军事和环境应用中很常见，潜在威胁（如敌方装备或危险材料）需要受到监视并在更近距离进行识别

距离比卫星图像更远。

在我们的案例研究中，MH-MR团队的目标是结合

[1]更多详情请见 https://sites.google.com/ view/ITA-AtRL

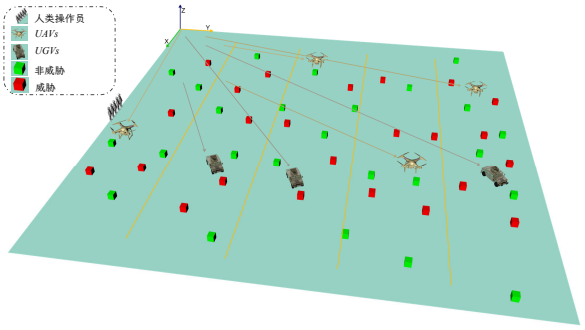


图 4. 仿真环境示意图，其中包含人类操作员、无人机、无人地面车辆、非威胁（绿色立方体）和兴趣点威胁（红色立方体）。

人类操作员与两种类型的机器人，对指定区域内的潜在威胁进行大规模监视与分类。任务始于卫星系统检测并生成一份包含多个兴趣点的清单。随后，指挥系统会将机器人分配至这些已报告的兴趣点，以收集近距离视觉信息；人类操作员则利用这些信息判断每个兴趣点是否构成潜在威胁。多台机器人和多名人类操作员可以同时工作，以加快这一流程。

为体现 MH-MR团队的异质性，我们考虑人类操作员、机器人和任务的多种属性。在人类因素方面，我们重点关注认知能力和操作技能水平，这两者是影响人机团队中人类绩效的两个主要因素 [24]。认知能力衡量个体抵御疲劳和工作负荷的能力，而操作技能水平指个体通过训练和经验获得的、在特定任务或一组任务中的技能或专业知识水平。在机器人特征方面，我们考虑两种类型的机器人：无人机（UAVs）和无人地面车辆（UGVs）。无人机能够快速移动，并从一定高度俯瞰兴趣点；无人地面车辆移动较慢，但能为人类提供更清晰的视野。最后，我们将兴趣点的位置信息和威胁分类的难度视为任务属性。

B. 仿真环境

为在上述任务场景中验证我们的 AtRL，我们基于 [18] 中使用的仿真环境，利用 GAMA 平台 [25]，创建了一个仿真环境。如图 4 所示，待监控区域是一片面积为 $2\ km\times 2\ km$ 的开放空间，其中包含若干兴趣点。每个兴趣点都包含一个需要识别的威胁或非威胁对象。

1) 机器人模型： 两种不同的机器人，即无人机和无人地面车辆，具有不同的规格，例如移动速度和所提供图像的质量，这些规格总结于表 I 中。然而，它们的移动模式和任务相同。

表 I 仿真中设置的无人机和无人地面车辆规格。

	UAV	UGV
速度	20m/s	8m/s
图像质量	low	high

在每项任务开始时，都会为机器人分配一个初始兴趣点。机器人从起点自动导航至该兴趣点，并在那里停留 3 秒以捕捉威胁的图像。随后，它将图像发送给一名人类操作员进行分类。机器人继续前往尚未访问过的最近兴趣点，并重复图像捕捉过程。当所有兴趣点都探索完毕后，机器人返回起点。

2) 人类绩效模型：我们将人类操作员建模为一个事件服务器，负责处理由无人机和无人地面车辆生成的图像分类任务。基于关于人类在复杂任务中绩效的实证研究结果 [26]，，我们假设人类服务器按序处理事件。假设二元威胁识别任务的绩效会受到人类操作员疲劳和工作负荷的影响。这种影响可以通过操作员的个体认知能力进行修正。图像分类任务的任务固有难度也会对其产生影响，而该难度取决于图像质量和威胁类型。任务难度水平可根据操作员的个人操作技能水平进行加权。操作员正确进行图像分类的概率会受到这些因素的非线性影响 [18]，具体如下：

$$Pr_c = \frac{1}{2} + \gamma(F_f F_w) \cdot \xi(F_s) \tag{11}$$

其中，γ，$\xi \in (0, \frac{\sqrt{2}}{2})$分别表示个体认知能力和操作技能水平的修正权重，而 F_f，$F_w \in [0,1]$和 F_s：$\in (0, 1)$分别表示疲劳、工作负荷和任务固有难度的修正因子。

这种公式使分类成功概率的最低值能够无限接近 0.5，这与二元选择中的随机猜测相一致。基于 [16]，[18]，修正因子的具体模式和值设定如下。有关理论和实证依据的更多细节，请参阅这些参考文献。疲劳和工作负荷的修正因子可定义为：

$$F_f(\hat{t}) = \begin{cases} 1 & 0 \le \hat{t} < 1 \\ -0.12\hat{t} + 1.12 & 1 \le \hat{t} \le 8 \end{cases} \tag{12}$$

$$F_w(u) = \begin{cases} -2.47u^2 + 2.22u + 0.5 & 0 \le u < 0.45 \\ 1 & 0.45 \le u < 0.65 \\ -4.08u^2 + 5.31u - 0.724 & 0.65 \le u \le 1.0 \end{cases} \tag{13}$$

其中，$\hat{t}$ 是人类智能体的工作时长，u 表示人类智能体的利用率，即操作员在过去 5 分钟内执行任务而处于忙碌状态的时间比例。

此外，反映任务难度的因子可定义为如下 S 形模式 [27]：

$$F_s(\bar{t}) = \frac{1}{1 + e^{0.05(\bar{t} - 150)}} \tag{14}$$

其中，$\bar{t}$ 作为任务难度的指标，以完成任务所需的最短时间（秒）来衡量。

t 的值取决于图像质量和威胁类型。在我们的案例中，考虑了两种图像质量：由无人机拍摄的低质量图像，以及由无人地面车辆拍摄的高质量图像。此外，我们考虑了三种待识别的物体类型，对应三个层次的抽象分类难度。t 的值定义见表 II。

表 II 不同场景下 t 的值。

图像质量	分类难度		
	Easy	中	Hard
Low	20	60	180
High	10	30	90

为了考虑绩效模型中的个体差异，我们引入两个修正权重：γ 和 ξ，分别反映人类认知能力和操作技能。γ 用于修正疲劳和工作负荷导致的绩效下降，而 ξ 用于修正任务难度导致的绩效下降。γ 和 ξ 的值定义如下：

$$\gamma = sin(h_c); \xi = sin(h_s) \tag{15}$$

其中，h_c，$h_s \in (0, \frac{\pi}{4})$分别表示个体的认知能力和操作技能水平。

我们将 h_c 和 h_s定义为三个等级：低 $\in (0, \frac{\pi}{12})$、中 $\in [\frac{\pi}{12}, \frac{\pi}{6}]$和高 $\in (\frac{\pi}{6}, \frac{\pi}{4})$。$h_c$或 h_s 的具体值从人类认知能力或操作技能水平的取值范围内随机抽取。

3）上下文多属性决策过程（*CMADP*）建模规范：在我们的案例研究中，状态空间，即多属性上下文 C，包含六个与任务、机器人或人类相关的子属性，如下所述：

- 任务属性：兴趣点的位置和分类难度
- 机器人特征：速度和图像质量
- 人类因素：认知能力和操作技能水平

动作空间 A 由两部分组成：将初始兴趣点分配给 j 机器人，以及将所有兴趣点的识别任务分配给 i 人类。请注意，我们根据位置信息使用 K均值聚类将兴趣点划分为 j 个感兴趣区域，并将每个聚类的质心兴趣点视为待分配的候选初始兴趣点，这与现实中的军事应用 [24]相符。分配初始兴趣点后，机器人会自动沿最短路径前往下一个最近的兴趣点。

对于奖励函数 f_R，我们将个体奖励 r_n 定义为由 Pr_c 驱动的威胁识别任务的平均表现，如公式 11 所述。具体而言，如果智能体正确识别出一个兴趣点，则根据分类等级（从低到高）获得 10、20 或 30 分。反之，如果识别错误，智能体将失去相同的分数。最后，计算单个环境步内所有兴趣点的平均得分，作为 r_n。

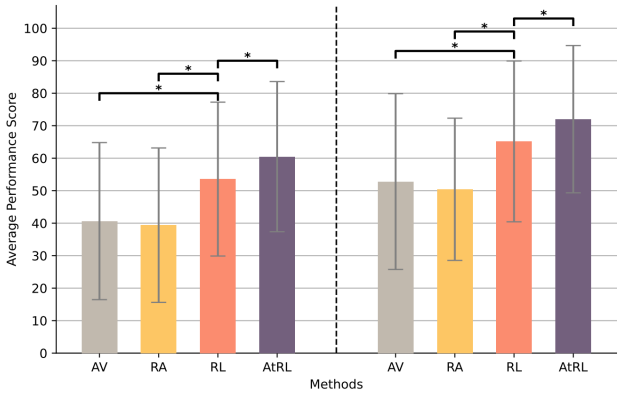


图5. 在设置（a）3个人类、4台机器人、20个威胁和20个非威胁（左），以及（b）5个人类、7台机器人、25个威胁和25个非威胁（右）下，各方法的平均性能得分及双样本t检验结果：$*$： $p < 0.001$。

C. 仿真实验与结果

1）实现细节：我们将每个跨属性注意力过程中的注意力头数设为 2，并将每个 Transformer 的隐藏单元大小设为 40。此外，门控循环单元中的隐藏单元数设为 64，且门控循环单元设为单层。对于 PPO，我们将裁剪比率设为 0.2，并将裁剪策略损失、价值函数损失和熵损失的损失系数分别设为 2、1 和 0.1。训练时使用 Adam优化器，初始学习率设为 2×10^{-4}。为提升并稳定训练过程，我们使用 10 个行为执行器并行收集交互经验。训练使用 NVIDIA Tesla V100 GPU 进行。

2）基线方法与消融模型：我们将AtRL模型与两种基线方法进行了比较：平均（AV）分配和随机（RA）分配。AV方法按照特定顺序将初始兴趣点分配给机器人，并将威胁分类任务平均分配给人类。相比之下，RA方法将机器人分配至兴趣点，并将威胁识别任务随机分配给人类。此外，为评估所提出的跨属性注意力模块的优势，我们构建了一个名为RL的消融模型，该模型移除了AtRL中的状态表征学习部分，并直接将属性序列拼接为状态。我们使用相同的PPO参数和 1×10^4个训练回合对AtRL和RL进行了训练。

3）评估：有两种评估设置：(a) 3个人类、4台机器人、20个威胁和20个非威胁；(b) 5个人类、7台机器人、25个威胁和25个非威胁。对于每种设置，随机选择人类、机器人和任务的属性，以生成多个场景，并使用每个模型测试500个未见过的场景。我们比较AtRL和RL的学习曲线，以及每个模型在每次运行中威胁分类任务所得平均分数，即平均性能得分。

4）结果：图5展示了每种设置下每次运行的平均性能得分。我们可以观察到，AtRL和RL的表现均在统计学上显著优于AV和RA（$p < 0.001$）。我们认为，原因在于AtRL和RL能够通过协同优化初始任务分配策略，持续与多属性上下文交互，而

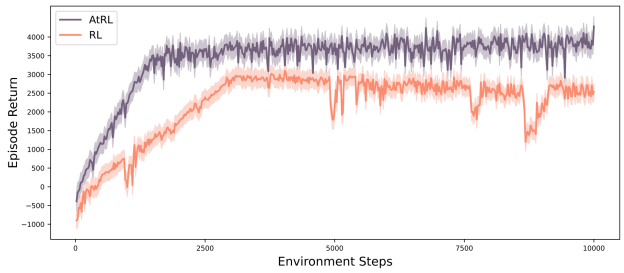


图6. 训练过程中 AtRL 和 RE 的回合回报学习曲线。

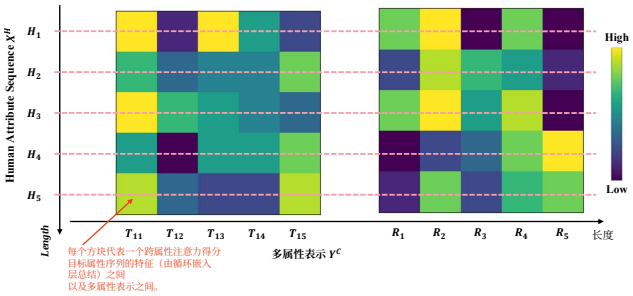


图7. 训练期间，在一个批次内设置 b 中，人类因素属性序列 X^H 与多属性表示 Y^C 之间学习到的部分示例跨属性注意力权重的可视化。请注意，原始注意力权重矩阵（见图3）的维度为 $i \times c$，即 5×69。

AV 和 RA 的任务分配策略是固定的，忽略了人类智能体、机器人和任务之间固有的异质性。这一比较凸显了异构多人人机团队中初始任务分配的重要性。此外，AtRL 的性能显著优于 RL（p <0。001）；当团队复杂度，即团队成员和任务的数量，更高时，这种性能提升更加明显。这是合理的，因为 AtRL 能够学习一种编码了多属性上下文中隐藏的潜在依赖关系的状态表示。此外，图 6 展示了设置 (b) 中 AtRL 和 RL 在训练期间的示例学习曲线。可以观察到，与 RL（橙色）相比，AtRL（紫色）以更高效、更稳定的模式获得了更多奖励。这进一步从定量角度证明了 AtRL 中引入的跨属性状态表示学习的优势。

为了说明我们的跨属性注意力如何工作，我们将注意力激活可视化。图 7 展示了一个部分示例：输入人类因素序列 X^H 和多属性表示 Y^C 后，在最终跨属性层学习到的跨模态注意力矩阵。我们可以观察到，我们的跨属性注意力成功地以自适应方式学会关注不同位置，这些位置在多属性表示中嵌入了人类特征与任务和机器人特征之间高度相关的信息。例如， H_3 （表示一名认知能力较低的人类操作员）与 T_{11} （表示位于较远位置的图像分类任务）以及 R_2 （表示速度较低的无人地面车辆）之间的交叉注意力分数相对较高。这是合理的

因为认知能力决定了人体抗疲劳能力，而人体抗疲劳能力对工作时间较为敏感；在我们的场景中，工作时间会受到机器人的导航速度和兴趣点位置的影响。

VI. 结论

在本研究中，我们研究了多人人机团队中的初始工作负荷分配问题，并考虑了人类、机器人和任务固有的异质性。我们对上下文多属性决策过程进行了通用的问题建模，并提出了一种新颖的基于注意力的深度强化学习方法。为了捕捉多属性上下文中多个属性和子属性之间的复杂依赖关系，我们引入了一个用于状态表征学习的跨属性注意力模块。我们通过威胁监视任务场景下的一项全面案例研究，展示了所提方法及其内部模块的优势。未来，我们计划在更复杂的场景中测试该方法，例如杂乱环境。我们还计划开展用户研究，以完善人类模型，并验证AtRL方法在现实应用中的有效性。此外，我们目前正致力于提升模型在规模更大的MH-MR团队中的适应能力。

致谢

本文基于美国国家科学基金会（NSF）资助的研究，资助编号为 IIS-1846221。本文所表达的任何观点、研究结果、结论或建议均仅代表作者的意见，不一定反映美国国家科学基金会的观点。

参考文献

[1] A. Ajoudani、A. M. Zanchettin、S. Ivaldi、A. Albu-Schäffer、K. Ko- suge 和 O. Khatib，《人机协作的进展与前景》，《*Autonomous Robots*》，第42卷，第957–975页，2018年。[2] A. Dahiya、A. M. Aroyo、K. Dautenhahn 和 S. L. Smith，《多智能体人机交互系统综述》，《*Robotics and Autonomous Systems*》，第161卷，第104335页，2023年。[3] S. D. Ramchurn、J. E. Fischer、Y. Ikuno、F. Wu、J. Flann 和 A. Waldock，《动态环境下人类—智能体协作的多无人机任务分配研究》，载于第二十四届国际人工智能联合会议（*IJCAI*）论文集，2015年，第1184–1192页。[4] A. Freedy、O. Sert、E. Freedy、J. McDonough、G. Weltman、M. Tambe、T. Gupta、W. Grayson 和 P. Cabrera，《面向多机器人、多人人机团队的多智能体可调自主性框架（MAAF）》，载于*2008*年国际协作技术与系统研讨会。IEEE，2008年，第498–505页。[5] A. Khamis、A. Hussein 和 A. Elmogy，《多机器人任务分配：最新进展综述》，协作机器人与传感器网络 *2015*，第31–51页，2015年。[6] M. Gini，《具有时间与顺序约束的多机器人任务分配》，载于第三十一届*AAAI*人工智能大会，2017年。[7] N. Creech、N. C. Pacheco 和 S. Miles，《动态多智能体系统中的资源分配》，*arXiv*预印本 *arXiv:2102.08317*，2021年。[8] H. Karami、K. Darvish 和 F. Mastrogiovanni，《产品缺陷检测场景中的人机协作任务分配方法》，载于*2020*年第*29*届*IEEE*机器人与人类交互通信国际会议（*RO-MAN*）。IEEE，2020年，第1127–1134页。

[9] L. Johannsmeier 和 S. Haddadin，《面向协作工业装配流程任务分配的分层人机交互规划框架》，*IEEE*机器人与自动化快报，第2卷，第1期，第41–48页，2016年。[10] H. Wu、A. Ghadami、A. E. Bayrak、J. M. Smereka 和 B. I. Epureanu，《多智能体团队中的负载管理与任务分配》，载于*2022*年国际机器人与自动化会议（*ICRA*）。IEEE，2022年，第8823–8830页。[11] T. Mina、S. S. Kannan、W. Jo 和 B.-C. Min，《面向独立同质任务的多人人机团队自适应工作负载分配》，*IEEE Access*，第8卷，第152 697–152 712页，2020年。[12] T. Yu、J. Huang 和 Q. Chang，《利用深度多智能体强化学习优化人机协作中的任务调度》，制造系统期刊，第60卷，第487–499页，2021年。[13] A. Ham 和 M.-J. Park，《人机任务分配与调度：波音777案例研究》，*IEEE*机器人与自动化快报，第6卷，第2期，第1256–1263页，2021年。[14] J. Patel 和 C. Pinciroli，《利用多人人机交互中的混合控制粒度提升人类绩效》，载于*2020*年第*29*届*IEEE*机器人与人类交互通信国际会议（*RO-MAN*）。IEEE，2020年，第1135–1142页。[15] Z. Xu，不确定多属性决策：方法与应用。Springer，2015年。[16] A. Tabrez、M. B. Luebbers 和 B. Hayes，《人机团队协作中的心理建模技术综述》，*Current Robotics Reports*，第1卷，第259–267页，2020年。[17] L. M. Ma、M. Ijtsma、K. M. Feigh 和 A. R. Pritchett，《人机团队设计指标：从团队协作视角评估人机团队》，*ACM*人机交互汇刊（*THRI*），第11卷，第3期，第1–36页，2022年。[18] J. Humann 和 E. Spero，《多无人机多操作员监视系统的建模与仿真》，载于*2018*年度*IEEE*系统会议（*SysCon*）。IEEE，2018年，第1–8页。[19] T. De Bruin、J. Kober、K. Tuyls 和 R. Babuška，《将状态表征学习融入深度强化学习》，*IEEE*机器人与自动化快报，第3卷，第3期，第1394–1401页，2018年。[20] F. Ni、J. Hao、J. Lu、X. Tong、M. Yuan、J. Duan、Y. Ma 和 K. He，《面向大规模混合流水车间调度问题的基于多图属性强化学习的优化算法》，载于第二十七届*ACM SIGKDD*知识发现与数据挖掘会议论文集，2021年，第3441–3451页。[21] J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov，《近端策略优化算法》，*arXiv*预印本 *arXiv:1707.06347*，2017年。[22] T. Zhang，《面向上下文赌博机与强化学习的“感觉良好”Thompson采样》，*SIAM*数据科学数学期刊，第4卷，第2期，第834–857页，2022年。[23] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、Ł. Kaiser 和 I. Polosukhin，《注意力就是你所需要的一切》，神经信息处理进展，第30卷，2017年。[24] C. E. Harriott 和 J. A. Adams，《人机系统的人类绩效建模》，人因与工效学评论，第9卷，第1期，第94–130页，2013年。[25] P. Taillandier、B. Gaudou、A. Grignard、Q.-N. Huynh、N. Marilleau、P. Caillou、D. Philippon 和 A. Drogoul，《利用GAMA平台构建、组合和实验复杂空间模型》，*GeoInformatica*，第23卷，第299–322页，2019年。[26] M. Watson、C. Rusnock、M. Miller 和 J. Colombi，《利用人类绩效建模为系统设计提供依据》，系统工程，第20卷，第2期，第173–187页，2017年。[27] R. W. Pew，《速度—准确性操作特征》，*Acta Psychologica*，第30卷，第16–26页，1969年。