arXiv:2601.01139v1 [cs.RO] 3 Jan 2026

面向多机器人探索的潜在空间强化学习

预印本

Sriram Rajasekar
航空航天工程系
印度科学研究所
印度班加罗尔
sriramrajase@iisc.ac.in

Ashwini Ratnoo
航空航天工程系
印度科学研究所
班加罗尔，印度
ratnoo@iisc.ac.in

2026年1月6日

摘要

未知环境的自主建图是一项关键挑战，尤其是在时间有限的场景中。多智能体系统能够通过协作提升效率，但运动规划算法的可扩展性仍是一项主要限制。强化学习已被用于探索解决这一问题，但现有方法受限于实现有效学习所需的有限输入规模，因此只能应用于离散环境。本研究通过利用自编码器进行降维，将高保真占据栅格地图压缩为潜在状态向量，同时保留关键空间信息，以解决这一限制。此外，我们提出了一种基于柏林噪声的新型程序化生成算法，用于生成拓扑结构复杂的训练环境，模拟小行星带、洞穴和森林。我们使用这些环境，借助分层深度强化学习框架训练自编码器和导航算法，以实现去中心化协同。我们还提出了一种加权共识机制，通过可调的信任参数调节对共享数据的依赖程度，从而确保系统能够抵御错误累积。实验结果表明，所提出的系统能够随智能体数量增加而有效扩展，并能很好地泛化至陌生且结构迥异的环境，在通信受限环境中也具有较强的鲁棒性。

关键词 自编码器 · 自主建图 · 去中心化协同 · 深度强化学习 ·降维 ·多智能体系统 · 程序化生成 · 加权共识

1 引言

多机器人自主探索是机器人学中的一项根本挑战 [1]，对于人员无法进入、进入危险或效率低下的场景中的应用至关重要。这类应用包括灾区的搜索与救援行动 [2, 3], 军事侦察、采矿资源勘探，以及面向地外定居点的行星探索 [4]。这些场景的主要目标是快速覆盖此前未知的空间，将原始传感数据转化为可用于行动的空间知识，并以地图的形式呈现 [5]。然而，要实现这一目标，仍需应对若干挑战。其中一些主要挑战包括：在任何给定环境中安全导航，包括本身复杂且无结构的环境；确保随着机器人数量或环境规模增加，整个系统仍能高效扩展；以及减轻含噪通信和传感器噪声所导致的误差影响。

为应对其中一些挑战而设计的多机器人探索策略，可分为集中式方法和去中心化方法。集中式方法虽然能够通过将所有信息路由至中央控制器来实现全局协调，但通常会受到计算瓶颈、单点故障以及对稳健且普遍可用的通信高度依赖等问题的困扰 [6]。相比之下，去中心化方法在可扩展性以及应对通信故障方面表现出更好的能力 [7]。传统解决方案无论采用集中式还是去中心化方式，主要

依赖手工设计的启发式方法，即预先编程的规则或效用函数 [5], ，例如引导智能体前往最近的未访问前沿 [8, 9], 、根据预计行进成本与信息增益之间的关系采用竞价机制进行任务分配 [10, 11], ，或利用人工势场来保持空间分离 [12]。此类传统方法需要基于领域特定知识进行明确的工程设计，例如关于环境布局的假设（如走廊宽度、房间尺寸）、对环境的语义理解 [13, 14], 以及对障碍物分布和类型的预期。因此，当面对多样且非结构化的环境时，这些策略往往难以实现泛化；此外，在缺乏全局视角的情况下，它们通常也难以协调复杂的多智能体协作。不同于依赖此类手工设计启发式方法的传统方法，深度强化学习（DRL）凭借能够直接从高维感知输入和交互经验中学习复杂行为的高级端到端策略，已成为一种强大的方法 [15], 。DRL 自动发现复杂策略的能力使其非常适合多机器人探索问题，尤其是在可扩展性和泛化能力至关重要时。

尽管将去中心化深度强化学习框架应用于基于地图的探索颇具前景，但要有效实现这一应用仍面临一系列重大挑战，主要源于输入数据的特性。为了准确决策，避免忽略环境中的关键细节，例如狭窄但可通行的间隙或细微的障碍物，高分辨率地图输入十分理想；尤其是随着多机器人系统日益承担探索大规模、复杂真实环境的任务，在这类环境中，相较于单机器人系统，多机器人系统的优势最为突出 [17]。然而，直接将此类高分辨率数据用作深度强化学习智能体的输入，会立即引发“维数灾难”，即状态空间会随着输入维度的增加而呈指数级扩张。这使策略网络极难同时学会稳健的特征提取，以及有效的导航或探索行为 [18]。

要训练出能够胜任大规模探索、具有鲁棒性和泛化能力的深度强化学习智能体，关键在于让它们接触大量多样且富有挑战性的环境。然而，过去将深度强化学习应用于高分辨率输入一直困难重重，因此许多方法转而依赖规模较小、复杂度较低的地图表示。这一实际限制无意间造成了一个显著的文献空白：成熟且可扩展的地图生成技术相对匮乏。由于智能体主要局限于较简单的输入，开发先进地图生成器的动力也随之减弱。因此，先前的研究往往采用简单的手工设计环境或假定已有地图集，而这些做法可能无法充分体现真实场景的多样性和复杂性，正如[19]所指出的。一些基准（如 Procgen [20], ）采用程序化内容生成来创建多样的游戏关卡。但据作者所知，深度强化学习多机器人探索领域的先前研究尚未提出专用且可定制的大规模地图生成器，以便对障碍物密度、结构和可导航性等特征进行精细控制。如今，缺乏高效方法来生成所需的复杂高分辨率训练场景，已成为一个显著瓶颈，严重阻碍了能够应对真实大规模探索任务的深度强化学习方案的开发与严格评估。Perlin 噪声 [21] 长期以来被用作计算机图形学等领域的程序化生成技术，因为它能够生成平滑、自然的纹理和地形。它会在网格上生成连续的伪随机值梯度，从而避免纯白噪声所带来的不连续性。其实现方式是在分配给网格点的梯度之间进行插值，从而形成频率和尺度可调的连贯空间模式。凭借这些特性，Perlin 噪声自然适用于生成兼具随机性和结构性的训练地图。

此外，先前尝试将强化学习应用于大型环境的工作，往往采用一些权宜之计，例如大幅降低网格分辨率，实际上主要关注离散环境 [22], 而这会牺牲关键的细粒度细节。其他工作仅将小范围裁剪的局部地图区域作为输入，这可能导致用于有效决策的必要全局上下文丢失 [23]。因此，先前基于强化学习的框架在可扩展性方面存在局限。由此，亟需能够有效进行特征提取和降维的方法，将高分辨率输入转换为紧凑且信息丰富的表示，使深度强化学习智能体能够处理这一学习问题，同时不损失关键信息。

此外，在去中心化多智能体环境中有效运行，需要强大的协调能力，尤其是在现实中存在噪声的环境中工作时。在传感器不完美或通信信道不可靠的情况下，智能体之间交换的信息可能有误；若没有机制对可能受损的数据进行批判性评估与整合，错误就极有可能在整个机器人群中传播。这会导致集体感知能力下降、决策失误，从而损害整体探索任务。这一挑战的理论基础源于概率映射，其中通过占据栅格对不确定性进行建模，并以算法方式融合信息 [24]。虽然这些原理能够处理单个智能体的传感器不确定性，但并未直接说明如何对多个智能体之间传递的整幅地图或地图片段进行最优融合；这仍是一个活跃的研究领域 [25], ，尤其是在感知和传递的信息受到不同类型、不同程度的噪声和错误影响时（将在第5节讨论）。因此，

需要一种信息融合算法，使每个个体能够控制自己对从邻居处接收的信息的信任程度。

除了上述挑战之外，自主多机器人探索本身具有复杂性高、时间跨度长的特点，这给传统的单策略深度强化学习系统带来了显著困难。单策略方法往往难以同时掌握高层战略决策和精细的低层运动控制，导致样本效率低下，并且在大规模任务中表现欠佳。对于多机器人探索这类复杂且时间跨度长的任务，分层控制结构往往能发挥更大优势。因此，有必要探索能够将问题分解为更易处理的不同抽象层次的分层控制结构。

为应对上述挑战，并提出一种具有泛化能力、可扩展且抗噪的多机器人探索问题解决方案，本文提出了一种包含多项关键创新的去中心化深度强化学习框架。首先，我们提出了一种先进的程序化地图生成算法，利用分层柏林噪声和可控的难度参数，生成多样化、大尺度、高分辨率的环境，为全面的深度强化学习训练提供必要条件。其次，为克服此类丰富地图输入带来的维数灾难，我们预训练了一个卷积自编码器，以有效提取特征并降维，从而为强化学习智能体提供紧凑的状态表示。第三，我们通过一种新颖的地图融合机制实现稳健的去中心化协同：智能体交换各自独立维护的地图的编码表示，并使用可配置的信任参数整合这些信息，以减轻含噪传感器数据、通信以及编解码循环所致误差的影响。最后，我们采用分层强化学习架构 [26, 27] 来管理探索任务的复杂性，将其分解为战略层面的高层决策和低层导航。该框架的设计能够促进隐式协同，因为智能体在两个层级进行决策时都会考虑邻居状态，从而为未知环境中的自主探索提供可扩展且自适应的解决方案。

本文其余部分的组织结构如下：第2节对问题进行描述。第3节介绍所提出的地图生成算法的详细信息；第4节介绍基于自编码器的特征提取方法的训练与使用，以支持处理大型占据栅格。第5节讨论智能体之间信息的流动与融合。用于训练的分层强化学习框架及训练循环本身将在第6节中介绍。最后，第7节呈现实验及其结果，第8节对工作进行总结，并讨论未来可能的研究方向。

2 问题描述

考虑 N 机器人（A_1， A_2，……， A_N）协作，对给定的未知静态环境进行全面探索。它们同时在未知环境中的随机位置启动，且彼此不重叠。每个机器人 A_i 周围都有一个感知半径 R_s，在该范围内，它可以准确感知障碍物，并与其他机器人进行通信。待探索环境被视为大小为 $L \times L$ 的方形占据栅格（下文称为“地面地图”， G），其中“1”表示障碍物，“0”表示自由空间。针对该环境和机器人，作出以下假设：

A1. 环境是静态的二维环境。A2. 机器人被建模为点代理，每个机器人在网格上占据一个像素。A3. 机器人配备的测距传感器没有噪声和误差。A4. 机器人始终知道自己在网格中的精确位置。

每个机器人 A_i 独立维护多个二值占据网格地图：

1. “自身障碍物地图”（O_i^s），用于存储仅由 A_i感知到的障碍物位置。2. “自身探索地图”（E_i^s），用于存储仅由 A_i探索过的位置。3. “融合障碍物地图”（O_i^f），用于存储通过感知和通信获得的累计障碍物位置。4. “融合探索地图”（E_i^f），用于存储通过感知和通信探索到的累计位置。

目标是利用强化学习训练一种基于神经网络的解决方案，供每台机器人单独使用，并将其作为共享策略，引导机器人协同探索环境。当

所有机器人的集体地图绘制完成时，即

$$\bigvee_{i=1}^{N} O_i^f(x,y) = G. \tag{1}$$

与 (1) 等价的一种条件是检查机器人的集体探索地图是否已完全填满，即

$$\forall x,y \in \{1,\dots,L\},\ \bigvee_{i=1}^{N} E_i^f(x,y) = 1. \tag{2}$$

与检查 (1) 相比，检查 (2) 更为实际，因为机器人起初并不知道 G。由于在未知环境中无法实现 100% 的探索率, [28], 因此将 80% 的探索进度作为软终止阈值。图 1 展示了仿真中探索进度不断提升时的一个示例场景。

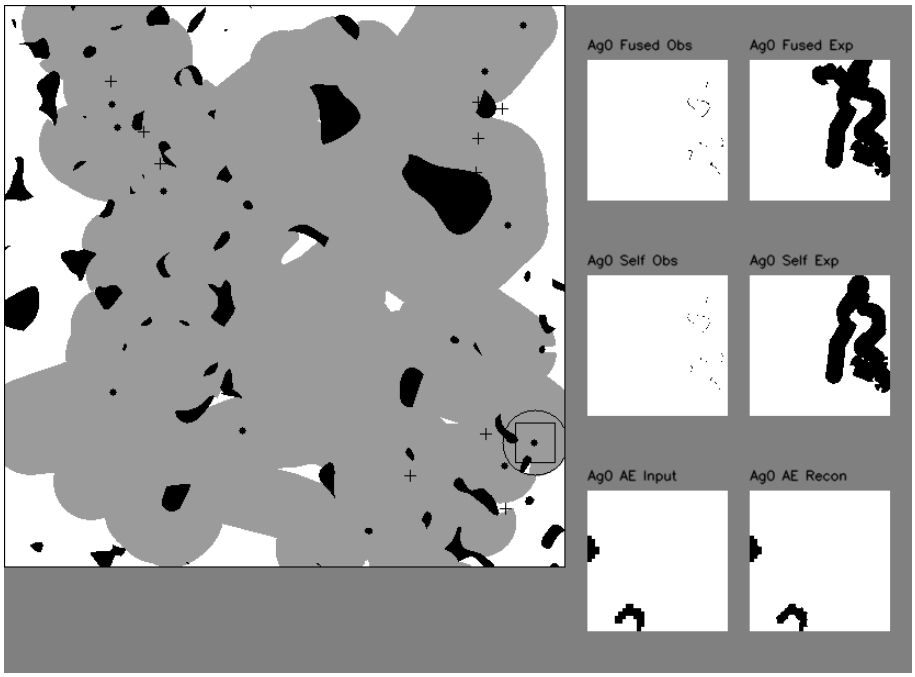


图 1：示例场景。(i) 左侧：10 台机器人正在探索地面地图。用圆圈标出一个智能体，以展示其感知半径。(ii) 右侧：从左上角开始，依次为该标出智能体的融合障碍物地图、融合探索地图、自身障碍物地图和自身探索地图。最后两幅图与自编码器有关，详见第4节。

3 基于柏林噪声的地图生成算法

本研究中使用的随机生成环境均表示为占据栅格，如前文所述。现实环境中的障碍物在形状、尺度和尺寸方面可能具有不可预测性。在生成训练数据时，聚焦于非结构化且完全随机的环境，有助于将解决方案泛化到任何环境，包括训练中未曾见过的环境。

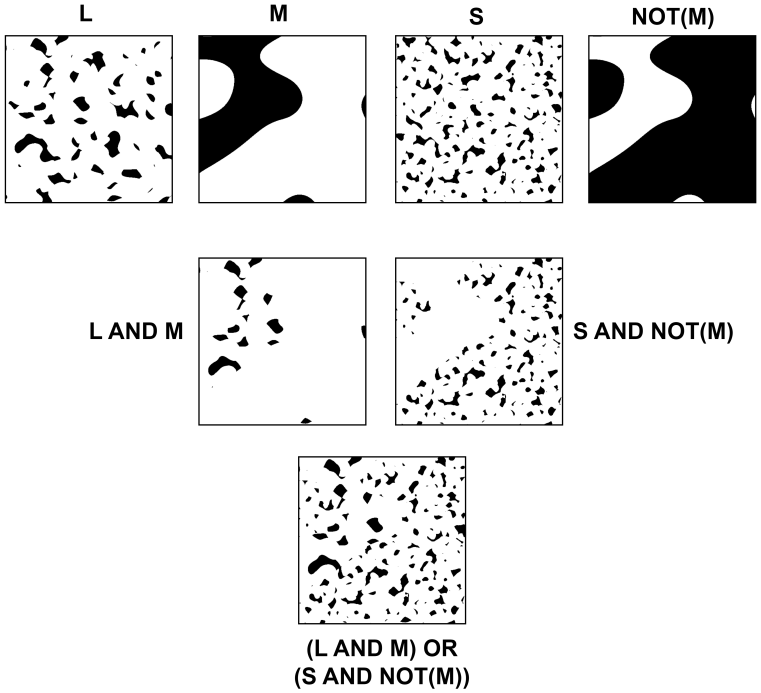
将网格扩展到更大规模（例如 512×512）并采用纯随机特征，会显著增加复杂度，尤其是在生成多样且有意义、适合训练强化学习（RL）智能体的环境时。为应对这一挑战，我们提出了一种快速灵活的地图生成算法，能够大规模生成随机化环境。该方法支持障碍物尺寸、形状和密度的完全可变，同时保持计算效率。因此，可以在实际可行的时间内生成大型训练数据集，以供第 IV 节和第 VI 节所述的学习框架使用。

算法1详细介绍了本文提出的地图生成过程。我们将柏林噪声的单个输出定义为层，它以特定的空间频率生成，以下称为倍频程数。每个层都是

512×512 实数值网格，取值范围为 $[0, 1],$，其中低于所选阈值（例如 0.5）的值表示自由空间，高于该阈值的值表示障碍物。

算法1使用掩膜 Perlin 层的地图生成

输入： 地图尺寸 L、裁剪半径 c、难度 $d \in [0, 1]$
$e \leftarrow 1 - d$ {阈值化的易化因子}
对于 $comp \in$ {L，S} **执行**
  对于 $type \in$ {基类，掩膜} **执行**
    八进制 $\leftarrow$ 随机函数(*comp*=L？ $[7, 12]$： $[15, 20]$)
    原始值 $\leftarrow$ Perlin(L，八进制)
    $layer_{type} \leftarrow$ 原始值 $>$ ($type =$ 基类？ e： 0。5)
  结束循环
  $Map_{comp} \leftarrow layer_{base} \wedge layer_{mask}$
结束循环
$M \leftarrow$ Perlin(L，随机函数($[1, 3]$)) > 0.5
地图 $\leftarrow$ (地图$_{\text{L}} \wedge M$) $\vee$ (地图$_{\text{S}} \wedge \neg M$)
将 Map 的中心设为掩膜，半径为 c
为 Map添加边框
返回 地图



Figu图 2：地图生成算法的使用示例。标签与算法中的用法一致 1.

单个柏林噪声层，记作 $layer_{base}$，阈值化后通常会导致障碍物密度较高，且可通行路径有限。为缓解这一问题，使用随机数量的倍频程生成第二个柏林噪声层 $layer_{mask}$。对两个层进行阈值化后，再执行二进制按位与运算，仅保留同时存在于 $layer_{base}$ 和 $layer_{mask}$ 中的障碍物区域。这能有效增加环境中的自由空间，同时保留柏林噪声特有的自然、有结构的形状。引入一个可控参数 d，通过在不同阈值下控制基类层的密度，间接调节难度。d 的值越低，阈值就越高，因此基类层中被障碍物覆盖的比例也越小。

虽然该组件地图提升了可导航性和障碍物真实感，但障碍物尺度的多样性可能不足。为此，使用随机选取的倍频程值生成两个此类组件地图（L 和 S）。第三个层

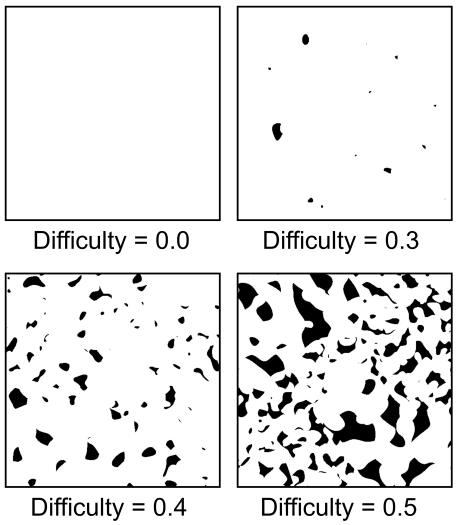


图 3：具有不同难度级别的生成地图示例。

M 则生成为低倍频程的纯 Perlin 噪声层，并通过算法1作为组合 L 和 S 的逻辑。聚合后得到的地图包含形状和大小各异的障碍物，且难度（密度）可控。地图中央设有一处空地，用于初始化机器人；地图边缘则添加由宽度为一个像素的障碍物构成的虚拟边框。图2展示了简化版地图生成算法的示意图，图3则展示了使用不同难度设置生成的一些地图示例。

4 特征提取

使用高分辨率占据栅格对于检测细粒度障碍物至关重要，但将此类高维数据直接输入强化学习智能体会引发“维数灾难”，使特征提取与控制策略的同步学习变得难以实现。作为端到端训练系统一部分的编码器已被用于规避这一问题，即使是在相对较小尺度的问题中也是如此 [29]。对于面向更大尺度问题、内在结构更深的网络，端到端训练可能会带来更多困难。为此，我们采用卷积自编码器，将高保真地图压缩为紧凑的潜在表示，从而有效地将视觉处理与决策解耦。该方法建立在既有研究的基础上：自编码器已被成功用于将复杂输入提炼为低维状态，以实现机器人控制 [30]。图 4 展示了使用自编码器压缩地图的可视化结果。

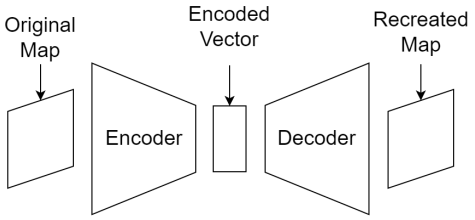


图 4：典型自编码器的框图。

该自编码器使用根据第 3 节所述地图生成算法、通过随机函数难度设置生成的 2000 张图像进行训练和测试，训练集与测试集的比例为 9:1。该自编码器架构由一个对称的编码器-解码器网络组成，用于处理 512×512 灰度占据栅格。编码器采用七个卷积块，逐步将空间尺寸从 512×512 压缩至 4×4，同时将通道深度从 1 增加到 128。由此得到一个大小为 2048 的展平潜在向量，该向量作为强化学习智能体压缩状态输入的一部分。详细的架构规格见表 1。

表 1：自编码器架构规格

层	输出维度
编码器	
输入	$1 \times 512 \times 512$
二维卷积块 1	$16 \times 256 \times 256$
二维卷积块 2	$32 \times 128 \times 128$
二维卷积块 3	$64 \times 64 \times 64$
二维卷积块 4	$128 \times 32 \times 32$
二维卷积块 5	$128 \times 16 \times 16$
二维卷积块 6	$128 \times 8 \times 8$
二维卷积块 7	$128 \times 4 \times 4$
展平 +双曲正切函数	**2048（潜在向量）**
解码器	
反展平	$128 \times 4 \times 4$
转置卷积块 1–6	上采样（对称）
转置卷积块 7	$1 \times 512 \times 512$
最终输出（Sigmoid函数）	$\mathbf{1 \times 512 \times 512}$

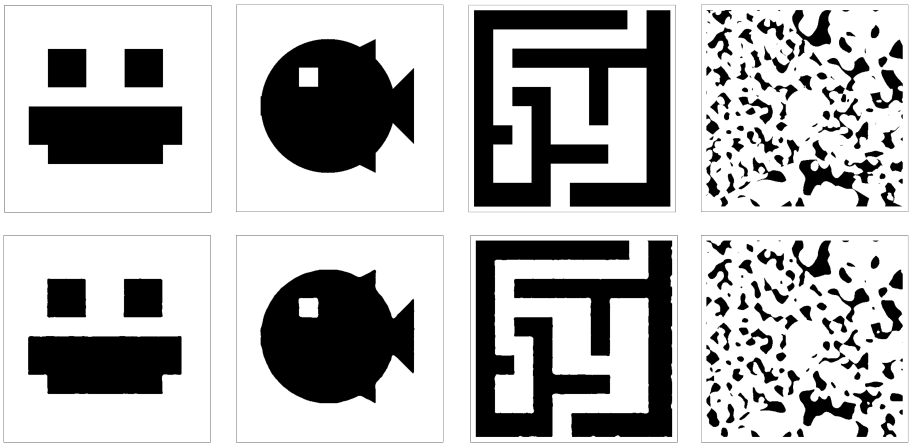


图5：对自编码器进行的泛化测试。顶行显示输入编码器的原始灰度图像。底行显示将编码后的潜向量输入解码器后重建的图像。

通过输入与训练数据在视觉上毫无关联的图像，对训练好的自编码器的泛化能力进行测试。图 5 中的结果表明，该自编码器能够很好地泛化，将任意黑白图像编码到具有信息量的潜在空间中，并将其重建为可识别的重建结果。图 5 中只有第四列代表用于训练该自编码器的训练数据。

5 信息流

给定机器人 A_i 通信半径内的 m机器人称为其邻居。在这些邻居中，与 A_i 通信的机器人称为通信邻居。通信邻居的最大数量是一个训练参数，也是一项设计选择，标记为 M。

根据 m 与 M 之间的关系，按照算法 2 选择 M通信邻居 $C_1, \ldots, C_M$，以确保所选通信邻居的数量始终为 M。这一点至关重要，因为来自通信邻居的信息将作为神经网络的输入，因此通信邻居的数量必须始终保持一致。自A_i 上次通信以来经过一定时间后，它将与接下来发现的机器人进行通信。在一次通信事件中，发起通信的机器人会向所选邻居请求并接收以下信息：邻居的标识符 (k)、当前位置、当前速度、当前目标，以及邻居自身障碍物地图 O_k^s和自身探索地图 E_k^s的编码向量。

算法 2 为智能体 A_i选择通信邻居 C_1、.....、 C_M

1: **如果** $m = 0$，**则**2: 将 M 位于对侧象限角落的虚拟智能体视为通信邻居3: **否则，如果** $m < M$，**则**4: 将 m 个最近的智能体视为通信邻居，并将 $M - m$ 个位于对侧象限角落的虚拟智能体也视为通信邻居5: **否则，如果** $m \geq M$，**则**6: 将 M 个最近的智能体视为通信智能体7: **结束条件判断**

虽然每个机器人的自身地图（O^s、 E^s）只会因该机器人自身的传感器噪声而累积误差，但融合地图（O^f、 E^f）累积的误差不仅来自该机器人的传感器噪声，还来自与其通信的每个机器人所产生的传感器噪声、通信噪声和编解码误差，并且还会进一步来自那些机器人所通信的每个机器人，依此类推。为防止误差在机器人群中级联传播，智能体在通信事件期间仅共享自身地图，而不共享融合地图。
融合过程必须整合这些信息（机器人的自身地图和融合地图，以及通信期间从其他机器人获得的自身地图），同时减轻信道和传感器不确定性的影响。为此，我们采用由信任参数 $\beta \in [0, 1]$ 控制的加权共识机制，用于调节对自我感知与邻居数据的依赖程度。该方法利用机器人群中的冗余来滤除随机误差，随着智能体数量增加而提高地图鲁棒性。融合算法作用于通信事件后机器人 A_i 可获得的地图集合：其自身的当前地图（O_i^s、 E_i^s、 O_i^f、 E_i^f）以及从邻居 $k \in C_1$、.....、 C_M 接收并解码的地图（O_k^s、 E_k^s）。以下小节将详细介绍如何组合这些输入，以计算中间融合地图，随后计算最终融合地图 O_i^s、 E_i^s、 O_i^f、 E_i^f。

5.1 中间地图计算

ˆ ˆ 临时中间探索地图（E）和障碍物地图（O）首先根据当前自身感知到的地图及接收到的邻居地图计算得出。由于假设定位理想，因此自身探索地图提供了可靠的信息。因此，坐标（x， y）处的中间探索状态，就是对所有可用探索地图在该坐标处执行逻辑或运算的结果。

$$\hat{E}(x, y) = E_i^s(x, y) \vee \bigvee_{k \in C} E_k^s(x, y). \tag{3}$$

ˆE 表示在当前步骤中，智能体的自身地图或其任一通信邻居的地图是否将某个像素视为已探索。

中间障碍物值的计算取决于智能体及其邻居对该像素的探索状态。令 C 为非虚拟通信邻居的标识符集合，并令 $n(x, y) = \sum_{k \in C} E_k^s(x, y)$。根据 $E_i^s(x, y)$和 $n(x, y)$的取值，分为以下四种情况：

情况 1： $E_i^s(x, y) = 0$ **且** $n(x, y) = 0$

位置（x, y）尚未被参与此次通信事件的任何智能体探索过，包括 A_i。关于位置（x, y）的障碍物状态，当前事件未提供任何新信息。

$$\hat{O}(x, y) = 0. \tag{4}$$

情况 2： $E_i^s(x, y) = 1$ **和** $n(x, y) = 0$

A_i 已探索（x, y），而此时间步中其他通信邻居均未探索。因此，机器人完全依赖自身感知到的障碍物状态。

$$\hat{O}(x, y) = O_i^s(x, y). \tag{5}$$

情况 3： $E_i^s(x,y) = 0$ **和** $n(x,y) > 0$

尽管 A_i 自身尚未探索过该区域（x，y），但一些通信邻居已经探索过。因此，机器人采用探索过该像素的邻居所提供的平均障碍物信念。

$$\hat{O}(x,y) = \frac{1}{n(x,y)} \sum_{k \in C} O_k^s(x,y). \tag{6}$$

情形 4： $E_i^s(x,\ y) = 1$ **和** $n(x,\ y) > 0$

（x，y）已由 A_i 以及一些通信邻居探索。信任权重比 W（x，y）定义为智能体的自我感知权重相对于邻居信息权重的比值，其根据探索邻居的数量 n（x，y）和信任参数 β 计算，公式为 W（x，y）$=$（n（x，y）-1）$\beta + 1$。信任参数的影响如图6所示。值得注意的是，当 $\beta = 0$、W（x，y）$= 1$ 时，这意味着 A_i 对通信邻居提供的数据与对自身收集的数据同样信任。这代表了它所能表现出的最大信任度。相反，当 $\beta = 1$、W（x，y）$= n$（x，y）时，这意味着只有当所有其他通信邻居都认定 A_i 有误时，A_i的信念才会被推翻。这代表了 A_i 所能表现出的最低信任度。

利用计算得到的信任权重比，随后可计算中间障碍物值：

$$\begin{aligned} \hat{O}(x,y) = &\frac{W(x,y)}{W(x,y) + n(x,y)} O_i^s(x,y) \\ &+ \frac{1}{W(x,y) + n(x,y)} \sum_{k \in C} O_k^s(x,y). \end{aligned} \tag{7}$$

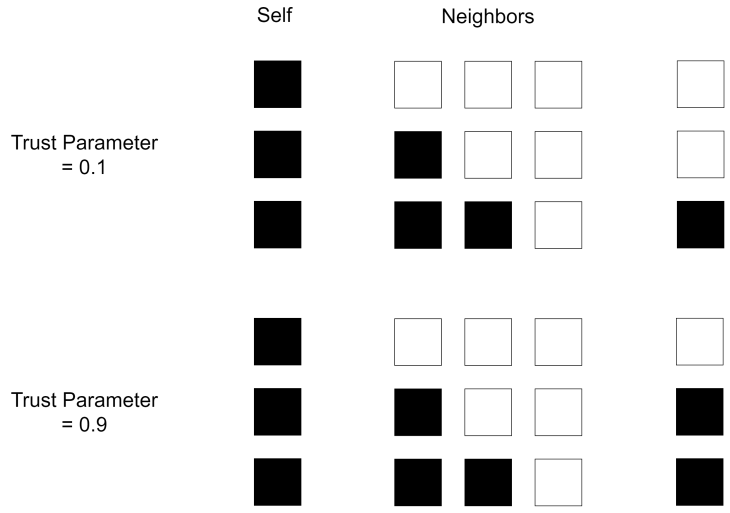


图 6：展示不同信任参数影响的示例。

ˆ ˆ 中间浮点地图 E 和 O 以 0.5 为阈值进行阈值化，从而转换为二值地图。这些地图表示当前时间步的融合共识，不包含过去从其他邻居处收集的任何信息。

5.2 阈值化与时间融合

持久融　　合地图（O_i^f 和 E_i^f）通过与E 和Oˆ中的新信息进行整合来更新　　.

- **探索　地图更新：** 更新后的探索地图仅包含任何新探索的区域。

$$E_i^f(x,y) = E_i^f(x,y) \vee \hat{E}(x,y). \tag{8}$$

• **障碍物地图更新：** 对于在此融合步骤中探索的像素（ˆ*O*(*x*， *y*) ˆ），障碍物地图更新优先采用当前融合步骤中的信息（(*x*， *y*)，使得 *E*(*x*， *y*) = 1）；而对于未参与当前融合更新的像素（(*x*， *y*)，使得 ˆ*E*(*x*， *y*) = 0），则保留其先前的障碍物状态（O_i^f(*x*， *y*)）。

$$O_i^f(x,y) = (\hat{O}(x,y) \wedge \hat{E}(x,y)) \vee (O_i^f(x,y) \wedge \neg\hat{E}(x,y)). \tag{9}$$

这一时序融合步骤确保 A_i 的融合地图反映最新的可用信息，同时保留有关当前感知或通信事件未覆盖区域的知识。信任参数 β 可用于调整系统对潜在噪声邻居信息的敏感度，以及对自我感知的依赖程度。

6 强化学习框架与训练

所提出的多机器人探索系统采用分层强化学习方法，将整体探索任务分解为两个子问题：（1）选择战略性探索航点；（2）高效、安全地导航至所选航点。为此，系统训练了两个不同的神经网络策略，分别称为网络 1（探索策略）和网络 2（导航策略）。本节将详细介绍训练环境、导航策略网络、探索策略网络，以及旨在促使网络产生有效行为的奖励结构。

6.1 多智能体强化学习环境

我们开发了一个自定义多智能体环境，用于模拟机器人探索任务。该环境模拟 N台机器人在由第 3 节所述算法生成的 512×512网格地图中运行。训练期间使用的关键环境和机器人参数汇总于表 2。

表 2：关键参数

参数	值
机器人数量（N）	10
最大通信邻居数（M）	3
地图尺寸	512×512 像素
感知半径（R_s）	30 像素
机器人碰撞半径（col_rad）	3 像素
机器人最大速度（max_vel）	10 像素/秒
机器人最大加速度（max_acc）	5 像素/秒^2
地图难度（d）	0.1、0.2、0.3、0.4
地图融合信任参数（β）	0.8
通信频率（$comm_freq$）	1 个时间步
时间步（dt）	0.1 秒

每个机器人都配备了一个模拟测距传感器，用于障碍物检测并更新其个体地图（O_i^s， E_i^s）。机器人之间的通信可以交换编码后的自身地图，随后使用第5节中详述的基于信任的机制，将这些地图融合到每个机器人的融合地图中（O_i^f， E_i^f）。第4节中介绍的自编码器用于在智能体间通信期间对地图进行编码。

6.2 导航策略网络

导航策略网络负责引导机器人安全地绕过障碍物，并隐式学习能够避免智能体间碰撞、同时高效抵达所选目标航点的行为。动作空间为连续空间，表示期望的加速度向量（$\ddot{x}_i$， $\ddot{y}_i$）。加速度的各个分量会经过裁剪，使其大小不超过设定的最大加速度，同时保持网络所选的方向。每个智能体 A_i 的导航策略网络所接收的观测是一个连续向量，由以下部分组成：

1. 将 A_i 的融合障碍物地图（O_i^f）裁剪并处理后，输入预训练编码器（第4节），生成一个 2048 维潜在向量。裁剪区域以智能体为中心，半径为$0.6 \times R_s$。对障碍物进行膨胀处理，以便在早期阶段障碍物地图信息稀疏时帮助编码器。

2. 一个 4 维向量，表示 A_i 的当前位置（x_i，y_i）和速度（$v_{x,\ i}$，$v_{y,\ i}$），分别按地图尺寸和最大速度归一化。3. 一个 2 维向量，表示 A_i 当前的导航目标（t_x，t_y），按地图尺寸归一化。在训练导航策略网络、再训练探索策略网络之前，该目标由环境随机分配。4. 对于每个 M 通信邻居（根据算法 2，可能是真实或虚拟邻居），使用一个 4 维向量表示其归一化后的位置和速度。由此得到一个 $4 \times M$ 维向量。

这些组成部分经过拼接，形成一个观测向量，其中部分元素的取值归一化到范围 $[-1, 1]$，其他元素则归一化到 $[0, 1]$。导航策略的演员和评论家组件均采用多层感知机（MLP）架构。策略网络（演员）和 Q 函数网络（评论家）均由四个隐藏层组成，每层分别包含 [1024, 512, 512, 128] 个单元，并使用 LeakyReLU 激活函数。导航策略的奖励函数旨在鼓励智能体高效到达目标，同时惩罚碰撞和危险行为。对于每个智能体 A_i，在每个时间步，奖励 R_i 由以下几个部分相加构成：

1. 如果智能体到达当前目标（定义为与目标的距离在 15 个单位以内），则给予较大的正奖励（$+1500$）。2. 奖励与相较于上一步到目标距离的减少量成正比（$4 \times \Delta d_{target}$）。这会鼓励持续前进。3. 根据智能体速度向量与指向目标的向量之间的余弦相似度给予奖励（$2 \times (\cos\theta)^3$）。这会与距离减少项共同鼓励智能体朝向目标前进。4. 如果智能体与静态障碍物发生碰撞（若其当前速度会导致其在下一步进入障碍物的碰撞半径内），或与另一机器人发生碰撞（若其当前速度会导致其与另一机器人的距离小于两倍碰撞半径），则给予负奖励（-100）。5. 根据较短时间范围（$T_{coll_window} = 15$ 个时间步）内预测的未来碰撞给予负奖励。该惩罚根据预测碰撞发生的时间远近进行缩放：

$$R_{predcoll} = S_{coll} \times \frac{T_{coll_window} - t_{coll} + 1}{T_{coll_window}}, \tag{10}$$

其中，t_{coll} 是与智能体或障碍物（$t_{coll} \leq T_{coll_window}$）发生碰撞的预测时间，$S_{coll}$ 是负缩放因子（-10）。

智能体 A_i 的总奖励为 $R_i = R_{term} + R_{prox} + R_{align} + R_{coll} + R_{predcoll}$。

6.3 探索策略网络

探索策略网络负责通过航点选择实现有效探索，并隐式学习能够与邻居间歇性接触以进行通信的行为。动作空间是连续的，用经过地图尺寸归一化的 x 和 y 坐标表示期望航点。每个智能体 A_i 的探索策略网络观测是一个连续向量，其分量如下：

1. 一个 2048 维潜在向量：将 A_i的融合探索地图（E_i^f）输入预训练编码器（第4节）生成。2. 一个 4 维向量，表示 A_i的当前位置（x_i，y_i）和速度（$v_{x,\ i}$，$v_{y,\ i}$），分别按地图尺寸和最大速度归一化。3. 对于 M 个通信邻居（根据算法 2，真实或虚拟），每个邻居均由一个 4 维向量表示其归一化后的位置和速度。由此得到一个 $4 \times M$ 维向量。

这些组件经过拼接，构成一个观测向量，其中一些元素的值被归一化到范围 $[-1, 1]$，其他元素则归一化到 $[0, 1]$。探索策略的演员和评论家组件均采用多层感知机（MLP）架构。策略网络（演员）和 Q函数网络（评论家）各由三个隐藏层组成，每层分别包含 [1024, 256, 64] 个单元，并采用 LeakyReLU 激活函数。探索策略的奖励函数旨在鼓励高效探索，同时惩罚时间浪费和选择低效航点。对于每个智能体 A_i，在每个时间步，其奖励 R_i 由多个部分组成：

1. 在回合结束时给予较大的正奖励，并按全局探索进度比例（f_{exp}）缩放为$100 \times (2 \times f_{exp} - 1)$，以鼓励探索的全局性。

2. 奖励与本步骤相较于上一步的自我探索进度比例（f_{self_exp}）增幅成正比，并按（$200 + 300 \times f_{exp}$）$\times f_{self_exp}$进行缩放。这旨在鼓励持续进展，并特别强调最后一公里探索。3. 如果智能体选择的下一个航点距离当前位置非常近（距当前位置≤ 40 像素），则给予负奖励（-3）；否则给予正奖励（$+1.5$）。

智能体 A_i 的总奖励是这些单项奖励之和。

6.4 分层训练流程

整体训练遵循两阶段分层流程：

1. 在导航策略的训练过程中，环境会为每个机器人随机分配有效的随机目标。此阶段的主要目标是学习稳健的避碰和目标到达行为。2. 随后，冻结导航网络的权重，并训练探索策略。探索网络将导航网络每 50 步视为一个宏步。此阶段的主要目标是在每个宏步中选择一个目标，以最大化探索效果。

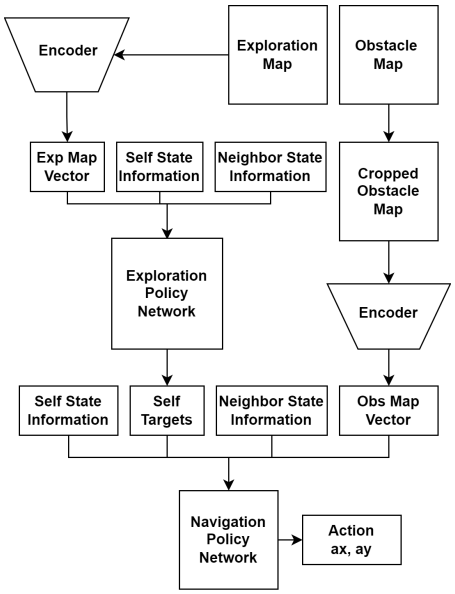


图 7：所提出的分层框架概览。

这种分阶段的方法使每个网络专门处理各自的任务，从而简化了学习问题。两个网络均采用软演员-评论家（SAC）算法 [31]进行训练。以往针对连续场景的分层探索框架主要采用深度确定性策略梯度（DDPG）[22]，而我们选择 SAC 是因为它采用最大熵形式。DDPG 是确定性算法，若不注入启发式噪声，就容易收敛到次优的局部极小值；与之不同，SAC 的随机策略能够自然地鼓励多样化的探索行为。此外，与脆弱的确定性策略相比，这种内在的随机性使其在去中心化多智能体环境常见的非平稳性面前具有更强的鲁棒性。在这两种情况下，训练均通过一个汇集式经验回放缓冲区进行：从单个智能体的角度看，该缓冲区将包含 N 个智能体的每个回合视为 N 个回合。这样做是为了让共享策略有效泛化，进而将其克隆到后续任意数量的智能体上。完整训练后的系统中，每个智能体内的两个网络均独立运行，其示意图见图 7。

7 结果

我们评估训练得到的策略网络在可扩展性、泛化能力和抗噪性方面的表现。评估指标包括在熟悉地图和陌生地图上的探索完成时间，以及通信条件存在噪声时的最终地图准确度。

7.1 在熟悉地图上的性能

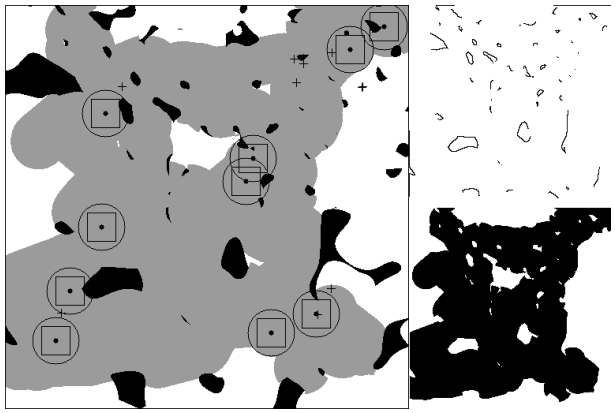


图8：在熟悉地图上的探索。左：地面地图，灰色区域为已探索区域。右上：累计障碍物地图。右下：累计探索地图。

熟悉地图根据第3节中的算法生成，其训练数据分布与训练时相匹配。图8展示了探索进行中的一个典型仿真回合。对于表3中列出的智能体数量和地图难度的每种组合，我们各进行了10个仿真回合。当探索范围达到80%或达到9000步（15分钟）上限时，回合结束。图9展示了一个示例运行的时间演变。

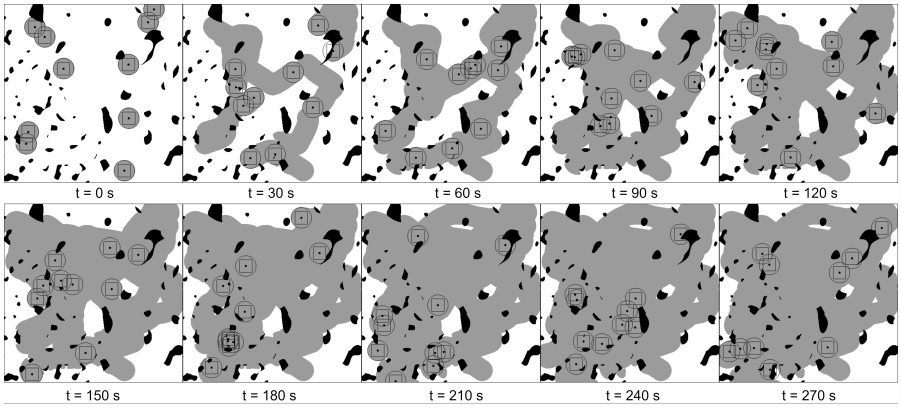


图9：探索进度快照。 $N = 10$， $d = 0.4$。

表3：评估参数设置

参数	值
机器人数量	1, 5, 10, 15, 20
难度等级	0.1、0.2、0.3、0.4
最大通信邻居数	3
地图尺寸	512×512 像素
感知半径	30 像素
机器人碰撞半径	3 像素
机器人最大速度	10 像素/秒
机器人最大加速度	5 像素/秒^2
信任参数（β）	0.8
时间步	0.1 秒
通信频率	1 个时间步
最大允许时间	9000 个时间步（15 分钟）
每组参数的回合数	10

表4：熟悉地图仿真的平均时间（秒）

N / d	0.10	0.20	0.30	0.40
1	900.00	900.00	900.00	900.00
5	305.70	321.50	322.10	391.17
10	166.22	162.33	169.50	250.33
15	122.00	126.00	131.67	196.00
20	114.00	102.33	103.67	169.00

表5：熟悉地图仿真的平均距离（像素）

N / d	0.10	0.20	0.30	0.40
1	5587.74	5549.19	5512.75	5518.47
5	9736.34	10162.66	10162.32	12148.90
10	10621.55	10453.99	11091.13	16015.63
15	12073.68	12486.27	13022.99	18914.38
20	15267.22	13408.26	13843.96	21551.27

表4和表5总结了结果。如图10所示，该方案具有良好的可扩展性，完成时间会随着智能体数量增加而缩短。不过，随着机器人编队规模扩大，碰撞规避开销也会增加，因此边际收益会逐渐减小。

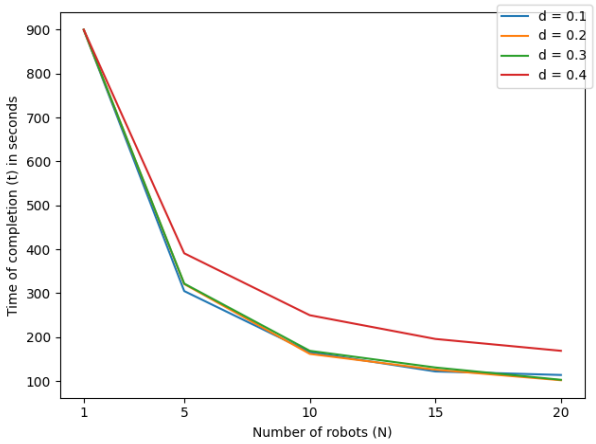


图10：完成时间与智能体数量的关系。

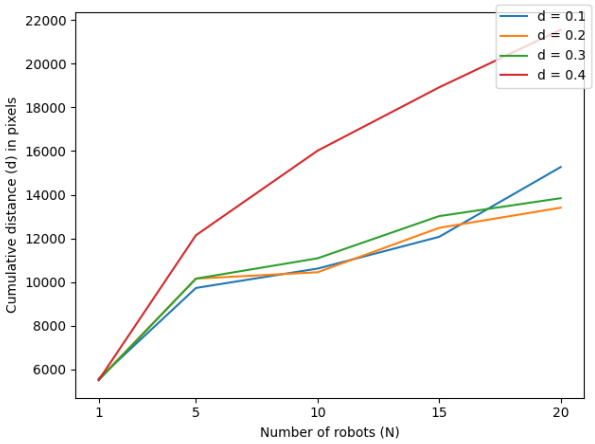


图11：累计距离与智能体数量的关系。

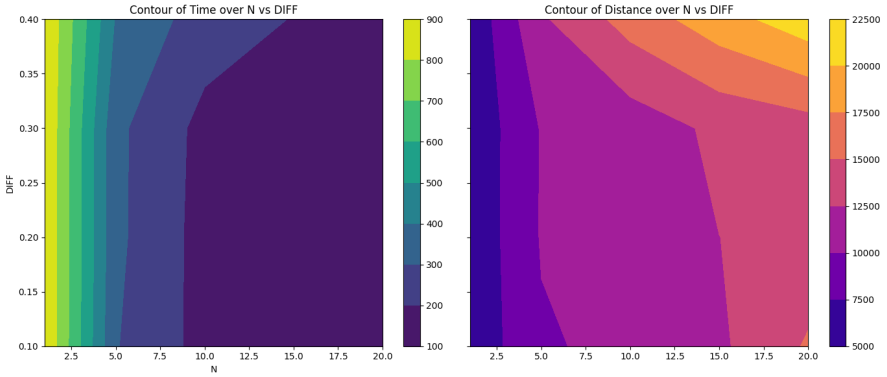


图12：完成时间（左）和累计距离（右）随智能体数量和难度变化的等高线图。

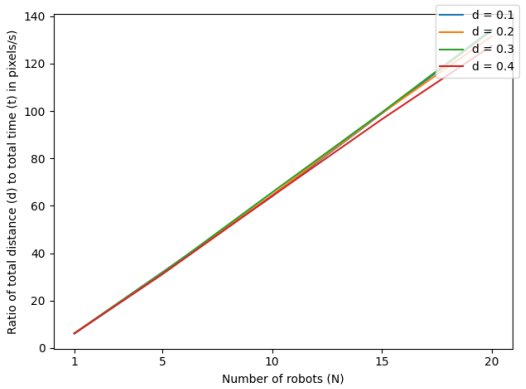


图13：单位完成时间的累计距离（像素/秒）。

图11和图12展示了这种权衡：机器人数量越多，总行驶距离就越长。图13显示，与高效探索带来的优势相比，距离惩罚会随机器人数量线性增长；此外，该惩罚不受难度级别影响。

7.2 陌生地图上的性能

为测试泛化能力，我们在包含训练数据中未出现的几何结构（圆形、三角形、尖角）的手工制作地图上评估了该系统。图 14 展示了陌生地图的示例。

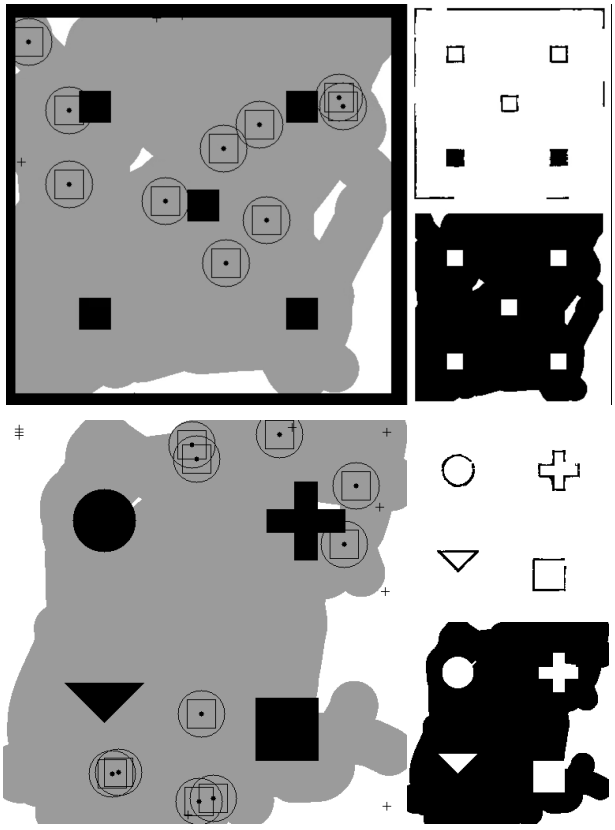


图14：两张陌生地图中的探索进度， $N = 10$。

这些地图包含不同的边界厚度和规则形状，用于检测过拟合。如图 15 所示，策略网络和自编码器具有良好的泛化能力。10 个机器人在这些地图上的表现与其在难度为 0.3–0.4 的熟悉地图上的表现非常接近。

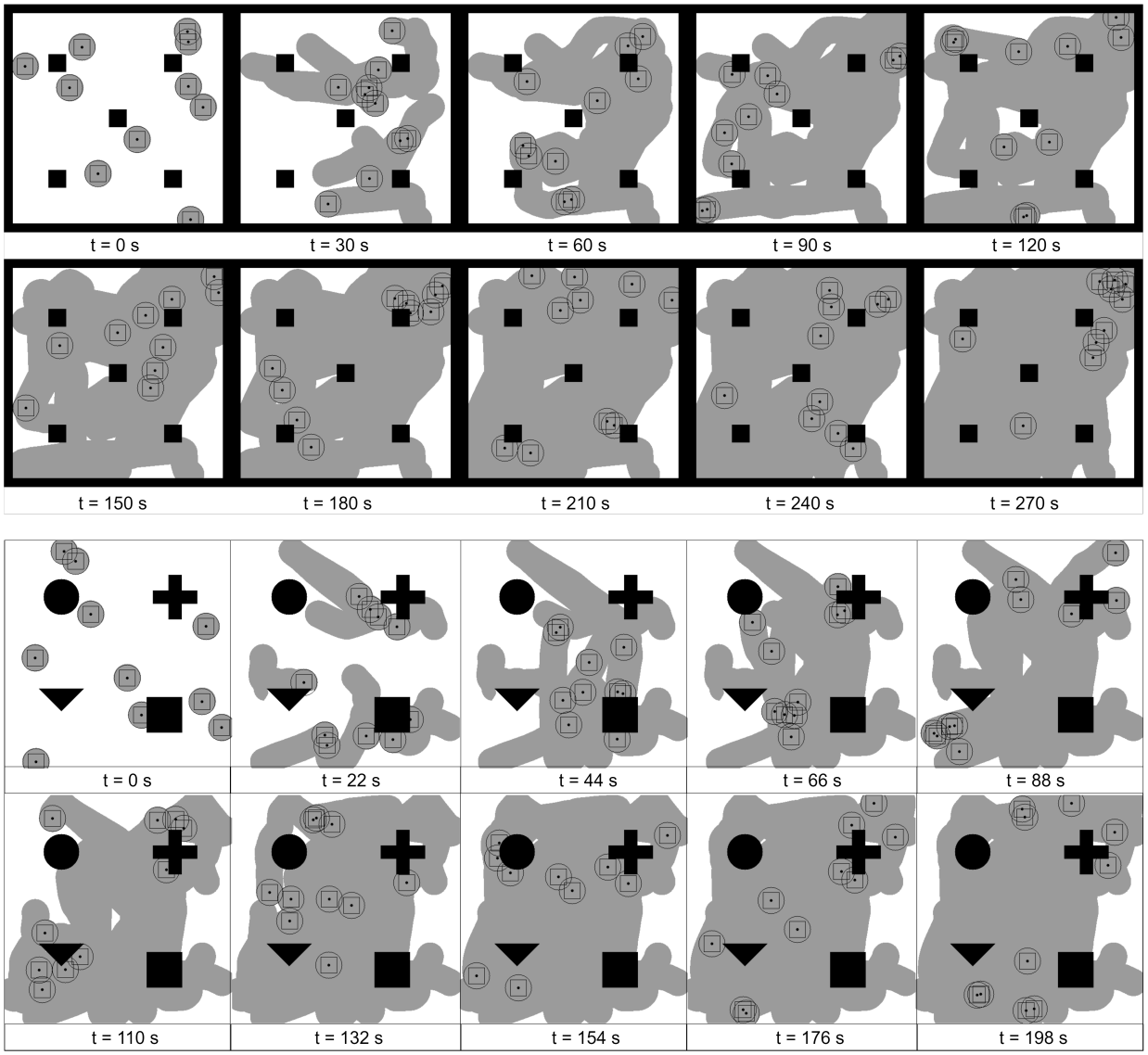


图 15：在陌生地图示例中进行探索的快照。

7.3 噪声的影响

除了研究能够实现完美通信的无噪声环境（第7.1和7.2小节）外，我们还通过在智能体间通信中注入高斯噪声，进一步评估了基于信任的融合机制的鲁棒性。相关参数见表5。

$$
\begin{aligned}
o_i^s &\leftarrow o_i^s + \mathcal{N}(0, \sigma^2) \\
e_i^s &\leftarrow e_i^s + \mathcal{N}(0, \sigma^2)
\end{aligned}
\tag{11}
$$

其中， o_i^s 和 e_i^s 是编码地图向量。地图准确度定义为真实值与已探索区域内集体障碍物地图之间的相似度：

$$\text{Accuracy} = \frac{\sum_{x,y}[(G(x,y) = O(x,y)) \wedge (E(x,y) = 1)]}{\sum_{x,y}[E(x,y) = 1]} \tag{12}$$

表 6：噪声测试的参数

参数	值
信任参数（β）	0.1, 0.3, 0.5, 0.7, 0.9
噪声标准差 (σ)	0，0.001，0.002，0.003，0.004
机器人数量	10
难度等级	0.35

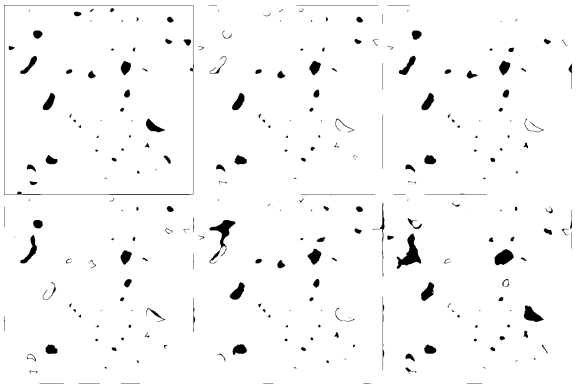


图 16：噪声对最终地图质量的影响， $\beta = 0.7$。

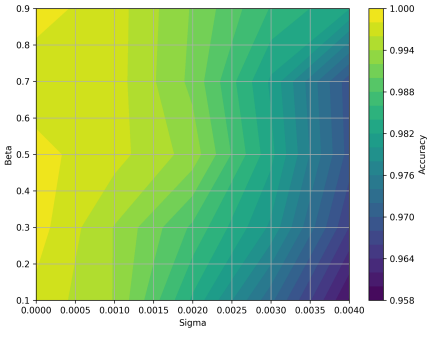


图 17： β 相对于 σ的准确率等高线图。

图 16 展示了噪声导致的障碍物形变。其定量影响如图 17 所示。在噪声水平较低时， β 的影响很小。然而，随着噪声增加，较高的 β 值会带来显著更高的准确率。这证实了在高噪声环境下，融合机制能够正确地优先采用自我感知，而非含噪通信，从而防止误差传播。

8 结论

本文提出了一个全面的多机器人探索框架，以应对可扩展性、泛化能力和抗噪性方面的关键挑战。通过将程序化环境生成与深度强化学习相结合，我们开发出一种能够在复杂未知环境中导航、且无需依赖手工设计启发式方法的系统。本文的一项重要成果是提出了一种采用掩码珀林噪声的全新分层地图生成算法，以支持鲁棒训练。该算法可生成多样且无结构的拓扑，并能全面控制障碍物尺寸和密度。针对地图数据的高维性，本文采用卷积自编码器将占据栅格压缩为潜向量，使强化学习的状态空间变得易于处理。此外，所提出的框架采用分层架构，将高层规划（探索策略）与底层规划（导航策略）解耦，进一步提高了强化学习解决方案的可训练性。为确保可扩展性和鲁棒性，信息共享仅限于自感知地图，并引入了基于信任的加权共识机制。该机制使用可调节的信任参数来控制潜在错误通信的影响。这有效防止了累积误差在机器人群中不断级联，从而确保即使在噪声通信条件下也能保持地图准确度。

在使用与生成训练地图相同的算法生成的地图上进行测试时，观察到随着机器人数量增加（1，5，10，15，20），系统完成探索所需的总时间稳步减少，但与此同时，所有机器人在结束探索前行进的累计距离稳步增加。这一趋势在多个难度等级的地图上均有所体现（0.1，0.2，0.3，0.4）。重要的是，研究发现，累计行进距离与所用时间之比仅随机器人数量的增加而线性增长。此外，即使难度级别不断提高，这一趋势也基本保持不变。这些观察结果证明了系统具有可扩展性，而在陌生的手工制作地图上进行的类似测试中

其中包含几何结构的地图也呈现出类似的模式，证明了所提出方案的泛化能力。在这些测试中，通过向通信链路注入不同程度的高斯噪声，对已实现的基于信任的加权共识机制缓解误差传播的能力进行了测试。研究发现，当噪声标准差为0.004时，调整信任参数使最终地图上的准确率从95%提升至98%。

未来的工作将侧重于利用三维卷积层和体素占据栅格，将该框架扩展到三维环境。此外，分层框架的模块化特性使高层策略能够适配纯探索之外的特定下游应用，例如搜救、自主采矿或协同运输作业。最后，Perlin 噪声地图生成算法、用于信息融合的基于信任的加权共识算法等单项成果，也可以独立应用于未来研究中，以解决性质相似的问题，例如在灾区使用机器人群进行搜救的导航算法，以及清洁或监控机器人实现地面覆盖的协调算法。

参考文献

[1] A. Quattrini Li, 《机器人群体的探索与建图：近期趋势》, *Curr. Robot. Rep.*, 第 1 卷, 第 4 期, 页 227–237, 2020 年 12 月。[2] R. R. Murphy, 《灾害机器人学》。美国马萨诸塞州剑桥：MIT出版社, 2014 年。[3] Y. Liu 和 G. Nejat, 《面向城市搜索与救援应用中半自主控制的多机器人协同学习》, *J. Field Robot.*, 第 33 卷, 第 4 期, 页 512–536, 2016 年 6 月。[4] B. Wilcox等, 《用于行星探索的机器人车辆》, 载于*1992* 年 *IEEE* 国际机器人与自动化会议论文集, 法国尼斯, 1992 年, 页 175–180。[5] W. Burgard、M. Moors、D. Fox、R. Simmons 和 S. Thrun, 《协作式多机器人探索》, 载于*IEEE* 国际机器人与自动化会议论文集, 美国加利福尼亚州旧金山, 2000 年, 页 476–481。[6] S. Saravanan、K. C. Ramanathan、M. M. Ramya 和 M. N. Janardhanan, 《多机器人系统动态任务分配策略最新进展综述》, *Ind.Robot*, 第 47 卷, 第 6 期, 页 929–942, 2020 年 8 月。[7] M. S. Talamali、A. Saha、J. A. R. Marshall、A. Reina, 《少即是多：机器人群体通过受限通信更好地适应变化》, *Sci. Robot.*, 第 6 卷, 第 56 期, 文章编号 eabf1416, 2021 年 7 月。[8] B. Yamauchi, 《一种用于自主探索的基于前沿的方法》, 载于*1997* 年 *IEEE* 计算智能与机器人自动化国际研讨会（*CIRA*）论文集, 美国加利福尼亚州蒙特雷, 1997 年, 页 146–151。[9] B. Yamauchi, 《使用多机器人进行基于前沿的探索》, 载于第二届自主智能体国际会议（*AGENTS*）论文集, 美国纽约州纽约, 1998 年, 页 47–53。[10] R. Zlot 和 A. Stentz, 《复杂任务的基于市场的多机器人协同》, *Int. J. Robot. Res.*, 第 25 卷, 第 1 期, 页 73–101, 2006 年 1 月。[11] Z. Yan、N. Jouandeau 和 A. A. Cherif, 《采用基于交易的方法进行多机器人去中心化探索》, 载于第八届信息学、控制、自动化与机器人国际会议（*ICINCO*）论文集, 荷兰诺德韦克豪特, 2011 年, 页 59–67。[12] J. H. Reif 和 H. Wang, 《社会势场：一种面向自主机器人的分布式行为控制方法》, *Robot.Autom.Syst.*, 第 27 卷, 第 3 期, 页 171–194, 1999 年 5 月。[13] C. Stachniss、O. M. Mozos 和 W. Burgard, 《使用移动机器人团队高效探索未知室内环境》, *Ann. Math. Artif. Intell.*, 第 52 卷, 第 2 期, 页 205–227, 2008 年 4 月。[14] C. Stachniss、O. M. Mozos 和 W. Burgard, 《通过考虑语义地点信息加快多机器人探索》, 载于*2006* 年 *IEEE* 国际机器人与自动化会议（*ICRA*）论文集, 美国佛罗里达州奥兰多, 2006 年, 页 1692–1697。[15] K. Arulkumaran、M. P. Deisenroth、M. Brundage 和 A. A. Bharath, 《深度强化学习简要综述》, *IEEE Signal Process. Mag.*, 第 34 卷, 第 6 期, 页 26–38, 2017 年 11 月。[16] F. Niroui、K. Zhang、Z. Kashino 和 G. Nejat, 《用于搜索与救援应用的深度强化学习机器人：在未知杂乱环境中的探索》, *IEEE Robot. Autom. Lett.*, 第 4 卷, 第 2 期, 页 610–617, 2019 年 4 月。[17] M. Kegeleirs、G. Grisetti 和 M. Birattari, 《群体同步定位与建图：挑战与展望》, *Front. Robot. AI*, 第 8 卷, 文章编号 618268, 2021 年 2 月。[18] M. Hu、J. Zhang、L. Matkovic、T. Liu 和 X. Yang, 《医学图像分析中的强化学习：概念、应用、挑战与未来方向》, *J.Appl.Clin.Med. Phys.*, 第 24 卷, 第 2 期, 文章编号 e13898, 2023 年 2 月。[19] N. Wang 、 Y. Wang 、 Y. Zhao 、 Y. Wang 和 Z. Li, 《仿真到现实：使用深度强化学习实现无人水面艇无地图导航》, *J. Mar. Sci. Eng.*, 第 10 卷, 第 7 期, 文章编号 895, 2022 年 7 月。

[20] K. Cobbe、C. Hesse、J. Hilton 和 J. Schulman，《利用程序化生成对强化学习进行基准测试》，载于第 *37* 届机器学习国际会议（*ICML*）论文集，虚拟会议，2020 年，页 2048–2056。[21] K. Perlin，《改进噪声》，*ACM Trans. Graph.*，第 21 卷，第 3 期，页 681–682，2002 年 7 月。[22] L. C. Garaffa、M. Basso、A. A. Konzen 和 E. P. de Freitas，《移动机器人探索中的强化学习：综述》，载于《IEEE神经网络与学习系统汇刊》，第 34 卷，第 8 期，页 3796–3810，2023 年 8 月。[23] S. Bai、F. Chen 和 B. Englot，《通过深度监督学习实现移动机器人的自主建图与探索》，载于*2017* 年 *IEEE/RSJ* 智能机器人与系统国际会议（*IROS*）论文集，加拿大不列颠哥伦比亚省温哥华，2017 年，页 2379–2384。[24] H. Moravec 和 A. Elfes，《利用广角声呐生成高分辨率地图》，载于*1985* 年 *IEEE* 国际机器人与自动化会议论文集，美国密苏里州圣路易斯，1985 年，页 116–121。[25] J. Weng、F. Xiao 和 Z. Cao，《多智能体信息融合系统中的不确定性建模》，载于第 *19* 届自主智能体与多智能体系统国际会议（*AAMAS*）论文集，美国南卡罗来纳州里奇兰，2020 年，页 1494–1502。[26] T. D. Kulkarni、K. R. Narasimhan、A. Saeedi 和 J. B. Tenenbaum，《分层深度强化学习：整合时间抽象与内在动机》，载于神经信息处理系统进展（*NIPS*），西班牙巴塞罗那，2016 年，页 3675–3683。[27] A. H. Tan、F. P. Bejarano、Y. Zhu、R. Ren 和 G. Nejat，《采用宏动作进行去中心化多机器人探索的深度强化学习》，*IEEE Robot.Autom.Lett.*，第 8 卷，第 1 期，页 272–279，2023 年 1 月。[28] S. Kriegel，《面向主动场景探索的未知物体自主三维建模》，博士论文，慕尼黑工业大学计算机科学系，德国慕尼黑，2015 年。[29] X. Liu 和 Y. Tan，《面向协同多智能体强化学习的封建式潜在空间探索》，载于《IEEE神经网络与学习系统汇刊》，第 34 卷，第 10 期，页 7775–7783，2023 年 10 月。[30] C. Finn、X. Y. Tan、Y. Duan、T. Darrell、S. Levine 和 P. Abbeel，《用于视觉运动学习的深度空间自编码器》，载于*2016* 年 *IEEE* 国际机器人与自动化会议（*ICRA*）论文集，瑞典斯德哥尔摩，2016 年，页 512–519。[31] T. Haarnoja、A. Zhou、P. Abbeel 和 S. Levine，《软演员-评论家：采用随机演员的离策略最大熵深度强化学习》，载于第 *35* 届机器学习国际会议（*ICML*）论文集，瑞典斯德哥尔摩，2018 年，页 1861–1870。