中图法分类号:TP391.9　　**文献标识码**:A　　**文章编号**:1006-8961(2023)07-2167-15

论文引用格式:Zhang C, Jiang W Y, Chen S Y, Zhou W and Yan F T. 2023. Multi-agent path planning based on improved double DQN. Journal of Image and Graphics, 28(07):2167-2181(张晨, 蒋文英, 陈思源, 周文, 闫丰亭. 2023. 基于双层DQN的多智能体路径规划. 中国图象图形学报, 28(07):2167-2181)[DOI:10.11834/jig.211239]

基于双层DQN的多智能体路径规划

张晨[1],蒋文英[1],陈思源[1],周文[1*],闫丰亭[2]

1. 安徽师范大学计算机与信息学院,芜湖　241000; 2. 上海工程技术大学电子电气工程学院,上海　201620

摘　要:**目的**　随着虚拟现实技术的发展,在虚拟场景中,基于多智能体的逃生路径规划已成为关键技术之一。与传统的火灾演习相比,采用基于虚拟现实的方法完成火灾逃生演练具有诸多优势,如成本低、代价小、可靠性高等,但仍有一定的局限性,为此,提出一种改进的双层深度Q网络(deep Q network,DQN)架构的路径规划算法。**方法**　基于两个结构相同的双Q网络,优化了经验池的生成方法和探索策略,并在奖励中增加火灾这样的环境因素对智能体的影响。同时,为了提高疏散的安全性和效率,提出了一种基于改进的K-medoids算法的多智能体分组策略方法。**结果**　相关实验表明提出的改进的双层深度Q网络架构收敛速度更快,学习更加稳定,模型性能得到有效提升。综合考虑火灾场景下智能体的疏散效率和疏散安全性,使用指标平均健康疏散值(average health evacuation value, AHEP)评估疏散效果,相较于传统的路径规划方法A-STAR(a star search algorithm)和DIJKSTRA(Dijkstra's algorithm)分别提高了84%和104%;与基于火灾场景改进的扩展A-STAR和Dijkstra-ACO(Dijkstra and ant colony optimization)混合算法比较,分别提高了30%和21%;与考虑火灾影响的DQN算法相比,提高了20%,疏散效率和安全性都得到提高,规划的路径疏散效果更好。通过比较不同分组模式下的疏散效果,验证了对多智能体合适分组可以提高智能体疏散效率。**结论**　提出的算法优于目前大多数常用的方法,显著提高了疏散的效率和安全性。

关键词:虚拟现实;火灾逃生演练;多智能体;深度强化学习;分组策略

Multi-agent path planning based on improved double DQN

Zhang Chen[1], Jiang Wenying[1], Chen Siyuan[1], Zhou Wen[1*], Yan Fengting[2]

1. School of Computer and Information, Anhui Normal University, Wuhu　241000, China;

2. School of Electronic and Electrical Engineering, Shanghai University of Engineering and Technology, Shanghai　201620, China

Abstract: **Objective** Rescue-oriented evacuation drills like fire escape drills have often been structured to optimize rehearsal training effect and firefighting awareness. To get sufficient evacuation experience, multiple drills are costly for related organizers. The requirement of that is based on evacuation drills, emergency drill venue, the physical condition of participants, and position information in real-time. The emerging virtual reality technology can be used to guide virtual fire escape in relevance to lower cost and risk and higher reliability. Moreover, to simulate its emergency drills in virtual scenarios, multi-agent path planning has been recognized and developed nowadays. **Method** We develop an improved double deep Q network (DQN) framework. Specifically, this virtual scenario analysis is developed through collecting enough campus information, including multiple agents, obstacles, exits, fire affected areas, and other related factors. Since all agents

收稿日期:2022-01-10;**修回日期**:2022-05-18;**预印本日期**:2022-05-25

*通信作者:周文　w.zhou@ahnu.edu.cn

基金项目:国家自然科学基金项目(61902003)

Supported by: National Natural Science Foundation of China (61902003)

are assumed on the same plane, we can convert them into two-dimensional grid diagrams via transformation gridding and coordination. Furthermore, different grids are colored and utilized in two-dimensional grid plane $\boldsymbol{m}$ to represent obstacles, fire affected areas, exits and locations of agents. According to the location of the agent in the virtual scene, the grid plane $\boldsymbol{m}$ is layered, and the grid plane $\boldsymbol{m}_1$ and the grid plane $\boldsymbol{m}_2$ can be obtained in terms of the sizes of 64 × 100 and 48 × 100 of each. In the double deep Q network, we use two double Q networks with the same structure, i. e. , $Q1$ and $Q2$, which consists of two category of convolution and full connection layers. Furthermore, input size can be interlinked to the grid planes with the same size as $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$ after environmental stratification. For the grid planes with the same size as $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$, trainable grid planes $\boldsymbol{m}'_{1t}$ and $\boldsymbol{m}'_{2t}$ can be obtained by randomly assigning the same number of black blocks with size of 1×1 to represent the duplicable location of the obstacle, and generating planes corresponding to all different starting positions to represent all status of the agent in the scene, which are used to initialize experience pools $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ and train networks $Q1$ and $Q2$. For the actual evacuation drills, the evacuation of the crowd is not completely independent and discrete. Nevertheless, due to the sociality of people, there is a certain social relationship between the people involved in evacuation, and there is often a certain phenomenon of "gathering and following" in crowd evacuation. In addition, to achieve the evacuation process of the crowd better in an actual evacuation drill, the organizer often arrange a certain number of guiders at different locations to assist the participants to complete the process of evacuation. Hence, our framework can add this guide into the virtual scenario and an improved k-medoids algorithm based multi-agent grouping strategy method is implemented. Agent-based location and relationship are involved in and the related grouping of the agents are accomplished as well, i. e. , the selection of corresponding guiding agents, and the evacuation-led of other agents in the group, and the improved path planning algorithm of double deep Q network architecture mentioned above. A reliability and efficiency of evacuation are improved further. **Result** Extensive experiment is carried out to validate our proposed methods. In the training process, the network $Q3$ of the traditional DQN method converge 24 000 batch sizes, while the $Q1$ and $Q2$ networks converge about 3 000 batch size as well. In detail, it demonstrates that the convergence performance of proposed method is significantly faster than the traditional DQN method and more stable. Additionally, to improve the evacuation efficiency and evacuation safety of the agent in fire scenarios, average health evacuation value (AHEP) is used to evaluate the evacuation effect. In AHEP criterion, it is about 84% and 104% higher than each traditional path planning methods of A-STAR, DIJKSTRA. Compared to the extended A-STAR and Dijkstra-ACO hybrid algorithm based on changeable fire scene, hybrid algorithm can be improved by 30% and 21%; Compared to DQN algorithm, it can be reached 20% higher. What is more, evacuation efficiency and safety are improved more, and evacuation effect of the planned path is much better. Furthermore, to verify the evacuation effect under different groups, we compared the AHEP values under the four groups of 4, 5, 6 and 7. When the group is 6, its value is the highest, which is 17%, 13% and 6% higher than those three cases of 4, 5 and 7. Finally, the results show that the appropriate grouping of multi-agent can improve the evacuation efficiency of agent. **Conclusion** The proposed method has its potentials to improve the evacuation efficiency and security to a certain extent.

Key words: virtual reality; fire drill; multi-agent; deep reinforcement learning; grouping strategy

0 引 言

人群大量聚集的公共场所,如学校、体育馆、火车站和大型商场等环境,一旦发生紧急情况,如地震、火灾等突发状况(Tan 等,2015),人们往往无法及时疏散,造成一些不必要的伤亡(Haghani 和 Sarvi,2016)。为此,开展科学有效的应急疏散演习是一种提高人们防范意识的主要方式,例如火灾逃生演习在很多地方经常开展,如学校、商场等。研究表明,在紧急情况下影响人群安全疏散的一般不是突发的灾难事件,而是由于缺乏疏散经验,极易使人群在疏散过程中恐慌并发生拥堵和踩踏事故(Liu 等,2021),导致人群无法安全疏散。为了解决这一问题,政府和一些部门组织了相关的疏散演练,比如在学校中常见的火灾逃生演练、地震逃生演练等。这些演练的最主要目的就是为了训练参与者积累疏散经验,寻找安全的疏散路径,以减少伤亡。然而这样的演练本身存在一定的局限性。由于不同灾害以及不同场景之间的差异性,导致其相应的安全策略

不同。并且一次演练参与的人数有限，参与者难以通过一次演练获取充分的疏散经验。为了得到充分的疏散经验，需要进行多次演练，耗费的成本极高。一些灾害如地震、洪水等难以复现，而另一些易复现的灾害如火灾等往往缺乏真实性，对于参与者来说真实度不够，缺乏指导意义。另外疏散演练需要实时跟踪参与者的移动轨迹和位置以及身体状况等信息，理论上通过手机等智能设备的辅助是可以实现的。而在实际情况下，由于疏散现场的复杂环境，往往难以实现。由于人的行为高度复杂且随机，即使是在同一场景下，不同场次的演练也会有不同的表现（Xie等，2021），因此通过疏散演练收集的数据并不完全可靠。综合上述在真实场景中演练所存在的弊端，一些研究者（Yan等，2019a，b，2020）通过构建虚拟场景替代真实场景，研究在紧急情况下人群的疏散。

由于在真实世界中人群的精神状态和行为比较复杂，而在虚拟环境中使用传统方法难以表达，虚拟场景往往缺乏真实感和逼真度。为了突破这一瓶颈问题，提出很多不同的解决方案，其中引入人工智能中有关智能体的概念是一种有效的方法，形成了基于智能体的建模方法。另外，随着机器学习的快速发展，强化学习作为人工智能领域最活跃的方向之一，日益成为一个研究热点，被用来解决很多瓶颈问题。该方法通过智能体和环境的交互，获得环境反馈给智能体的回报信息，智能体根据此信息进行决策和行动，这样的模式更加接近人类的思维过程。结合强化学习的方法在虚拟疏散演练中引入智能体的概念，体现了个体对象的自主性、自治性和智能性等特点。在智能体疏散路径的研究上，传统的强化学习如MC（Monte Carlo）、Q-Learning等有着一些缺陷。然而，由于虚拟场景环境的复杂性，其状态空间较大，可能出现维度爆炸问题。为此，考虑从深度强化学习的角度提出相关的解决方案。深度强化学习相关算法结合了深度学习的感知能力和强化学习的决策能力，使用神经网络来更新参数，逼近行为状态值函数，在一定程度上缓解了维度爆炸问题。

综上所述，本文从深度强化学习的角度出发，为多智能体在复杂虚拟场景中应急路径规划提供解决方案，主要贡献如下：1）在基于真实场景建模的虚拟场景中，模拟火灾情况下的人群疏散效果，提高了模拟演练的真实性，对人们在真实世界火灾情况下的疏散具有一定的指导意义；2）结合真实世界疏散情况，考虑群体之间的社会关系，提出一种改进的K-medoids分组策略，对智能体进行分组，并选择引导智能体，引导组内其他智能体疏散，提高了疏散的安全性和效率；3）针对火灾场景下多智能体疏散的路径规划问题，提出一种改进的双层深度Q网络（deep Q network，DQN）架构的路径规划方法，提高了模型的学习速度和稳定性，同时优化了经验池的生成方法和探索策略，并在奖励中增加火灾这样的环境因素对智能体的影响，在路径规划中综合考虑疏散效率和疏散安全性。

1 相关工作

1.1 智能体疏散模型

早期研究通常将人群视为一个整体进行疏散，基于这种假设，许多研究者提出了相关模型，如流体力学模型（Li等，2012）、势能场模型（Zhou 等，2019）等，这些都是宏观模型。在这些模型中人群受到来自外部统一方向、数量和大小的力影响，因此可以不考虑个体之间的相互作用，从而减少了人群在疏散时受影响的因素，使人群疏散的路径变得简洁，同时优化了模型的计算量。然而这样的宏观模型存在一定的局限性。在灾害发生的真实场景，通常人们的行为不仅受到灾害等外部因素的影响，在疏散过程中还会受到拥挤、踩踏等群体内部因素（Tian等，2020）的影响。所以使用宏观模型的相关方法来疏散人群不能完整地模拟疏散过程。

因此，一些研究者考虑群体内部因素，并提出了一些模型，如元胞自动机模型（Miyagawa和Ichinose，2020）、基于智能体的模型（Le等，2017）以及基于粒子系统的模型（Zhang等，2020）等，这些都是微观模型。微观模型能够从个体的角度考虑每个个体与环境的相互作用，可以弥补宏观模型描述个体行为细节的不足。如Helbing等人（2003）提出的社会力模型，是一种典型的粒子系统模型，将行人运动描述为社会力模型中力的结果。行人运动是由其内在驱动力、个体之间的相互作用、个体与环境之间的相互作用所驱动的。在模拟疏散中，既表现了个体与环境之间的交互（Li等，2020），又反映了个体之间的内在联系。韩延彬和刘弘（2018）提出一种基于避障策略的社会力模型构建疏散路径集合的方法。然而这种

基于社会力模型的方法却无法体现个体在面临环境变化时的决策能力(Li等,2019)。近年来,元胞自动机模型非常热门,被许多商业仿真软件所使用。元胞自动机以计算机建模和仿真的方法,研究类似于生物细胞等由大量并行单元个体组成的复杂系统的宏观行为与规律,适用于大规模的疏散模拟。一些研究者使用元胞自动机模型模拟人群疏散过程(Yuan和Tan,2009;Wang等,2012;Vihas等,2012;Spartalis等,2014)。由于在真实灾害发生时,人的行为具有一定的社会性,而将人群看做生物细胞却无法体现这种社会性,并且对环境变化的感知能力依旧不足,同时也无法表现出模拟演练的智能性。因此有研究者提出使用基于智能体的模型来改善上述缺陷。基于智能体的模型将所有个体视为具有一定智能性的离散实体,能够感知环境以及和其他个体交流,最终得到正确的疏散策略。

研究表明,基于智能体的模型用于人群疏散演练仿真具有一定优势。Wagner和Agrawal(2014)提出一种基于智能体的人群疏散仿真系统,对火灾情况下的人群疏散进行仿真,并研究了多种灾害场景下的疏散性能。Zhang等人(2018)利用基于智能体的模型模拟行人在标志影响下的运动,开发了用于优化大型公共设施的标识系统设计的模拟系统。在这些系统中,由于采用了基于智能体的模型,导致其中每个个体都被视做独立的,从而使仿真的计算成本增加。随着现代计算机的高速发展,对于一般二维场景的仿真来说,这种增加的计算量仍在可控范围之内。然而对于三维虚拟场景的模拟演练来说,将所有个体视作单独的智能体,进行路径规划是不现实的。而且在真实场景中,一旦发生紧急事故,由于灾害的突发性和高度危险性,部分人群的疏散会出现跟随现象。

基于这种跟随现象,有研究者在疏散仿真中提出了领导者—跟随者模型。一些研究者(Aubé和Shield,2004;Yang等,2013;Fang等,2016)提出了基于智能体的领导者模型,在这些模型中,人群跟随领导者进行疏散。Pelechano和Badler(2006)揭示了领导者的数目位置对疏散效率的影响,恰当的比例和位置可以减少模型的计算量,优化疏散效率。然而在真实灾害事件中,由于恐慌等心理因素的影响,往往很难有明确的领导者,但是在一些特殊场景中,如学校的地震火灾演练,训练有素的老师可以发挥类似领导者的作用,协助人群安全疏散。综合上述经验和研究,本文考虑在虚拟场景中设定一定数量的引导智能体,这样既能提高智能体疏散效率,又符合实际情况。

1.2 智能体路径规划

在人群疏散演练中,最主要的是规划出一条安全、快捷的疏散路径。在计算机领域,基于人群疏散的路径规划问题是重要的研究方向。早期的研究主要基于势能场的方法(Yuan和Liu,2019),从环境和群体之间的受力关系来规划路径。这样的方法不能体现群体之间的社会性和个体的智能性。于是有研究者提出将启发式搜索算法应用到路径规划中。启发式搜索是利用启发函数对搜索进行指导,从而实现高效搜索,启发式搜索是一种智能搜索,典型算法包括A-STAR(a star search algorithm)算法等。靳海亮等人(2019)提出一种改进的A-STAR算法,用于火灾逃生的路径规划。然而启发式搜索表现出的自学习、自适应和自搜索的特性,体现了个体的智能性,却忽视了群体之间的相互作用和影响。因此有人对其进行改进,提出了启发式群体智能算法,如蚁群算法(Lissovoi和Witt,2015)、蜂群算法(Wang等,2019)以及粒子群算法(Mac等,2017)等。艾子豪等人(2019)提出一种基于虚拟足迹聚类的蚁群优化算法。虽然这些群体式的智能算法保留了启发式搜索的智能性,但是在应用中往往会出现早熟和停滞现象,并且只能简单表现群体之间的关系,无法体现群体的智能性。

为了突出路径规划的智能性,提出了结合强化学习的相关算法,研究人群疏散的路径规划问题。具体来说,通过设计恰当的奖励函数来模拟智能体与环境之间的交互,使智能体能够在正确寻找目标点的同时避免与障碍物发生碰撞,这种寻路方式与人类的决策模式相似度极高,典型的是表格型算法,如Q-learning方法(Konar等,2013),该方法广泛应用于智能体疏散的路径规划。然而在真实场景中,一次大规模疏散演练的人数较多并且环境复杂,使用传统表格型的强化学习算法解决此类问题时,容易造成状态空间较大,导致维度灾难的问题。

由于深度强化学习结合了深度学习的高维数据处理能力,有研究人员提出使用深度强化学习的算法来处理高维度的状态空间。DQN(Mnih等,2015)

是第 1 个被提出的深度强化学习的框架，它是基于 Q-learning 的改进型，通过经验值回放来训练智能体。在训练初期，由于智能体缺少对环境的充分探索，导致经验值的深度和广度不够，影响了模型的学习速度和稳定性。本文以 DQN 算法框架为基础，提出一种改进的双层深度 Q 网络架构的方法，解决火灾场景下多智能体路径规划问题。

2 本文方法

本文方法框架如图 1 所示，针对火灾虚拟场景中的多智能体疏散，提出改进的 K-medoids 方法，对多智能体进行分组，并选择引导智能体，结合改进的双层深度 Q 网络方法为引导智能体规划疏散路径，同时为组内其他智能体疏散提供指引。

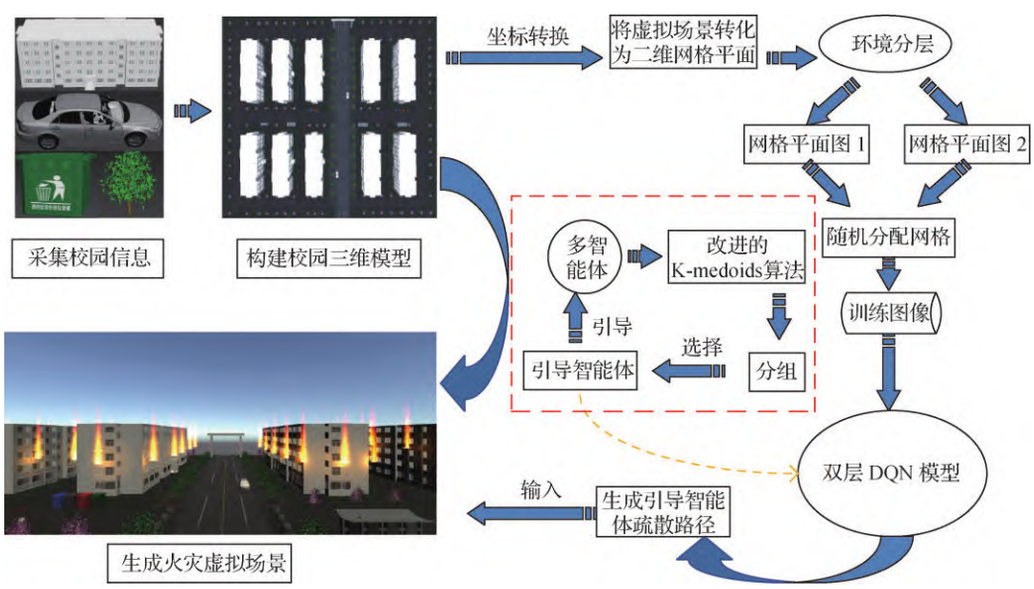


图 1　双层深度 Q 网络架构的路径规划方法框架

Fig. 1　Path planning method framework of double Q network architecture

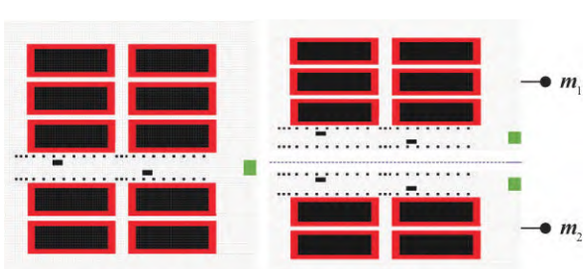
2.1 双层深度 Q 网络架构

深度强化学习算法大都是通过学习智能体与环境交互产生的数据来训练模型，构建与智能体交互的环境是关键点之一。

本文对真实场景的校园宿舍使用基于智能体的建模方法得到虚拟场景 $\boldsymbol{M}(\boldsymbol{P},\boldsymbol{G},\boldsymbol{W})$，其中包含多个智能体、单出口 $\boldsymbol{G}$、火灾影响范围 $\boldsymbol{W}$ 和多个障碍物 $\boldsymbol{P}$（包括房子、树木和汽车等）。虚拟场景 $\boldsymbol{M}$ 中包含 X，Y，Z 的三维空间的坐标信息，而本文所要解决的路径规划的问题考虑所有智能体在同一平面，因此对 $\boldsymbol{M}$ 进行降维，忽略 Z 轴的坐标信息，将其转换为二维平面图 $\boldsymbol{m}(\boldsymbol{p},\boldsymbol{g},\boldsymbol{w})$，再根据智能体的尺寸将平面图 $\boldsymbol{m}(\boldsymbol{p},\boldsymbol{g},\boldsymbol{w})$ 网格化。平面图的实际尺寸为 20 000 × 20 000 像素，以单个智能体尺寸 200 × 200 像素为单位网格的尺寸，为此得到分辨率为 100 × 100 像素的 RGB 图像，其中包含 4 种不同颜色的点，如图 2(a)所示。其中，黑色点代表障碍物 $\boldsymbol{p}$，白色点表示可通行区域，绿色点表示出口位置 $\boldsymbol{g}$，红色点表示火灾影响的范围 $\boldsymbol{w}$。当智能体遭遇障碍物 $\boldsymbol{p}$ 时，将停止移动。当智能体在火灾影响范围 $\boldsymbol{w}$ 内，可以继续运动，但是会受到一定的伤害。

根据智能体在虚拟场景中的相对位置，对环境进行分层，将 $\boldsymbol{m}$ 划分成 $\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$ 两部分，具体为

$$\boldsymbol{m}=\begin{cases}\boldsymbol{m}_1 & H\in[1,64],\quad W\in[1,100]\\ \boldsymbol{m}_2 & H\in[53,100],W\in[1,100]\end{cases}\tag{1}$$



(a) 2 维平面图 $\boldsymbol{m}$　　(b) 环境分层的结果 $\boldsymbol{m}_1$ 和 $\boldsymbol{m}_2$

图 2　可训练的网格图

Fig. 2　Trainable grid diagram

((a) the grid plane $\boldsymbol{m}$; (b) layered $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$)

式中，H和W表示网格图的行号和列号，$\boldsymbol{m}_1$表示$\boldsymbol{m}$中[1, 64]行的环境，网格图的尺寸为64 × 100像素，$\boldsymbol{m}_2$表示$\boldsymbol{m}$中[53, 100]行的环境，网格图的尺寸为48 × 100像素。划分结果如图2(b)所示。

本文改进了DQN算法中的网络结构，提出一种双层深度Q网络架构，如图3所示，使用两个结构相同的双Q网络，分别是$Q1$和$Q2$，都由两个卷积层和两个全连接层组成，网络层数较少，它们的输入分别对应与环境分层后$\boldsymbol{m}_1$和$\boldsymbol{m}_2$的尺寸相同的网格图。$Q1$和$Q2$的具体参数详见表1。

在$Q1$网络中，输入为64 × 100 × 3的网格图。第1层是1个卷积层，包含10个2 × 2 × 3的卷积核，步长为1，使用ReLU(linear rectification function)函数作为激活函数，并通过矩阵补充可以得到64 × 100 × 10的输出。第2层也是1个卷积层，包含20个2 × 2 × 10的卷积核，步长为1，通过ReLU函数和输入矩阵补充可以得到64 × 100 × 20的输出。第3层是1个全连接层，将第2层卷积的输出64 × 100 × 20打平成1 × 128 000的张量作为输入，包含100个神经元，输出是1 × 100。第4层也是1个全连接层，包含4个神经元，输出4个状态动作值。

在$Q2$网络中，输入为48 × 100 × 3的网格图。第1层包含10个2 × 2 × 3的卷积核，步长为1，使用ReLU函数作为激活函数。通过输入矩阵补充，得到48 × 100 × 10的输出。第2层同样是1个卷积层，包含20个2 × 2 × 10的卷积核，步长为1，通过ReLU函数和输入矩阵补充可以得到48 × 100 × 20的输出。第3层是1个全连接层，将卷积层2中的输出打平成1 × 96 000的张量作为输入，包含100个神经元，输出是1 × 100。第4层也是一个全连接层，包含4个神经元，同样输出4个状态动作值。

表1　$Q1$和$Q2$网络的相关参数

Table 1　Relevant parameters of $Q1$ and $Q2$ networks

层	$Q1$输入尺寸	$Q2$输入尺寸	参数
Cov1	64 × 100 × 3矩阵补充	48 × 100 × 3矩阵补充	10个2 × 2 × 3卷积核，Stride为1
Cov2	64 × 100 × 3矩阵补充	48 × 100 × 3矩阵补充	20个2 × 2 × 10的卷积核，Stride为1
FC1	拉平成1 × 128 000张量	拉平成1 × 96 000张量	输出：1 × 100
FC2	1 × 100	1 × 100	输出：1 × 4

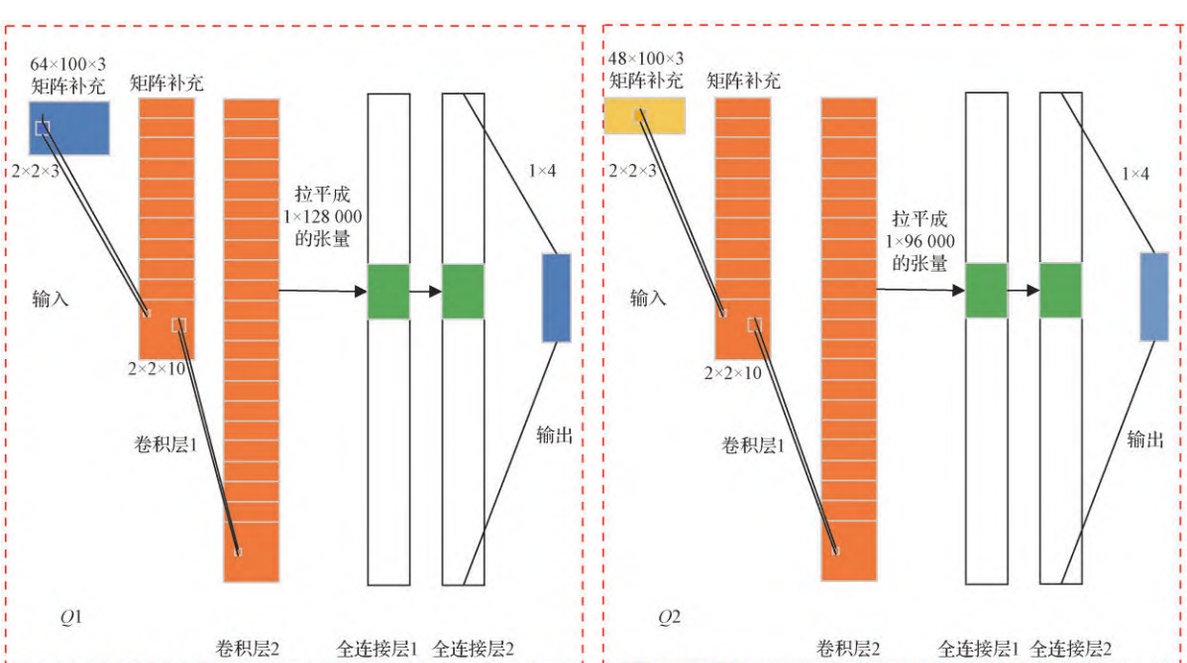


图3　双层深度Q网络结构

Fig. 3　The structure of double deep Q network

2.2 双层深度Q网络架构的路径规划算法

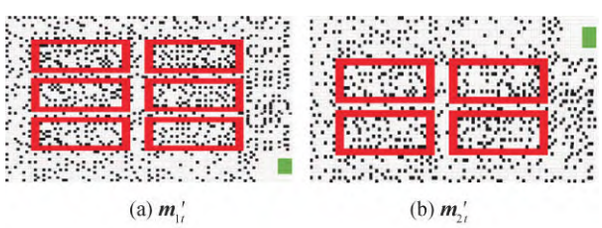
本文提出一种改进的基于双层Q网络架构的路径规划方法。为了解决在火灾虚拟场景下多智能体的疏散问题，即找到一个从当前位置S到出口位置G的最佳路径，有别于传统的路径规划算法中对最佳路径定义的距离最短或使人群疏散最快的概念，为了最大限度模拟真实火灾场景中人群的疏散情况，这里的最佳路径考虑了智能体在疏散过程中所受到的伤害，规划了一条使智能体受到伤害最低同时疏散效率尽可能高的逃生路径。

DQN算法使用深度卷积网络$Q(S_t, a_t, \theta)$拟合$Q(S_t, a_t)$，通过更新Q函数中的θ来更新Q函数，具体为

$$\theta_{t+1} = \theta_t + \alpha \times \left[R_t + \gamma \max_{a_{t+1} \in A} Q(S_{t+1}, a_{t+1}, \theta^{\mathrm{TD}}) - Q(S_t, a_t, \theta)\right] \quad (2)$$

式中，$R_t + \gamma \max_{a_{t+1} \in A} Q(S_{t+1}, a_{t+1}, \theta^{\mathrm{TD}})$是时间差分(temporal-difference)目标，网络更新权重使$Q(S_t, a_t, \theta)$接近TD目标。

对于传统的DQN算法来说，受限于单一场景的训练环境，导致智能体能获取的经验不够充分，因此本文使用多个场景作为训练环境。为了得到多个不同场景的网格图，通过固定出口位置$\boldsymbol{g}$、火灾影响的范围$\boldsymbol{w}$，在与$\boldsymbol{m}_1$和$\boldsymbol{m}_2$相同尺寸的图像中，通过e次随机分配相同数量1×1的黑色块，表示障碍物所在位置，最终可得到e幅可训练的网格图$\boldsymbol{m}'_{1t}(\boldsymbol{p}', \boldsymbol{g}', \boldsymbol{w}')$和$\boldsymbol{m}'_{2t}(\boldsymbol{p}', \boldsymbol{g}', \boldsymbol{w}')$，如图4所示。



(a) $\boldsymbol{m}'_{1t}$　(b) $\boldsymbol{m}'_{2t}$

图4　不同场景的可训练网格图

Fig. 4　Trainable grid diagram of multiple different scenes ((a) $\boldsymbol{m}'_{1t}$; (b) $\boldsymbol{m}'_{2t}$)

在网格化的路径规划问题中，通常有两种不同的动作空间，即四邻域和八邻域。本文不考虑智能体的运动轨迹，采用四邻域作为动作空间，包括上、下、左、右4个动作。奖励是智能体与环境交互并反馈给模型的唯一信息，模型根据奖励进行学习。因此奖励的设计决定了模型学习的能力和训练的效率，也是算法的关键点之一。为了找到疏散的最佳路径，智能体要在躲避障碍物的同时尽可能远离火灾的影响范围，并使所受伤害最低。基于上述需求，将奖励函数定义为

$$r_t = \begin{cases} r_p & z = p \\ r_g & z = g \\ r_w & z = w \\ 0 & \text{其他} \end{cases} \quad (3)$$

式中，r_p代表智能体遇上障碍物时，受到来自于环境的反馈。由于智能体在疏散过程中避让障碍物，将其设置为一个较大的负值-100，作为一个负反馈。r_g表示智能体到达出口位置时的反馈，为了使智能体能够寻找到出口，将其设置为一个较大的正值100，作为一个正反馈。r_w表示智能体进入火灾影响的范围内的反馈，将其设置为一个较小的负值-5。其他情况下奖励为0。

显然，式(3)中，整个奖励函数包含4种状态，分别是智能体到达障碍物位置$z = p$，到达出口位置$z = g$，到达火灾影响范围$z = w$和其他情况。图5反映了智能体在疏散过程中可能表现出的状态以及相关的奖励值。

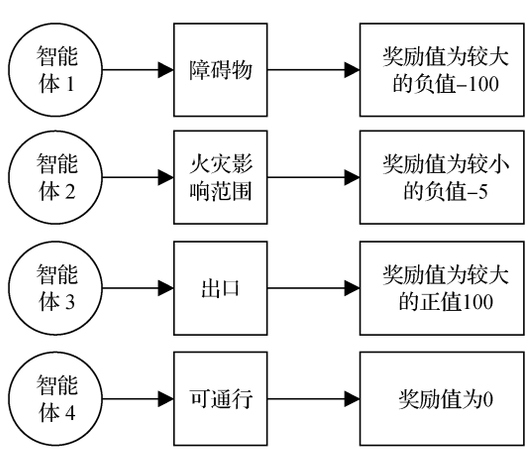


图5　智能体的状态和相关的奖励值

Fig. 5　Agent status and related reward value

DQN算法是一种异策略时间差分方法，其中使智能体与环境交互产生数据的策略称为行动策略，而所需评估的策略称为目标策略。通常对于一般的DQN算法，为了能够探索环境，使用ε贪婪策略作为行动策略，具体为

$$\pi(s_t)=\begin{cases}a_t' & \text{其他}\\ \arg\max\limits_{a_t\in A} q(s_t,a_t) & \mu\leqslant\varepsilon\end{cases}\tag{4}$$

式中，μ是每轮从[0,1]中产生的一个随机数，ε值一般设置0.01。当μ小于贪婪值ε时，在t轮选取使$q(s_t,a_t)$最大的动作a_t。反之，从剩下的动作空间中随机选取一个动作a_t'。基于这样的行动策略，将产生大量的数据，也称为经验，并使用一个五元组$\boldsymbol{e}_t(s_t,a_t,r_t,s_{t+1},E_t)$来表示。初始化经验池$\boldsymbol{D}$的大小为$N$，将$\boldsymbol{e}_t$存入其中，得到经验池。然而在模型训练初期，由于经验池中包含的$\boldsymbol{e}_t$数量较少，训练效果不好。为了克服以上问题，本文提出一种包含环境中全部状态信息的经验池生成方法，即在模型开始训练之前，将智能体在环境中能够表现出的全部状态输入到经验池。对e幅不同场景的可训练网格图$\boldsymbol{m}_{1t}'$和$\boldsymbol{m}_{2t}'$生成对应所有不同起始位置的图像，表示智能体在该场景中的全部状态，分别得到$e\times64\times100$幅$\boldsymbol{m}_{1t}'(\boldsymbol{s}_{it}',\boldsymbol{p}',\boldsymbol{g}',\boldsymbol{w}')$和$e\times48\times100$幅$\boldsymbol{m}_{2t}'(\boldsymbol{s}_{it}',\boldsymbol{p}',\boldsymbol{g}',\boldsymbol{w}')$网格图，$\boldsymbol{s}_{it}'$表示智能体位置。这些图像包含了五元组中的$(s_t,a_t,s_{t+1},E_t)$4种信息，表示不同场景中的智能体的全部状态。此外，定义了一个函数ψ，用来表示每幅图像对应的分数，具体为

$$\psi_{it}=\begin{cases}1+r_p & z=p\\ 1+r_g & z=g\\ 1+r_w & z=w\\ 1 & \text{其他}\end{cases}\tag{5}$$

式中，参数z表示当前位置，p表示障碍物位置，g表示出口位置，w表示火灾影响范围。根据式(5)，每幅图像的不同状态赋予了不同的分数。此外，将所有生成的图像分成两个模块，分别是ψ值等于1和$1+r_w$的可通行模块$\boldsymbol{\Psi}_1$，以及ψ值等于$1+r_p$和$1+r_g$的模块$\boldsymbol{\Psi}_2$，即智能体到达目标点或遭遇障碍物。并且，定义目标函数为

$$y_{it}=\begin{cases}\psi_{it+1}-\psi_{it} & \psi_{it+1}\in\boldsymbol{\Psi}_2\\ \psi_{it+1}-\psi_{it}+\gamma\max\limits_{a_{it+1}\in A} q_j(s_{it+1},a_{it+1},\theta^{\mathrm{T}}) & \psi_{it+1}\in\boldsymbol{\Psi}_1\end{cases}\tag{6}$$

式中，ψ_{it+1}是ψ_{it}图像在动作a_{it}作用下的下一图像对应的分数，即智能体的当前位置图像的分数和下一位置的分数，而$\psi_{it+1}-\psi_{it}$的差值表示最终受到环境所反馈的奖励值。y_{it}表示时间差分目标，同时也是目标函数，当没有碰撞或到达目标位置时，y_{it}为$\psi_{it+1}-\psi_{it}+\gamma\max\limits_{a_{it+1}\in A} q_j(s_{it+1},a_{it+1},\theta^{\mathrm{T}})$。而发生碰撞或到达目标的情况下，其值为$\psi_{it+1}-\psi_{it}$。

综合以上改进，本文提出一种基于改进的双层深度Q网络架构的路径规划算法。初始化经验池$\boldsymbol{D}_1$和$\boldsymbol{D}_2$，将$e\times64\times100$幅$\boldsymbol{m}_{1t}'$和$e\times48\times100$幅$\boldsymbol{m}_{2t}'$网格图，分别输入到经验池$\boldsymbol{D}_1$和$\boldsymbol{D}_2$中，在开始训练之前，得到包含全部环境状态的经验池。在训练时将每个经验池分成$\boldsymbol{\Psi}_1$和$\boldsymbol{\Psi}_2$两个不同的模块，设定参数J，从$\boldsymbol{\Psi}_1$中取样数量为m的图像作为一个批次(batch size)，计算本文设定的目标函数y_{it}。训练完成后，根据智能体在图像中的起始位置选择其对应的目标网络，为其规划疏散路径。具体步骤如下：

输入：$Q1$为$64\times100\times3$；$Q2$为$48\times100\times3$；

输出：$Q1$为1×4；$Q2$为1×4；N_1为$e\times64\times100$；N_2为$e\times48\times100$。

1)分别初始化经验池$\boldsymbol{D}_1$和$\boldsymbol{D}_2$大小为N_1和N_2。

2)根据环境生成$\boldsymbol{m}_{1t}'$和$\boldsymbol{m}_{2t}'$以及其对应的ψ值，分别输入到经验池$\boldsymbol{D}_1$和$\boldsymbol{D}_2$中。

3)在经验池$\boldsymbol{D}_1$和$\boldsymbol{D}_2$中将图像分成$\boldsymbol{\Psi}_1$和$\boldsymbol{\Psi}_2$两个不同的模块。

4)用随机的权重θ_0初始化$Q1$和$Q2$网络的权重θ和目标网络$Q1'$和$Q2'$的θ^{T}，设定batch size大小为m。

5)同时训练$Q1$和$Q2$双网络。

6)For episode in range(EPISODE)

(1)从经验池$\boldsymbol{D}_1$和$\boldsymbol{D}_2$中选择其对应$\boldsymbol{\Psi}_1$模块中的所有图像，数量为η。

(2)分别在$Q1$和$Q2$网络中初始化$J=C\times\eta/m$。

(3)For j in range(J)。

a)从$\boldsymbol{\Psi}_1$模块中随机采样数量为m的图像，根据式(6)计算目标函数y_{it}。

b)使用均方损失误差$\sum\limits_{it=1}^{m}(y_{it}-q_j(s_{it},a_{it},\theta))^2$来更新损失函数。

c)End for。

(4)将θ分别复制给其对应的目标网络θ^{T}。

(5)清空经验池$\boldsymbol{D}$。

7)End for。

2.3 基于改进的K-medoids分组策略算法

在实际的疏散演练中，人群的逃生并不完全是独立分散的。由于人具有社会性，参与疏散的人群

之间存在着一定的社会关系，如朋友、亲人等，受到这种社会关系以及现场复杂环境的影响，人群的疏散往往会出现一定的聚集跟随现象。此外，在一般的疏散演练中组织者为了更好地疏散人群，会在不同位置安排一定数量的安全员，协助参与者完成疏散。因此，本文将这种真实世界的安全员引入到虚拟场景中，同时综合考虑疏散中存在的聚集跟随现象，对智能体进行分组，选择相应的引导智能体并引导组内其他智能体疏散。

采用聚类方法对多智能体进行分组和选取引导智能体。事实上，聚类方法是按照某个特定标准（如距离准则）将一个样本集划分成不同的类，使同一个类中的对象具有一定的相似性，使不同的类中数据对象具有明显区分度。即聚类后同一类的对象尽可能聚集到一起，不同的对象尽量分离。在本文中，所需划分的对象是虚拟场景中的多智能体，将所有智能体位置的坐标信息(X,Y)作为数据集。此外考虑智能体之间的社会性，本文提出一种基于亲密度改进的K-medoids方法。此方法综合考虑智能体位置和智能体关系的影响，对两类不同类型的特征值进行加权。在智能体疏散的过程中，需要在每个群体中选择一个中心点作为引导智能体，这个中心点使集群中的其他点和其差异最小。本文定义了一个差异函数来综合评估群体之间的亲密度和距离的影响。具体为

$$T(I,J)=\frac{k_1\times D(I,J)}{\nu}+\frac{k_2\times R(I,J)}{\zeta} \tag{7}$$

$$R(I,J)=\left|V_I-V_J\right| \tag{8}$$

$$G_t=\sum_{J=1}^{k}T(P,J) \tag{9}$$

式中，$k_1+k_2=1$，是距离和亲密度的权重比，$T(I,J)$表示差异函数，$D(I,J)$表示两个智能体之间的距离，本文采用欧氏距离度量智能体之间的距离，$R(I,J)$表示智能体之间的亲密度。V_I和V_J分别表示两个智能体之间的亲密值，$R(I,J)$的值越小，代表智能体之间越亲密。ν和ζ分别是距离和亲密度的归一化因子。P表示中心点，k表示该中心点所属群体的剩余点数量，t表示迭代的轮数，G_t表示中心点到组内剩余点的差异函数的总和。

本文将定义的差异函数作为中心点更替的标准。综合考虑智能体之间距离和亲密度的关系，选择引导智能体并对智能体进行分组。整个改进后的K-medoids算法的具体步骤如下：

输入：聚类数K（引导智能体的数量），以及所有智能体坐标信息的集合$\boldsymbol{Z}(X,Y)$。

输出：满足所有约束条件的智能体分组情况。

约束条件：1）每个小组包括1个引导智能体和1个智能体；2）每个智能体只属于其中一个小组；3）在每一轮迭代中，每组中心点满足$G_t\leqslant G_{t-1}$。

初始化：从$\boldsymbol{Z}$中随机选取K个点作为中心点。

Repeat

For each I do

1）计算群体中剩余点到各个中心点的差异函数$T(I,J)$，将其分配给差异函数最小的中心点，形成分组。

2）判断该分组是否满足约束条件，不满足则重新初始化。

3）满足约束条件，在每个组内更新中心点。

4）将组内每个点当作中心点计算每个点的G_t值，选择其值最小的点，作为新中心点。

End for。

当E_t值不再变化的时候，智能体分组完成。

3 实　验

为了在火灾虚拟场景下完成对智能体的疏散，本文提出一种改进的双层深度Q网络架构的路径规划方法，为智能体规划一条最佳的疏散路径，结合改进的K-medoids分组策略算法对多智能体分组，每组内选择对应的一个引导智能体，并引导组内智能体疏散。

3.1 实验设置

虚拟场景是智能体疏散和模型训练的环境。本实验使用Unity 3D构建一个基于真实学校宿舍环境的虚拟场景，模拟宿舍区域发生火灾时人群的疏散过程。运行实验程序的计算机配置如下：Windows 10操作系统，AMD R7-4800H处理器，NVIDIA GeForce RTX 2060显卡，DDR4 16 GB内存。

为了最大限度提高虚拟场景的真实性质，通过调查收集了学校宿舍环境的相关信息，使用3ds MAX软件先构造一个3D模型。图6给出了3D模型的二维平面图，蓝色区域表示宿舍楼，共10栋宿舍楼。灰色区域表示可通行区域，包含一条主干道路和若干校园人行区域。黑色区域表示主干道路边停

放的两辆汽车，而绿色区域表示树木和植物，白色表示石头。其中红色区域表示火灾对智能体的影响范围，如图所示集中在宿舍楼周围。在这样的场景中，除了灰色表示的区域，红色区域表示的火灾影响范围也是可通行的。显然，相比于正常的主干道路和人行区域，智能体在疏散过程中经过火灾影响的范围时会受到伤害。

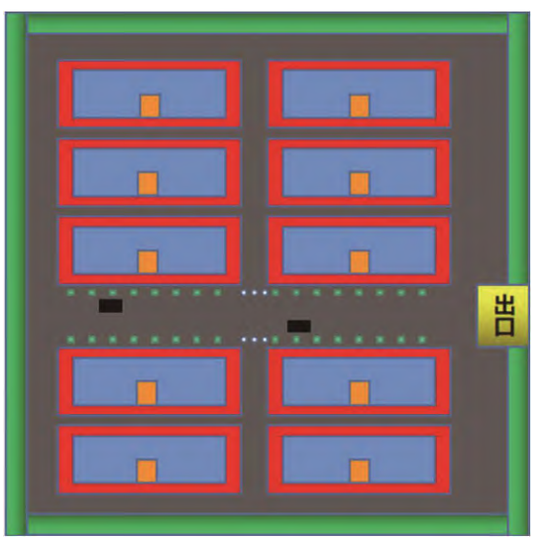


图6　校园虚拟场景的二维平面图

Fig. 6　2D plane map of campus virtual scene

本文在虚拟场景中初始化智能体的数量为180，并作为改进的K-medoids分组策略中的智能体总数。图7展示了分组为6时的智能体分布情况，6种不同颜色对应6个不同的分组，每个分组中的星型点表示该分组的引导智能体，横坐标X表示网格化平面图$\boldsymbol{m}$的行数，纵坐标Y表示$\boldsymbol{m}$的列数，对应智能体的初始位置。

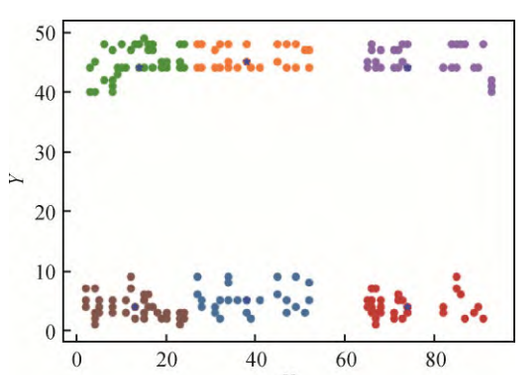


图7　分组为6时智能体的分布情况

Fig. 7　Distribution of agents when grouped into 6

3.2　实验结果与分析

为了验证改进的双层深度Q网络架构的性能，与基于相同改进的单层深度Q网络进行比较。实验中，将batch size大小设置为50，初始学习率(learning rate)设定为0.001，按每2 000步的衰减速度和系数0.9进行衰减。设定折损系数γ为0.9，更新频率C为4，场景数量e为200。

损失值的变化如图8所示，包括双层深度Q网络中的网络$Q1$和$Q2$以及单层深度Q网络中的网络$Q3$。其中，$Q1$网络对应经验池$\boldsymbol{D}_1$，包含200×64×100幅不同的可训练图像$\boldsymbol{m}'_{1t}$，$Q2$对应经验池$\boldsymbol{D}_2$，包含200×48×100幅不同的可训练图像$\boldsymbol{m}'_{2t}$。$Q3$对应经验池D，包含200×100×100幅不同的$\boldsymbol{m}'$，$\boldsymbol{m}'$是未经过环境分层的可训练图像。在$Q1$和$Q2$网络中，大约在3 000 batch size处收敛，在$Q3$网络中，大约在24 000 batch size处收敛。可以看出，双层深度Q网络架构在训练早期拥有更陡峭的梯度，收敛速度更快。在整个训练过程中，损失振荡幅度较低并且更加稳定。因此，本文提出的双层深度Q网络架构提高了模型的性能。

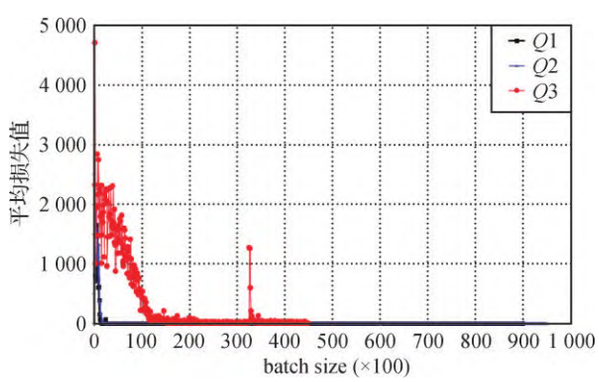


图8　训练过程中的损失值

Fig. 8　Loss value during training

为了评估规划路径的疏散效果，通常使用疏散时间作为指标进行比较。然而不同于一般的人群疏散，智能体在疏散过程中通过火灾影响范围会受到一定的伤害。因此在考虑疏散时间的同时，也要考虑疏散的安全性，即在疏散过程中尽可能使智能体受到的伤害最小。

基于以上因素，定义了健康值、平均健康值和平均健康疏散值等3个指标，用于评估规划路径的疏散效果。

1)健康值(health value，HP)指单个智能体在疏

散过程中其自身健康状态的评估值。

2)平均健康值(average health value，AHP)指多个智能体在疏散过程中其自身健康状态的平均值。具体为

$$O_t = \sum_{i=1}^{c} \frac{HP_{it}}{c} \tag{10}$$

式中，HP_{it}表示单个智能体在某时刻的HP值，c表示智能体的数量。O_t值越大，表明当前时刻所有智能体的健康状态越好。

3)平均健康疏散值(average health evacuation value，AHEP)指多个智能体疏散完成时的平均健康值与全员完成疏散时间的比值。具体为

$$K = \frac{O_{t=l}}{l} \tag{11}$$

式中，l表示疏散完成所用的总时间。K值越大，表示该方法越兼顾疏散的效率和安全性。

本文使用指标AHEP评估多智能体场景下的疏散效果，其值越大，表示规划路径的疏散效果越好。同时，与传统的路径规划方法A-STAR、DIJKSTRA(Dijkstra's algorithm)算法、考虑火灾影响的DQN方法、基于火灾场景改进的扩展A-STAR算法(程鹏举等，2020)和Dijkstra-ACO(Dijkstra and ant colony optimization)混合算法(曹祥红 等，2020)进行比较，验证提出的双层深度Q网络架构的路径规划方法的疏散效果。

疏散过程中，智能体健康值的变化情况如图9所示。其中，纵轴表示180个智能体的平均健康值，每个智能体的初始健康值设为100。横轴表示疏散时间。可以看出，双层深度Q网络架构的路径规划方法在所有智能体疏散完成后，所剩的平均健康值(AHP)远高于其他方法。考虑火灾影响的DQN方法和扩展A-STAR以及Dijkstra-ACO混合算法相较于A-STAR和DIJKSTRA有一定提升，但仍低于本文提出的双层深度Q网络架构的路径规划方法。

表2给出了详细参数和不同方法的平均健康疏散值(AHEP)。A-STAR和DIJKSTRA的平均健康值(AHP)分别为42.8和37，扩展的A-STAR和Dijkstra-ACO混合算法的AHP分别为72.5和76，考虑火灾影响的DQN方法的AHP为81.5，都低于提出的双层深度Q网络架构的路径规划方法。并且提出的双层深度Q网络架构的路径规划方法的AHEP高出传统的A-STAR和DIJKSTRA方法84%和104%；高出扩展A-STAR和Dijkstra-ACO混合算法30%和21%，相较于考虑火灾影响的DQN方法也高出20%，表明其在火灾虚拟场景中的疏散效果更好。

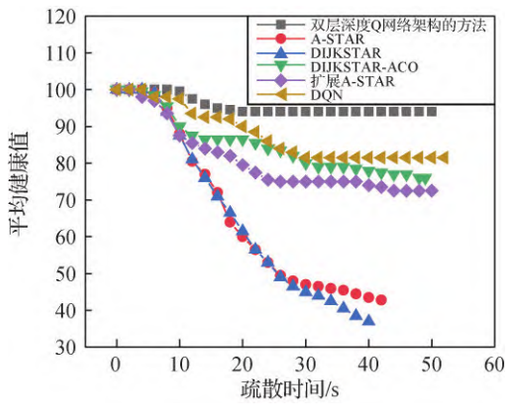


图9　不同方法的平均健康值对比

Fig. 9　Comparison of average health values for different methods

表2　不同方法的疏散时间和健康值

Table 2　Average health evacuation value for different methods

方法	疏散时间/s	AHP	AHEP
A-STAR	42	42.8	1.02
DIJKSTRA	40	37	0.92
扩展A-STAR	50	72.5	1.45
Dijkstra-ACO	49	76	1.55
DQN	52	81.5	1.56
本文	50	**94**	**1.88**

注：加粗字体表示各列最优结果。

在火灾虚拟场景中，疏散效率同时受疏散路径和有序性的影响。在本文方法中，通过选择引导智能体来协助其他智能体有序的疏散。为了验证引导智能体的数量对疏散有序性的影响，基于改进的K-medoids算法，对智能体进行不同数量的分组，并比较不同分组的AHEP值，结果如图10所示。图10(a)反映了4个引导智能体的分组情况，4种颜色将其分为4组，每组中的星型点表示该组的引导智能体，图10(b)反映了5个引导智能体的分组情况，5种颜色分为5组，每组包含1个引导智能体，图10(c)反映了7个引导智能体分组的情况，7种颜色分为7组，同样每组内包含1个引导智能体。表3中给出了提出的双层深度Q网络架构的路径规划方法对应

不同分组的AHEP，当分组为6时，引导智能体的数量为6，其结果最佳。

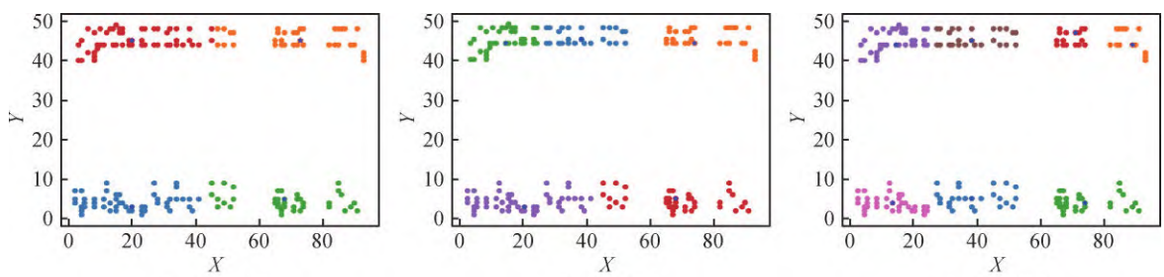
表3　不同分组的AHEP

Table 3　Average evacuation health value for different groups

引导智能体	疏散时间/s	AHP	AHEP
4	60	96	1.6
5	58	**96.5**	1.66
6	50	94	**1.88**
7	50	88.4	1.76

注：加粗字体表示各列最优结果。

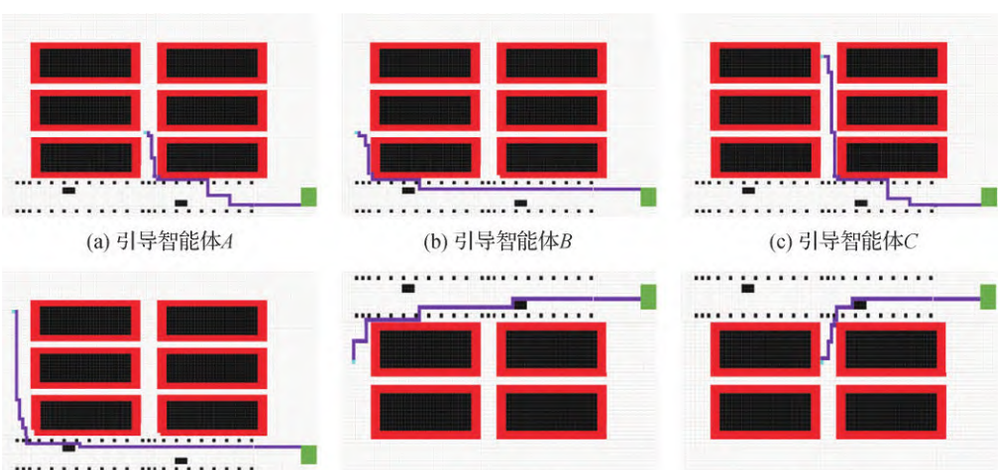
图11分别给出了智能体分组为6时，基于改进的双层深度Q网络架构的路径规划方法为引导智能体规划的疏散路径，黑色点表示障碍物，红色点表示火灾影响范围，蓝色点表示智能体初始位置，紫色点构成的路径就是智能体疏散的路线。可以看出，所有的引导智能体均成功到达目标位置，并且在疏散过程中最大限度绕开火灾影响范围，同时实验也表明了智能体所受到的伤害最少，疏散效果较好。图12给出了多智能体疏散的可视化效果，当疏散时间约为48 s时，智能体疏散基本完成，所有智能体全部疏散完成约用时50 s。



(a) 分组为4　(b) 分组为5　(c) 分组为7

图10　不同分组的智能体和引导智能体分布情况

Fig. 10　Distribution of agents in different groups and guiding agents((a) 4 groups; (b) 5 groups; (c) 7 groups)



(a) 引导智能体A　(b) 引导智能体B　(c) 引导智能体C

(d) 引导智能体D　(e) 引导智能体E　(f) 引导智能体F

图11　6个引导智能体的疏散路径

Fig. 11　Six evacuation paths guiding agent((a) guiding agent A; (b) guiding agent B; (c) guiding agent C; (d) guiding agent D; (e) guiding agent E; (f) guiding agent F)

3.3　方法的不足

根据上述实验结果，通过与相同改进的单层深度Q网络比较，证明了双层深度Q网络架构的性能优异性。并与一些常用的路径规划方法比较，结果表明提出的双层深度Q网络架构的路径规划方法优于一些常见方法。然而在基于DQN框架的相关方法中，智能体会出现路径漂移现象(path wandering phenomenon)(Lyu等，2019)，表现为训练过程中智能体在某位置重复来回，影响学习的效率，提出的方法中也出现了此类现象。

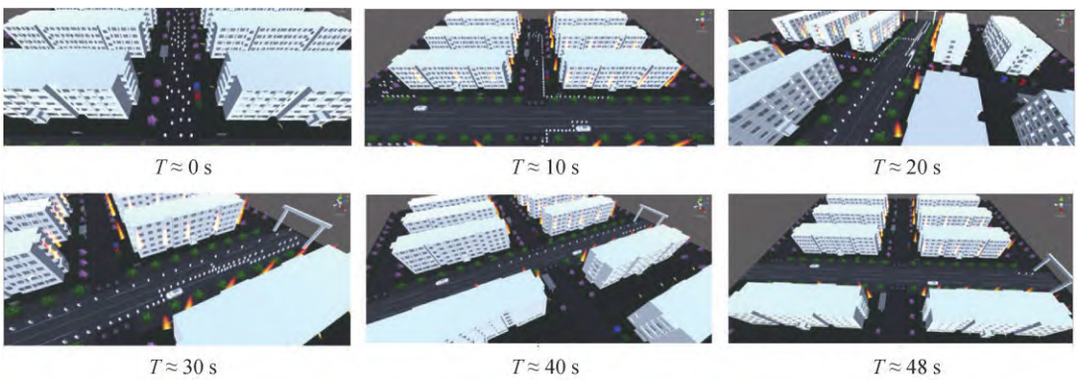


图12　多智能体疏散可视化效果

Fig. 12　Visualization of evacuation by multi-agents

为了评估这个现象，定义了一个指标路径漂移率，即在训练过程中每个回合完成后，测试存在路径漂移现象的次数占EPISODES总数的比例。表4反映了6个引导智能体在$Q1$、$Q2$和$Q3$网络中的路径漂移率，其中，A、B、C、D四点只存在于$\boldsymbol{m}_1$和$\boldsymbol{m}$中，对应$Q1$和$Q3$网络，E、F只存在于$\boldsymbol{m}_2$和$\boldsymbol{m}$中，对应$Q2$和$Q3$网络。可以看出，基于双层深度Q网络的架构大幅降低了此类现象出现的次数。

表4　$Q1$、$Q2$和$Q3$网络中引导智能体的路径漂移率

Table 4　Path wondering rate of the locations of the bootstrapping agents in three networks

网络	A	B	C	D	E	F
$Q1$	**0.1**	**0.08**	**0.08**	**0.02**	–	–
$Q2$	–	–	–	–	**0.02**	**0.06**
$Q3$	0.28	0.24	0.26	0.2	0.22	0.28

注：“–”表示在$Q1$和$Q2$中无该指标数值。

4　结　论

本文提出一种改进的双层深度Q网络架构的路径规划方法，同时提出了一种改进的K-medoids方法对智能体分组，选择引导智能体，并引导组内其他智能体疏散。实验表明，双层深度Q网络架构相比于单层深度Q网络架构，收敛速度大幅提升，并且在学习过程中更加稳定。综合考虑智能体的疏散时间和健康程度，使用指标AHEP评估智能体疏散的效果，相比传统的A-STAR、DIJKSTRA算法，提高了84%和104%；与基于火灾场景改进的扩展A-STAR和Dijkstra-ACO混合算法比较，提高了30%和21%；相较于考虑火灾影响的传统DQN方法提高了20%。此外，本文发现合适的分组可以有效提高疏散的效率和安全性，当分组为6时对应的AHEP值高于其他分组，其疏散效果最好。提出的方法在火灾场景下的疏散效果，优于大多数常用方法。

然而，提出的方法也存在一定的局限性，路径漂移现象得到改善并未完全消失。由于目标位置较远，智能体在探索未知环境的前期往往无法获得正奖励，当周围环境充满负奖励时，智能体可能会在局部位置来回徘徊，出现路径漂移现象，直到该回合结束。在以后的工作中，将考虑结合局部路径规划的方法，增加多个正奖励点，使路径漂移现象完全消失。

参考文献（References）

Ai Z H, Hu Y H, Yan F T, Zhang H J, Wang D Q, Qing S L, Zhu H H and Jia J Y. 2019. Key technology of lightweight Web3D online planning of metro fire escape. Scientia Sinica Information, 49(4): 405-421 (艾子豪, 胡永豪, 闫丰亭, 张惠娟, 王冬青, 青胜蓝, 朱合华, 贾金原. 2019. 轻量级Web3D地铁火灾逃生在线规划关键技术. 中国科学: 信息科学, 49(4): 405-421) [DOI: 10.1360/N112018-00275]

Aubé F and Shield R. 2004. Modeling the effect of leadership on crowd flow dynamics//Proceedings of the 6th International Conference on Cellular Automata. Amsterdam, the Netherlands: Springer: 601-611 [DOI: 10.1007/978-3-540-30479-1_62]

Cao X H, Li X Y, Wei X G, Li S, Huang M X and Li D L. 2020. Dynamic programming of emergency evacuation path based on Dijkstra-ACO hybrid algorithm. Journal of Electronics and Information Technology, 42(6): 1502-1509 (曹祥红, 李欣妍, 魏晓鸽, 李森, 黄梦溪, 李栋禄. 2020. 基于Dijkstra-ACO混合算法的应急疏散路径动态规划. 电子与信息学报, 42(6): 1502-1509)

[DOI: 10.11999/JEIT190854]

Cheng P J, Wu N, Meng F K and Li S. 2020. Fire escape path planning based on extended A* algorithm. Communications Technology, 53(12): 3012-3016 (程鹏举, 吴楠, 孟凡坤, 李爽. 2020. 扩展A*算法的火灾逃生路径规划研究. 通信技术, 53(12): 3012-3016) [DOI: 10.3969/j.issn.1002-0802.2020.12.021]

Fang J S, El-Tawil S and Aguirre B. 2016. Leader - follower model for agent based simulation of social collective behavior during egress. Safety Science, 83: 40-47 [DOI: 10.1016/j.ssci.2015.11.015]

Haghani M and Sarvi M. 2016. Pedestrian crowd tactical-level decision making during emergency evacuations. Journal of Advanced Transportation, 50(8): 1870-1895 [DOI: 10.1002/atr.1434]

Han Y B and Liu H. 2018. Research on route choice model based on evacuation route set and its application in crowd evacuation simulation. Chinese Journal of Computers, 41(12): 2653-2669 (韩延彬, 刘弘. 2018. 一种基于疏散路径集合的路径选择模型在人群疏散仿真中的应用研究. 计算机学报, 41(12): 2653-2669) [DOI: 10.11897/SP.J.1016.2018.02653]

Helbing D, Buzna L, Johansson A and Werner T. 2005. Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. Transportation science, 39(1): 1-24 [DOI: 10.1287/trsc.1040.0108]

Jin H L, Wang Y L, Yuan M and Chen M L. 2019. Research on escape path planning algorithm for high-rise buildings based on A*. Bulletin of Surveying and Mapping, (11): 17-21, 25 (靳海亮, 王赢乐, 袁鸣, 陈梦龙. 2019. 改进A*的高层建筑逃生路径规划算法研究. 测绘通报, (11): 17-21, 25) [DOI: 10.13474/j.cnki.11-2246.2019.0344]

Konar A, Chakraborty I G, Singh S J, Jain L C and Nagar A K. 2013. A deterministic improved Q-learning for path planning of a mobile robot. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 43(5): 1141-1153 [DOI: 10.1109/TSMCA.2012.2227719]

Le V M, Vinh H T and Zucker J D. 2017. Reinforcement learning approach for adapting complex agent-based model of evacuation to fast linear model//Proceedings of the 7th International Conference on Information Science and Technology (ICIST). Da Nang, Vietnam: IEEE: 369-375 [DOI: 10.1109/ICIST.2017.7926787]

Li X L, Kuang H and Fan Y H. 2012. Lattice hydrodynamic model of pedestrian flow considering the asymmetric effect. Communications in Nonlinear Science and Numerical Simulation, 17(3): 1258-1263 [DOI: 10.1016/j.cnsns.2011.07.034]

Li Y, Chen M Y, Dou Z, Zheng X P, Cheng Y and Mebarki A. 2019. A review of cellular automata models for crowd evacuation. Physica A: Statistical Mechanics and its Applications, 526: #120752 [DOI: 10.1016/j.physa.2019.03.117]

Li Z W, Huang H, Li N, Chu M L C and Law K. 2020. An agent-based simulator for indoor crowd evacuation considering fire impacts. Automation in Construction, 120: #103395 [DOI: 10.1016/j.autcon.2020.103395]

Lissovoi A and Witt C. 2015. Runtime analysis of ant colony optimization on dynamic shortest path problems. Theoretical Computer Science, 561, 73-85 [DOI: 10.1016/j.tcs.2014.06.035]

Liu H, Lu D J, Zhang G J, Hong X and Liu H. 2021. Recurrent emotional contagion for the crowd evacuation of a cyber-physical society. Information Sciences, 575: 155-172 [DOI: 10.1016/j. ins. 2021.06.036]

Lyu L H, Zhang S J, Ding D R and Wang Y X. 2019. Path planning via an improved DQN-based learning policy. IEEE Access, 7: 67319-67330 [DOI: 10.1109/ACCESS.2019.2918703]

Mac T T, Copot C, Tran D T and De Keyser R. 2017. A hierarchical global path planning approach for mobile robots based on multi-objective particle swarm optimization. Applied Soft Computing, 59: 68-76 [DOI: 10.1016/j.asoc.2017.05.012]

Miyagawa D and Ichinose G. 2020. Cellular automaton model with turning behavior in crowd evacuation. Physica A: Statistical Mechanics and its Applications, 549: #124376 [DOI: 10.1016/j.physa.2020.124376]

Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, Graves A, Riedmiller M, Fidjeland A K, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S and Hassabis D. 2015. Human-level control through deep reinforcement learning. Nature, 518(7540): 529-533 [DOI: 10.1038/nature14236]

Pelechano N and Badler N I. 2006. Modeling crowd and trained leader behavior during building evacuation. IEEE Computer Graphics and Applications, 26(6): 80-86 [DOI: 10.1109/MCG.2006.133]

Spartalis E, Georgoudas I G and Sirakoulis G C. 2014. CA crowd modeling for a retirement house evacuation with guidance//Proceedings of the 11th International Conference on Cellular Automata. Krakow, Poland: Springer: 481-491 [DOI: 10.1007/978-3-319-11520-7_50]

Tan L, Hu M Y and Lin H. 2015. Agent-based simulation of building evacuation: combining human behavior with predictable spatial accessibility in a fire emergency. Information Sciences, 295: 53-66 [DOI: 10.1016/j.ins.2014.09.029]

Tian Z N, Zhang G J, Hu C Y, Lu D J and Liu H. 2020. Knowledge and emotion dual-driven method for crowd evacuation. Knowledge-Based Systems, 208: #106451 [DOI: 10.1016/j. knosys. 2020.106451]

Vihas C, Georgoudas I G and Sirakoulis G C. 2012. Follow-the-leader cellular automata based model directing crowd movement//Proceedings of the 10th International Conference on Cellular Automata. Santorini Island, Greece: Springer: 752-762 [DOI: 10.1007/978-3-642-33350-7_78]

Wagner N and Agrawal V. 2014. An agent-based simulation system for concert venue crowd evacuation modeling in the presence of a fire disaster. Expert Systems with Applications, 41(6): 2807-2815 [DOI: 10.1016/j.eswa.2013.10.013]

Wang S N, Liu H, Gao K Z and Zhang J X. 2019. A multi-species artificial bee colony algorithm and its application for crowd simulation. IEEE Access, 7: 2549-2558 [DOI: 10.1109/ACCESS.2018.2886629]

Wang X L, Zheng X P and Cheng Y. 2012. Evacuation assistants: an extended model for determining effective locations and optimal numbers. Physica A: Statistical Mechanics and Its Applications, 391(6): 2245-2260 [DOI: 10.1016/j.physa.2011.11.051]

Xie W, Lee E W M, Li T, Shi M, Cao R F and Zhang Y C. 2021. A study of group effects in pedestrian crowd evacuation: experiments, modelling and simulation. Safety Science, 133: #105029 [DOI: 10.1016/j.ssci.2020.105029]

Yan F T, Hu Y H, Jia J Y, Ai Z H, Tang K, Shi Z C and Liu X. 2020. Interactive WebVR visualization for online fire evacuation training. Multimedia Tools and Applications, 79(41/42): 31541-31565 [DOI: 10.1007/s11042-020-08863-0]

Yan F T, Hu Y H, Jia J Y, Guo Q H, Zhu H H and Pan Z G. 2019b. RFES: a real-time fire evacuation system for Mobile Web3D. Frontiers of Information Technology and Electronic Engineering, 20(8): 1061-1074 [DOI: 10.1631/FITEE.1700548]

Yan F T, Jia J Y, Hu Y H, Guo Q H and Zhu H H. 2019a. Smart fire evacuation service based on internet of things computing for Web3D. Journal of Internet Technology, 20(2): 521-532 [DOI: 10.3966/160792642019032002019]

Yang Y C, Dimarogonas D V and Hu X M. 2013. Optimal leader-follower control for crowd evacuation//Proceedings of the 52nd IEEE Conference on Decision and Control. Firenze, Italy: IEEE: 2769-2774 [DOI: 10.1109/CDC.2013.6760302]

Yuan T and Liu Y. 2019. Potential energy field based pedestrian behavior model for crowd evacuation simulation in airport terminal. Journal of Physics Conference Series, 1345(4): #042023 [DOI: 10.1088/1742-6596/1345/4/042023]

Yuan W F and Tan K H. 2009. Cellular automata model for simulation of effect of guiders and visibility range. Current Applied Physics, 9(5): 1014-1023 [DOI: 10.1016/j.cap.2008.10.007]

Zhang G J, Lu D J and Liu H. 2020. Strategies to utilize the positive emotional contagion optimally in crowd evacuation. IEEE Transactions on Affective Computing, 11(4): 708-721 [DOI: 10.1109/TAFFC.2018.2836462]

Zhang H, Liu H, Qin X and Liu B X. 2018. Modified two-layer social force model for emergency earthquake evacuation. Physica A: Statistical Mechanics and Its Applications, 492: 1107-1119 [DOI: 10.1016/j.physa.2017.11.041]

Zhou M, Dong H R, Zhao Y B, Ioannou P A and Wang F Y. 2019. Optimization of crowd evacuation with leaders in urban rail transit stations. IEEE Transactions on Intelligent Transportation Systems, 20(12): 4476-4487 [DOI: 10.1109/TITS.2018.2886415]

作者简介

张晨，男，硕士研究生，主要研究方向为强化学习和机器学习。E-mail：956609626@qq.com

周文，通信作者，男，副教授，硕士生导师，主要研究方向为虚拟现实、计算机视觉、人工智能和数据可视化。E-mail：w.zhou@ahnu.edu.cn

蒋文英，女，硕士研究生，主要研究方向为强化学习和虚拟现实。E-mail：jiangwenying@ahnu.edu.cn

陈思源，男，硕士研究生，主要研究方向为目标跟踪和机器学习。E-mail：chensiyuan@ahnu.edu.cn

闫丰亭，男，博士，讲师，硕士生导师，主要研究方向为虚拟现实、计算机视觉和强化学习。E-mail：yanfengting2008@163.com