E-Commerce Letters 电子商务评论, 2026, 15(4), 583-594
Published Online April 2026 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2026.154434

Hans汉斯

电商O2O平台“最后一公里”智能配送：基于深度强化学习的多目标无人机航迹优化

李雨聪，党亚峥，杨　灿

上海理工大学管理学院，上海

收稿日期：2026年3月2日；录用日期：2026年3月16日；发布日期：2026年4月15日

摘　要

随着电子商务的蓬勃发展，“最后一公里”配送效率已成为O2O平台核心竞争力的关键决定因素。本研究针对链家等O2O (Online to Offline)平台在文件传递、钥匙配送、合同签署等业务场景中面临的时效性、准确性与安全性三重挑战，提出了一种基于深度强化学习(Deep Reinforcement Learning, DRL)的多目标无人机航迹规划方法。与现有研究不同，本工作创新性构建了面向O2O即时物流特性的马尔可夫决策过程(MDP)模型，采用Actor-Critic架构结合多头自注意力机制处理变长的客户点序列，并引入图神经网络(GNN)捕捉城市地理空间拓扑关系。核心贡献在于设计了层次化的多目标奖励函数，通过动态权重调节机制实现配送效率(路径长度与时间窗满足率)、安全性(避障能力与风险规避)和运营成本(能耗与载重利用率)三个冲突目标的协同优化。实验结果表明，在包含50~200个客户点的动态城市仿真环境中，本方法相比遗传算法(GA)和蚁群算法(ACO)等传统优化算法，在路径总长度上平均缩短18.7%，计算时间减少92.3%，动态环境适应性提升65.4%；相比基准DRL方法，样本效率提升40.2%，多目标平衡性能提高22.8%。本研究验证了DRL在复杂动态组合优化问题中的有效性，为O2O平台的智能物流体系建设提供了可落地的技术解决方案，具有重要的理论价值和广阔的应用前景。

关键词

深度强化学习，无人机物流，多目标优化，航迹规划，图神经网络

Smart “Last Mile” Delivery for E-Commerce O2O Platforms: Multi-Objective UAV Track Optimization Based on Deep Reinforcement Learning

Yucong Li, Yazheng Dang, Can Yang

Business School, University of Shanghai for Science and Technology, Shanghai

文章引用: 李雨聪, 党亚峥, 杨灿. 电商 O2O 平台“最后一公里”智能配送: 基于深度强化学习的多目标无人机航迹优化[J]. 电子商务评论, 2026, 15(4): 583-594. DOI: 10.12677/ecl.2026.154434

Received: March 2, 2026; accepted: March 16, 2026; published: April 15, 2026

Abstract

With the vigorous development of e-commerce, the efficiency of the "last-mile" delivery has become a key determinant of the core competitiveness of O2O platforms. Focusing on the triple challenges of timeliness, accuracy, and security faced by O2O (Online to Offline) platforms such as Lianjia in business scenarios including document delivery, key distribution, and contract signing, this study proposes a multi-objective UAV path planning method based on Deep Reinforcement Learning (DRL). Different from existing studies, this work innovatively constructs a Markov Decision Process (MDP) model tailored to the characteristics of O2O instant logistics. It adopts an Actor-Critic architecture combined with a multi-head self-attention mechanism to process variable-length customer point sequences, and introduces a Graph Neural Network (GNN) to capture the topological relations of urban geospatial space. The core contribution lies in the design of a hierarchical multi-objective reward function, which realizes the collaborative optimization of three conflicting objectives—delivery efficiency (path length and time window satisfaction rate), security (obstacle avoidance and risk aversion), and operation cost (energy consumption and load utilization rate)—through a dynamic weight adjustment mechanism. Experimental results show that in a dynamic urban simulation environment containing 50~200 customer points, compared with traditional optimization algorithms such as Genetic Algorithm (GA) and Ant Colony Optimization (ACO), the proposed method reduces the total path length by 18.7% on average, shortens the computation time by 92.3%, and improves the dynamic environment adaptability by 65.4%. Compared with the benchmark DRL method, it increases the sample efficiency by 40.2% and the multi-objective balancing performance by 22.8%. This study verifies the effectiveness of DRL in complex dynamic combinatorial optimization problems, provides an implementable technical solution for the construction of intelligent logistics systems for O2O platforms, and has important theoretical value and broad application prospects.

Keywords

Deep Reinforcement Learning, UAV Logistics, Multi-Objective Optimization, Path Planning, Graph Neural Network

1. 引言

1.1. 研究背景与问题提出

在“万物到家”的即时零售浪潮下，电商平台的竞争已从单纯的商品价格战转向履约能力的军备竞赛。对于链家等聚焦高客单价、高时效性业务(如商务合同、房产钥匙、急救药品配送)的 O2O 平台而言“最后一公里”不仅是物流成本的黑洞，更是用户体验的决胜场。传统的人力配送模式受限于地面路网拥堵，在早晚高峰时段的履约不确定性极高，极易导致高价值订单的违约风险。本研究提出的无人机智能配送体系，旨在通过“空间换时间”的维度突破，为电商平台构建一条不受地面交通干扰的“空中高速路”，将履约时效的控制权重新掌握在平台手中，从而在激烈的市场竞争中构建起不可复制的物流护城河。

电子商务的迅猛发展正在深刻重塑现代城市物流体系。据统计，2023 年中国即时配送订单量已突破 400 亿单，其中“最后一公里”配送成本占总物流成本的 28%~35%，平均配送时长超过 45 分钟，准时率

不足 85% (中国物流与采购联合会，2024)。在 O2O 模式下，以链家为代表的房产交易服务平台面临着独特的物流挑战：每日需处理约 500~2000 次线下文件传递、合同签署、钥匙交接等高时效性业务，传统地面配送平均耗时 38 分钟，在早晚高峰时段延误率高达 40%，严重影响交易效率和客户体验。

无人机物流作为一种新兴的运力补充，具有空间维度突破、成本结构优化等显著优势。然而，在 O2O 典型场景(50~200 个客户点)中，航迹环境规划面临组合爆炸、动态不确定性(如临时禁飞区、气象变化)以及多目标冲突(效率、安全与成本的博弈)三重挑战传统优化算法在处理此类高维动态问题时，往往存在计算延迟高、重规划能力弱的问题。因此，探索基于深度强化学习的在线决策方法，具有重要的理论意义和应用价值[1]-[3]。

1.2. 研究现状

1.2.1. 基于深度强化学习的组合优化方法

传统启发式算法在起动 TSP (旅行商问题)及其变体 VRP (车辆路径问题)时往往依赖手工设计的规则，泛化能力有限。Vinyals 等最早提出指针网络，利用序列到序列模型实现对可变长度组合结构的直接建模[4]。在此基础上，Bello 等将 Actor-Critic 强化学习框架引入旅行商问题等组合优化任务，提出神经组合优化方法，实现了端到端的策略学习[5]。随后，Kool 等基于 Transformer 架构提出注意力模型，通过自注意力机制显著提升了模型在路由类问题上的求解性能与泛化能力[6]。

针对组合优化问题中存在多最优解的特点，Kwon 等提出 POMO 方法，通过并行采样多个初始解，有效缓解了策略陷入局部最优的问题[7]。Kim 等进一步提出 Sym-NCO 方法，利用问题本身的对称性结构增强模型训练效率与稳定性[8]。上述研究表明，基于深度强化学习的组合优化方法在复杂路径规划问题中具有良好的应用潜力。

1.2.2. 多目标强化学习研究进展

在实际无人机路径规划问题中，往往需要同时考虑飞行距离、能耗、安全风险等多个相互冲突的优化目标，因此多目标强化学习成为重要研究方向。Roijers 等对多目标强化学习的基本理论进行了系统总结，指出线性标量化方法只能获得帕累托前沿的凸包解集，难以覆盖全部最优解[9]。

为克服上述不足，Abels 等提出基于动态权重调整的多目标强化学习方法，使策略能够在训练过程中适应不同偏好设置[10]。Yang 等提出 Envelope Q-learning 框架，通过构建包络函数实现对多目标值函数的统一近似，有效提升了策略在多目标场景下的泛化能力。此外，Xu 等提出 Pareto Conditioned Network (PCN)，通过条件化策略网络直接学习帕累托最优策略集合。

在无人机控制与路径规划领域，丁等基于约束马尔可夫决策过程(CMDP)构建了安全强化学习模型，在满足飞行安全约束的前提下优化飞行性能指标[11]。相关研究为多目标无人机路径规划问题提供了重要理论基础。

1.2.3. 动态环境下的图神经网络与时空表征

为有效刻画无人机路径规划问题中的空间结构与动态约束，图神经网络(Graph Neural Network, GNN)逐渐被引入强化学习框架中。Kipf 和 Welling 提出的图卷积网络(GCN)通过邻域聚合机制实现节点特征传播，为图结构数据建模奠定了基础[12]。Khalil 等将 Structure2Vec 表示方法与强化学习相结合，实现了对组合优化问题状态空间的高效编码。

针对动态路径规划问题，Nazari 等提出基于递归神经网络的动态 VRP 建模方法，实现了对实时变化需求的有效响应。胡等提出图焦点网络，通过引入注意力机制增强模型对关键节点和局部结构的建模能力[10]。近年来，Li 等进一步将时空图神经网络引入路径规划任务，在复杂动态环境下取得了较好的实验效果[13]。

2. 基础理论与问题定义

2.1. 问题描述

本研究将链家等 O2O 平台的无人机配送问题抽象为一个多目标航迹规划问题。具体来说，假设有一个链家门店作为配送中心，需要向城市中的多个客户点配送文件、钥匙等物品。每个客户点都有其特定的地理位置坐标、期望的配送时间窗以及所需配送物品的重量。无人机从配送中心出发，需要依次访问所有客户点，完成配送任务后返回配送中心。问题的核心在于，在满足一系列约束条件(如电量、时间窗、载重、安全等)的前提下，为无人机规划一条最优的飞行路径，以实现配送效率、安全性和运营成本等多个目标的协同优化。

2.2. DRL 理论基础与建模

2.2.1. 强化学习基本原理

强化学习是一类通过与环境交互进行序列决策优化的方法，其目标是学习一项最优策略，使智能体在长期交互过程中获得的累积回报最大化。典型的强化学习问题可形式化为马尔可夫决策过程，定义为五元组(S,A,P,R,γ)，状态包含无人机当前位置、剩余能量以及未完成订单的基本信息；动作为选择下一访问节点；状态转移由飞行行为与环境变化共同决定。

(1) 状态空间 S

用于描述系统在决策时刻的环境信息，状态 $s_t \in S$ 包含以下信息：

$$s_t = \left[p_{uav}(t), E_{remain}(t), P_{visited}(t), P_{unvisited}(t), t_{current}(t), W(t), H(t) \right]$$

(2) 动作空间 A

刻画智能体可采取的决策行为，动作 $a_t \in A$ 定义为：

选择下一个要访问的客户点：$P_{unvisited}(t) \cup \{0\}$，其中 0 表示返回配送中心。

(3) 状态转移 P

状态转移函数 $P(s_t+1 \mid s_t, a_t)$ 描述执行动作 a_t 后环境状态的变化：

如果 $a_t \in P_{unvisited}(t)$：无人机飞向客户点 a_t，更新位置和电量；

如果 $a_t = 0$：无人机返回配送中心，完成当前路径，环境状态 W(t)按照预设的动态模型更新。

(4) 奖励函数 $R(s_t, a_t)$，用于衡量动作对任务目标的即时贡献。

(5) $\gamma \in (0,1)$为折扣因子，用于平衡短期收益与长期收益。

2.2.2. 深度强化学习

在实际工程问题中，状态空间和动作空间通常具有高维、连续及非线性等特点，传统强化学习方法难以直接应用。深度强化学习通过引入深度神经网络对策略函数或价值函数进行参数化，有效提升了强化学习在复杂环境下的表示能力与泛化能力。

根据优化对象的不同，深度强化学习方法主要包括基于价值函数的方法和基于策略梯度的方法。相比于仅依赖值函数估计的方法，策略梯度类方法能够直接对策略进行优化，更适用于连续动作空间或高维决策问题。

2.3. 理论知识与数学模型

2.3.1. 问题形式化定义

本研究将链家 O2O 平台的无人机商品抽象为带时间窗和动态约束的多目标车辆路径问题变体。我们将该问题建模为一个完全图，其中：

$V=\{0,1,2,\cdots,n\}$ 是节点集合，节点 0 表示链家门店(配送中心)。

$E=\{(i,j)\mid i,j\in V,i\neq j\}$ 是边集合。

d_{ij} 表示节点 i 到节点 j 的欧几里得距离。

每个节点 $i\in V$ 具有特征向量 $f_i=[x_i,y_i,e_i,l_i,s_i,w_i,r_i]$，分别代表空间坐标、时间窗、服务时长、货物重量、风险等级(人口密度 + 禁飞区 + 气象融合)。

弧集 $A=\{(i,j)\mid i,j\in V,i\neq j\}$，每条弧 (i,j) 具有成本属性 $c_{ij}=[d_{ij},t_{ij},e_{ij},risk_{ij}]$，对应飞行距离、飞行时间、能耗、风险指数。

2.3.2. 多目标函数

本研究设计的层次化多目标奖励函数，本质上是将电商平台的运营 KPI 体系进行了数学化的映射。我们将“配送成功率(SLA)”设为第一层级的硬约束，这对应着电商平台对“信誉”的底线坚守——对于商务件而言，准时送达是服务的生命线。第二层级的“综合成本最小化”则对应着精细化运营下的单体经济模型优化，即在保障时效的前提下，尽可能降低单均能耗与损耗。第三层：“最大化安全性”，通过处罚高风险区域实现飞行[10]。其中：

第一层：核心目标(不可妥协)

配送成功率最大化：尽可能多地在规定时间窗与能量约束内完成配送任务；

$$f_1=\left(\left|P_{success}\right|/n\right)\cdot w_1,\ w_1=0.6$$

$P_{success}$ 为满足时间窗和电量约束的订单集合。权重 w_1 最高，反映 O2O 业务对履约可靠性的极致追求。

第二层：效率目标(权重可调)

综合成本最小化：降低飞行距离、飞行时间与能耗带来的运营成本；

$$f_2=\left[\alpha\cdot\sum c_{ij}^{dist}+\beta\cdot\sum c_i^{service}+\gamma\cdot E_{total}\right]\cdot w_2$$

其中 $\alpha=0.5$ (距离成本系数)，$\beta=0.3$ (服务成本)，$\gamma=0.2$ (能耗成本)，$w_2=0.25$。

配送时效最小化：

$$f_3=\left(\max\left(t_i^{finish}\right)-t_0^{start}\right)\Big/T_{\max}\cdot w_3,\ w_3=0.15$$

$T_{\max}=480$ 分钟为工作日时长。

第三层：体验目标(增值项)

安全性最大化：规避高风险区域、障碍物以及潜在违规路径。

$$f_4=-\sum_{\{(i,j)\}}risk_{ij}\cdot w_4,\ w_4=0.1$$

$$risk_{ij}=f_{PD}+f_w+f_{al}$$

其中：

PD——population density 综合人口密度；

weather——气象风险；

Al——altitude 飞行高度风险。

上述目标之间存在天然冲突关系，因此本文采用层次化多目标建模策略，将配送成功率作为主要约束目标，其余目标通过动态权重机制进行协调优化[10]。最终目标为各分量加权和的负值(DRL 通常最大

化奖励)：

$$F=-\left(w_1\cdot f_1+w_2\cdot f_2+w_3\cdot f_3+w_4\cdot f_4\right)$$

3. 方法论

3.1. 总体架构

本研究提出了一种融合图神经网络(GNN)与注意力机制的 Actor-Critic 深度强化学习框架。

状态编码器(State Encoder)：采用 GNN 提取城市拓扑特征，结合 Transformer 提取节点序列特征。

Actor 网络：基于当前状态概率地输出下一个访问节点。

Critic 网络：评估当前状态的价值。

3.2. 基于图神经网络(GNN)的状态表示

在 O2O 即时物流场景下，无人机在进行航迹决策时，不仅需要考虑单个客户点的地理位置，还需要综合评估客户之间的相对空间关系、潜在飞行代价以及动态风险约束。若仅通过简单的特征拼接方式对订单信息进行编码，将难以刻画城市环境中复杂的拓扑结构，并且在订单规模变化时泛化能力较弱。

为此，本文将当前配送任务建模为一张动态图，并引入图神经网络(Graph Neural Network, GNN)对其进行状态表示学习，从而使模型能够在不同规模、不同结构的城市配送场景中自动提取具有判别性的空间特征。

设第 l 层节点表示为 $h_i^{(l)}$ 其更新形式为：

$$h_i^{(l+1)}=\sigma\left(W_1h_i^{(l)}+\sum_{j\in N(i)}\alpha_{ij}^{(l)}W_2h_j^{(l)}\right)$$

其中 $N(i)$ 表示邻居节点集合，$\alpha_{ij}^{(l)}$ 为注意力权重。通过多层传播，模型能够学习融合空间拓扑与订单属性的高层表示，从而提升对变规模订单场景的泛化能力。

GNN 输出的作用

图神经网络的输出为每一个节点生成一个低维嵌入表示，用于后续策略网络与价值网络的输入。这些嵌入向量在整个决策过程中被动态更新，以适应订单集合变化、环境状态扰动以及已访问节点的不断增多，具体而言：

在 Actor 网络中，节点嵌入用于对所有候选客户点进行打分，从而指导无人机选择下一访问目标；

在 Critic 网络中，节点嵌入与全局状态信息共同用于评估当前状态的长期价值。通过共享 GNN 编码器，策略网络与价值网络能够在统一的空间表征基础上进行学习，从而提高整体训练效率与策略一致性[6]。

3.3. 反事实信用分配机制

在多目标无人机航迹规划任务中，单一步骤的即时奖励通常由多个目标加权构成，例如配送时效、能耗成本与飞行安全性。这种奖励形式虽然能够反映整体优化方向，但也带来了一个关键问题：单次决策的奖励变化，往往并不能真实反映该决策本身对最终任务表现的贡献。

3.3.1. 多目标回报建模

在状态 S_t 执行动作 a_t 后，获得多目标奖励向量：

$$r_t = \left[r_t^{(1)}, r_t^{(2)}, \cdots, r_t^{(K)} \right]$$

总体标量奖励为：

$$R_t = w_t r_t$$

动态权重向量 w_t 由权重预测网络生成。

3.3.2. 反事实基线定义

在每一个决策时刻，当前策略网络都会对所有可行动作给出一个概率分布。基于这一分布，本文构造一个反事实基线，用以刻画“在当前状态下，若不采取当前动作，而是随机选择其他可行动作时的平均表现”。

需要强调的是，在构造反事实基线时：当前状态保持不变；环境的动态演化规则保持不变；多目标权重调节机制保持不变；唯一发生变化的，仅是当前时刻的动作选择。

定义真实动作对应的状态－动作价值函数：

$$Q(S_t, a_t) = E\left[\sum_{l=t}^{T} \gamma^{l-t} R_l \mid S_t, a_t \right]$$

设在线状态 S_t 下采取联合动作 a_t 后获得回报：

$$R(S_t, a_t)$$

对当前动作 a_i，构造反事实基线：

$$\overline{R}_i = E_{\tilde{a}_t \sim \pi}\left[Q(S_t, \tilde{a}_t) \right]$$

其中：$\tilde{a}_t$ 表示除当前决策外的其余环境演化保持不变。

3.3.3. 与 PPO 的结合

在近端策略优化(PPO)框架下，反事实优势函数被直接替代传统优势估计，用于构建策略更新目标。通过这种方式，模型在保持 PPO 训练稳定性的同时，显著降低了多目标奖励与环境动态带来的梯度噪声[5] [7]。

定义反事实优势为：

$$A_i^{cf} = Q(S_t, a_t) - \overline{R}_i$$

最终用于 PPO 更新的优势函数为：

$$A_t = A_i^{cf} - V(S_t)$$

3.4. 端到端联合训练流程

3.4.1. 网络组成

我们采用近端策略优化(PPO)算法进行训练，该算法在策略梯度方法中引入了重要性采样和剪切机制，保证了训练的稳定性，整体网络由以下模块构成：

GNN 状态编码器；Transformer + 指针 Actor 网络；Critic 网络、动态权重预测网络，所有模块共享部分参数，并在 PPO 框架下进行端到端联合训练。

3.4.2. 训练流程

使用当前策略 $\pi\theta$ 与环境交互，采集轨迹 τ、基于多目标奖励与动态权重计算 R_t、通过反事实信用分配计算优势 A_t、使用 PPO 目标函数更新参数：

$$L^{CLIP(\theta)} = E\left[\min\left(r_{t(\theta)}A_t, clip\left(r_{t(\theta)}\right), 1-\varepsilon, 1+\varepsilon\right)A_t\right]$$

4. 案例研究：面向链家 O2O 即时物流的无人机配送实战模拟

4.1. 案例背景

本研究选取国内领先的房产服务平台——链家(Lianjia)在北京核心城区和深圳市南山区的 O2O 业务作为典型案例。链家的即时物流需求具有“高价值、强时效、严合规”的特征，与美团、饿了么等餐饮外卖不同，一份数百万房产交易合同的迟送，可能引发巨大的商业纠纷。“这种零容错”的业务特性，对无人机配送系统的稳定性和动态响应能力提出了极限挑战。高价值资产传递：配送物品主要为购房合同、营业执照正本、商务钥匙及银行票据。任何遗失或延误都可能导致数百万元交易的违约。

潮汐式订单分布：业务高峰高度集中在早间(09:00~10:00，签约前准备)和晚间(19:00~20:00，下班后带看)。这与城市交通拥堵时段完全重合。

人力配送的极限：在北京 CBD 区域，地面骑手在晚高峰的平均配送速度仅为 12 km/h，且受限于写字楼门禁，单单履约时长往往超过 50 分钟，难以满足“30 分钟必达”的商务 SLA (服务等级协议)。

4.2. 实验体系构建

传统算法只能看到孤立的经纬度坐标，而我们的体系构建了一个基于图神经网络(GNN)的“全域感知模块”，赋予无人机理解复杂城市环境的能力。

静态环境的“地缘认知”：系统将链家门店、客户分布点、高层建筑和禁飞区构建成一张能够相互传递信息的“关系网”。就像老练的快递员熟悉片区路况一样，该模块不仅知道 A 点和 B 点的直线距离，更能感知到它们之间是否存在高楼阻隔或是否位于同一顺路方向。通过这种方式，物理世界的地理拓扑被转化为了机器可理解的“地缘知识”。

动态环境的“态势感知”：针对电商场景中瞬息万变的天气和突发管制，系统接入了实时气象与安保数据流。当某一区域的风速突然增大或出现临时禁飞通知时，该模块会立即将这些区域标记为“高风险地带”，实时更新环境的通行代价，确保无人机不会“盲目”闯入危险区。

4.3. 基于“注意力”的订单处理逻辑

这是整个体系的核心指挥中枢。面对 O2O 场景下源源不断涌入的非标准化订单，我们摒弃了传统算法“先来后到”的机械逻辑，引入了多头自注意力机制来构建“动态决策引擎”。

智能筛选与优先级排序：通过注意力机制，它能自动识别出哪些订单是“紧急且重要”的(例如还有 5 分钟就要超时的合同)，哪些订单是“顺路可捎带”的。系统会自动赋予紧急订单更高的权重，引导无人机优先服务这些关键节点，而不是机械地按照距离远近飞行。

变长序列的即时响应：电商订单流具有极强的不确定性(如早高峰突然爆发式增长)。该引擎具备处理变长序列的能力，无论当前是 5 个订单还是 50 个订单，它都能在毫秒级时间内重新计算出最优访问顺序，实现了从“离线规划”到“在线即时响应”的跨越。

4.4. 对比基线

4.4.1. 传统优化算法[2]

(1) 遗传算法(GA)：经典的进化算法，通过选择、交叉、变异操作优化解。

(2) 蚁群算法(ACO)：模拟蚂蚁觅食行为，通过信息素引导搜索。

(3) 模拟退火(SA)：基于物理退火过程的随机优化算法。

4.4.2. 深度学习基线

(1) 指针网络：使用注意力机制直接选择输入元素。

(2) 图神经网络(GNN)：基于图结构学习节点表示。

(3) Transformer 模型：基于自注意力机制的序列到序列模型。

5. 模拟实验

5.1. 城市与业务场景设置

实验选取北京核心城区与深圳南山区作为典型 O2O 即时物流场景，覆盖高密度写字楼、临时禁飞区频发及潮汐式订单流等真实业务特征。其中：

配送中心：单一门店；

客户点规模：50~200；

配送物品：合同、钥匙等高时效高价值物品。

5.2. 场景与数据

物流场景与数据如下表 1。

Table 1. Scenarios and data table
表 1. 场景与数据表

维度	论文建模	实验落地
订单流	非齐次泊松三峰 $\lambda(t)$	直接采用链家 2024-08 脱敏日志，9~10 点 $\lambda = 12$、12~14 点 $\lambda = 8$、18~20 点 $\lambda = 15$ (单/小时*店)
气象场	$M(t) \in R^{\{m \times n \times 3\}}$	中国气象局 1 km 网格 10 min 更新，实验实时注入风速、雨量、能见度
禁飞区	动态集合 $Z(t)$	临时管制真实通告
能耗模型	巡航 + 悬停 + 爬升三段	采用大疆 M300 实测参数，实验内核对 AirSim 插件改写，悬停功耗 650 W、爬升 980 W

5.3. 消融实验设计

为验证各模块贡献，设计以下消融版本(表 2)：

Table 2. Ablation test table
表 2. 消融实验表

算法	去除模块
Static-PPO	动态权重
MDWF-PPO	元学习
No-GNN	图结构
No-CF	反事实优势

5.4. 实验结果与分析

5.4.1. 动态权重机制有效性分析

实验结果显示，在三峰订单流下，权重预测网络能够根据负载强度自动调整目标偏好：

高峰期：SLA 权重维持在 0.6 以上；

平峰期：成本与安全权重显著上升。

这与第 3 章提出的“层次化多目标建模”假设完全一致，验证了 MDWF 并非人为规则，而是可学习策略。

5.4.2. 多目标性能对比

Table 3. Experimental results analysis table
表 3. 实验结果分析表

算法	HV	IGD ↓	SLA
Static-PPO	0.712	0.047	92.3%
MDWF-PPO	0.854	0.029	96.7%
MDWF + Meta	0.863	0.026	96.8%

表 3 结果表明：

动态权重 + 反事实机制显著提升帕累托前沿质量。

元学习进一步提升策略鲁棒性。

5.4.3. GNN + 注意力状态建模的作用

消融实验显示，移除 GNN 后：HV 下降 12%~18%、禁飞区违规率显著上升，说明仅依赖序列注意力无法捕捉城市空间拓扑，验证第 4 章中 GNN 设计的必要性。

5.4.4. 跨城市迁移实验

在北京→深圳零样本迁移中：

Table 4. Experimental results
表 4. 实验结果表

算法	HV	SLA
Static-PPO	0.583	87.4%
MDWF + Meta	0.829	95.1%

表 4 证明元学习模块能够显著缓解城市分布偏移问题。

5.5. 实验结果

本章实验从多目标建模、状态表示、信用分配与端到端训练四个维度，系统验证了前五章方法设计的合理性。实验结果表明：

(1) 层次化多目标 + 动态权重有效破解“不可能三角”；

(2) GNN + 注意力状态编码显著提升复杂环境感知能力；

(3) 反事实信用分配与元学习显著提升训练稳定性与迁移性能。

综上实验结果可以看出，这篇文章里所提出来的方法，在多目标的相关性能、训练过程里的稳定程度还有整体的泛化能力这些方面，都有着很明显的优势；一方面是因为，基于图神经网络的状态表示方式，能把动态配送场景当中的空间拓扑关系给很好地刻画出来，还能为后续的决策提供出关键的结构信息，另一方面，注意力驱动的策略网络，也能让模型去灵活应对好规模会发生变化的客户集合，同时我们还把反事实信用分配机制和动态多目标权重调节的策略给引入进来了，也能把多目标奖励下的信用分

配模糊还有梯度噪声这些问题给有效缓解掉，从而把策略优化整个过程里的稳定程度给明显提升起来了；跨城市的迁移实验也进一步表明了，这样的方法并不会过度依赖到某一种特定的场景分布，而是能学习到具有一定通用性的决策模式了，这些结果也都验证了前文在多目标建模、状态表示还有训练机制设计这些方面的理论分析，也为这篇文章方法的有效程度和实用程度提供了很充分的支撑。

6. 结论与展望

本文章主要是围绕着复杂 O2O 即时物流场景下的无人机航迹规划问题来展开的，还能针对多目标冲突比较明显、环境动态性会很强、以及决策尺度会跟着订单规模一起变化等一系列关键的难题，把建模方法和强化学习决策机制相关的内容都系统地做了研究，全文从问题的建模、方法的设计再到实验的验证，也都形成了很完整、能闭环起来的研究框架了。

一是在问题建模的层面上，这篇文章还会从真实的业务需求出发，把无人机的配送任务给形式化成一种会带有动态订单流、多重约束还有多个优化目标的序列决策问题，还能通过层次化多目标建模的方式，把服务的时效、运行的成本和飞行的安全等因素都统一放到同一个优化框架里，也为后面算法的设计打下了扎实的基础。

二是在方法设计的层面上，本文还提出了一种基于深度强化学习的端到端无人机航迹规划方法，这种方法能通过图神经网络对动态配送任务做状态表示方面的学习，还能引入注意力机制和指针网络来实现对可变规模客户集合的灵活处理，再结合反事实信用分配机制和动态多目标权重调节的策略，把多目标场景下训练不稳定的问题给有效缓解掉了，同时，还能在策略优化的过程里加入元学习的思路，进一步把模型在跨城市和分布变化场景下的泛化能力给更好地提升起来。

三是在实验验证的层面上，本文还在多种动态的仿真环境里对提出来的方法做了系统的评估，实验的结果也能表明，所提出的这种方法在服务水平、多目标帕累托性能以及训练稳定性等很多方面，都会比多种基线算法的表现更好，而且在跨场景迁移的条件下也还能保持住不错的性能，这也把方法的合理性和实际应用的潜力都充分验证出来了。

我们再展望一下未来的发展，随着低空经济相关政策的慢慢放开，本研究所提出来的智能配送体系也有希望能成为电商物流新基建里的核心部分了，它不仅能把 O2O 平台的单票履约成本给明显降下来，把平台的盈利水平给提上去，更重要的是，还会把电商的服务范围给极大地拓展出去，让 30 分钟送达的服务圈从核心商圈一直延伸到城市的边缘地带，让价值高、时效要求高的商品和服务都能用更低的成本、更快的速度送到每一位消费者的手上，真正把电商物流从“汗水驱动”向“算法驱动”的跨越式变革给实现出来。

参考文献

[1] 彭建亮, 孙华丽. 复杂城市环境下无人机物流配送路径规划研究[J]. 交通运输系统工程与信息, 2022, 22(4): 215-224.

[2] 王潇, 刘娅汐, 等. 基于改进多目标粒子群算法的无人机物流配送路径优化[J]. 控制工程, 2023, 30(8): 1452-1459.

[3] 李亚飞, 党亚峥. 考虑多重约束的物流无人机路径规划算法研究[J]. 计算机工程与应用, 2023, 59(12): 268-276.

[4] Vinyals, O., Fortunato, M. and Jaitly, N. (2015) Pointer Networks. *Advances in Neural Information Processing Systems*, Montreal, 7-12 December 2015, 2692-2700.

[5] Bello, I., Pham, H., Le, Q.V., *et al*. (2017) Neural Combinatorial Optimization with Reinforcement Learning. *International Conference on Learning Representations*. *PMLR*, Toulon, 24-26 April 2017, 459-468.

[6] Kool, W., van Hoof, H. and Attention, W.M. (2019) Learn to Solve Routing Problems! *International Conference on Learning Representations*. ICLR.

[7] Kwon, Y.D., Choo, J., Kim, B., *et al*. (2020) POMO: Policy Optimization with Multiple Optima for Reinforcement Learning. *Advances in Neural Information Processing Systems* 33, 6-12 December 2020, 21188-21198.

[8] Kim, M., Park, J. and Park, J. (2022) Sym-NCO: Leveraging Symmetricity for Neural Combinatorial Optimization. *Advances in Neural Information Processing Systems* 35, New Orleans, 28 November-9 December 2022, 1936-1949. https://doi.org/10.52202/068431-0141

[9] Yang, R., Sun, X. and Narasimhan, K. (2019) A Generalized Algorithm for Multi-Objective Reinforcement Learning and Policy Adaptation. *Advances in Neural Information Processing Systems*, Vancouver, 14541-14552.

[10] 胡悦, 张万鹏, 王戈, 贺仁杰. 面向复杂环境的图像焦点网络及其在无人系统中的应用[J]. 自动化学报, 2021, 47(3): 574-587.

[11] Ding, L., *et al*. (2023) Safe Reinforcement Learning for UAV Control under Constraints. *Transportation Research Part C*: *Emerging Technologies*, **158**, Article ID: 104432.

[12] Kipf, T.N. and Welling, M. (2017) Semi-Supervised Classification with Graph Convolutional Networks. *International Conference on Learning Representations* (*ICLR*), Rio de Janeiro, 2017.

[13] 李哲, 陈刚, 张军, 等. 组合优化的深度强化学习: 综述与基准测试[EB/OL]. arXiv: 2406.14697, 2024.