Advances in Applied Mathematics 应用数学进展, 2026, 15(11), 223-235
Published Online November 2026 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2026.1511445

Hans汉斯

面向路径规划的改进图注意力强化学习算法

苏继满[1]，刘志刚[1]，焦玉全[2]

[1]南通科技职业学院能源交通学院，江苏 南通
[2]南通科技职业学院智能制造学院，江苏 南通

收稿日期：2026年10月17日；录用日期：2026年11月12日；发布日期：2026年11月18日

摘 要

节点遍历类路径规划广泛存在于智能制造、物流配送与交通运维等场景，需在约束下确定总通行代价最小的节点访问次序，属NP难组合优化问题。针对现有深度强化学习方法在几何特征刻画、拓扑感知与训练稳定性方面的不足，本文提出距离感知图注意力策略网络(DGAPN)：以几何边编码将通行代价显式注入注意力，用门控残差块抑制深层图网络的特征趋同，经双路池化与门控读出形成全局图嵌入，并采用课程式近端策略优化将训练分为探索与精修两段，配合自适应优势与策略平滑。以旅行商问题(TSP)为载体的测试表明，DGAPN在路径长度、跨规模与跨分布泛化及推理效率上均优于多数同类方法，可作为此类路径规划问题的通用框架。

关键词

路径规划，深度强化学习，几何边编码，门控残差块，课程式近端策略优化

Improved Graph Attention Reinforcement Learning Algorithm for Path Planning

Jiman Su[1], Zhigang Liu[1], Yuquan Jiao[2]

[1]School of Energy and Transportation, Nantong College of Science and Technology, Nantong Jiangsu
[2]School of Intelligent Manufacturing, Nantong College of Science and Technology, Nantong Jiangsu

Received: October 17, 2026; accepted: November 12, 2026; published: November 18, 2026

Abstract

Node-traversal path planning, common in intelligent manufacturing, logistics delivery and traffic operation and maintenance, seeks the minimum-cost node order under constraints, an NP-hard combinatorial optimization problem. To address weaknesses of deep reinforcement learning (DRL) in geometric depiction, topology perception and stability, this paper proposes a Distance-aware

文章引用: 苏继满, 刘志刚, 焦玉全. 面向路径规划的改进图注意力强化学习算法[J]. 应用数学进展, 2026, 15(11): 223-235. DOI: 10.12677/aam.2026.1511445

Graph Attention Policy Network (DGAPN). Geometric edge encoding injects travel cost into attention, a gated residual block avoids feature homogenization in deep graph networks, dual-branch pooling with gated readout yields a global graph embedding, and a curriculum proximal policy optimization splits training into two stages of exploration and refinement, with adaptive advantage and policy smoothing. Tests on the traveling salesman problem (TSP) show that DGAPN beats most DRL baselines in tour length and in cross-scale and cross-distribution generalization and inference efficiency, providing a general framework for such path planning problems.

Keywords

Path Planning, Deep Reinforcement Learning, Geometric Edge Encoding, Gated Residual Block, Curriculum Proximal Policy Optimization

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

在智能制造车间调度、城市物流配送、无人系统航迹规划以及交通运维巡检等工程活动中，普遍需要为一组分散的目标节点合理安排访问先后，以便在完成全部任务的同时把距离、时间或能耗等累计代价控制到最低。这类节点遍历类路径规划在数学上可归结为约束条件下的组合优化问题，且多数情形具有 NP 难特性[1]：可行访问次序的数量随节点数增长而急剧膨胀，难以通过穷举获得最优方案。旅行商问题(TSP)是其中最具代表性的抽象模型，它以节点为顶点、以节点间通行代价为边权、以闭合回路总长最短为目标，其图结构与求解逻辑对各类工程路径规划具有很强的普适性。随着节点规模扩大与运行约束增多，传统方法越来越难以同时兼顾解的质量与响应速度，因此发展高效、可泛化的新型求解途径具有明显的理论意义与工程价值[2]。

围绕 TSP 与一般路径规划，传统求解途径大体可分为精确方法、近似方法与启发式方法三类。精确方法能够给出可证明的全局最优解，其中以分支定界为核心的 Concorde 被公认为 TSP 的标杆求解器[3]，但其计算耗时随规模近似指数上升、资源占用大，难以满足大规模或在线动态场景的需要。启发式方法依据领域规则在解空间中搜索，可在有限时间内给出可行方案，例如 Shi 等[4]将粒子群优化用于 TSP，并借助不确定搜索策略与交叉消除技术加快收敛；然而此类方法高度依赖先验知识与人工调参，普适性有限，且容易过早收敛到局部最优，解的优劣受初始解影响较大。整体来看，传统方法长期受困于“求优”与“求快”难以兼顾，在规模扩大时泛化能力不足，这促使研究者把目光转向深度强化学习。

深度强化学习兼具特征提取与复杂模式拟合能力，为路径规划提供了新的思路，已有工作大致可划分为基于序列的方法与基于图神经网络的方法两类。基于序列的方法把路径生成视为一维序列输出：Vinyals 等[5]提出指针网络 PtrNet，以编码器–解码器结构直接构造组合优化问题的解，在小规模上验证了可行性，但规模增大后泛化明显变差；Bello 等[6]率先用强化学习训练神经组合优化模型，确立了以学习方式求解组合优化的基本训练范式；Kool 等[7]基于 Transformer 构建了“重型编码器加轻型解码器”的注意力模型 AM，在百节点以内表现突出并成为常用基线；Kwon 等[8]进一步提出基于多最优策略的 POMO，增强了推理能力，但在大规模下多策略并行探索带来大量冗余计算。此类序列模型本质上把二维拓扑压成一维序列，忽视了路径规划固有的空间结构先验。

基于图神经网络(GNN)的方法则直接面向问题的图结构进行建模。Dai 等[9]较早给出了在图上学习

组合优化策略的框架，说明了图网络对该类问题的适配性；Joshi 等[10]用图卷积网络求解 TSP，提高了图结构问题的求解效率；Pan 等[11]尝试以分层策略处理大规模 TSP，体现出图网络的结构潜力。但现有图方法仍存在两个突出短板：其一是特征过度平滑，网络层数增加后节点表示趋于同质、区分度下降；其二是几何感知不足，Fountoulakis 等[12]在对图注意力机制的回顾中指出，现有注意力对边所对应的实际空间距离缺乏显式刻画，权重分配与真实几何相脱节，容易产生相互交叉的次优路径，这一问题在高度依赖空间关系的路径规划中尤为明显。

针对上述不足，本文面向节点遍历类路径规划，提出距离感知图注意力策略网络(Distance-aware Graph Attention Policy Network, DGAPN)，主要工作包括：① 在节点表征层面，设计几何边编码与门控残差更新，将通行代价作为空间先验引入邻域消息传递，并以门控方式调节跨层信息流，在增强几何敏感性的同时缓解深层特征趋同；② 在整体结构层面，构建双路池化与门控读出相结合的全局图嵌入模块，并行刻画节点的显著特征与整体分布，为决策提供宏观视野，减少因短视而产生的交叉路径；③在策略训练层面，提出课程式近端策略优化，把训练组织为探索课程段与精修课程段，结合自适应优势函数与策略平滑更新，在大规模解空间中平衡探索与利用，提升训练稳定性与收敛效率。

2. 路径规划求解框架

为契合节点遍历类路径规划的图结构特性与复杂解空间，本文提出 DGAPN，以深度强化学习实现端到端的路径构造。整体采用自回归编码器–解码器范式：编码器负责抽取高阶节点特征与整体拓扑信息，解码器据此逐步给出节点决策，并通过课程式近端策略优化完成稳定训练，从而在特征刻画、结构感知与策略寻优三个方面协同改善求解效果。

2.1. 问题形式化与马尔可夫决策建模

节点遍历类路径规划可统一表述为：给定无向完全图 $G=(V,E)$，顶点集 $V=\{x_1,x_2,\cdots,x_m\}$ 对应 m 个待访问目标，边集 E 中每条边的权值 d_{ij} 表示由节点 i 到节点 j 的通行代价(可为距离、时间或能耗等)。优化目标是得到一条恰好访问每个节点一次并返回起点的闭合回路，使累计通行代价最小。本文以二维欧氏 TSP 作为标准验证载体，取平面欧氏距离为边权，记访问序列为 $\pi=(\pi_1,\pi_2,\cdots,\pi_m)$，并令 $\pi_{m+1}=\pi_1$，则回路总长度 $L(\pi)$ 定义为：

$$L(\pi)=\sum_{t=1}^{m-1}\left\|x_{\pi_t}-x_{\pi_{t+1}}\right\|_2+\left\|x_{\pi_m}-x_{\pi_1}\right\|_2 \tag{1}$$

参照序列决策问题的一般建模方式[5]，本文将路径生成刻画为马尔可夫决策过程。设当前图的整体状态为 s，参数为 θ 的策略给出完整回路的概率为 $p_\theta(\pi\mid s)$。依据概率链式法则，可将其分解为各步条件概率之积；在每个决策步，模型结合当前状态与候选节点特征计算适配度，再经 Softmax 归一化为未访问节点上的概率分布，如式(2)所示：

$$\begin{aligned} p_\theta(\pi\mid s)&=\prod_{t=1}^{m}p_\theta(\pi_t\mid\pi_{<t},s)\\ p_\theta(\pi_t=h_i\mid\pi_{<t},s)&=\frac{\exp\left(f(s,h_i)\right)}{\sum_{j\notin\pi_{<t}}\exp\left(f(s,h_j)\right)} \end{aligned} \tag{2}$$

其中 $\pi_{<t}$ 表示第 t 步之前已访问的节点序列，$f(s,h_i)$ 为候选节点 i 的适配度打分。该自回归建模无需一次性给出整条回路，可随决策过程动态评估已访问状态与剩余节点之间的关系，因而能够灵活适配不同节点规模的路径规划场景。

2.2. DGAPN 总体框架

DGAPN 在逻辑上由“特征嵌入–图编码–自回归解码–课程式训练”四部分组成，如图 1 所示。与

基于序列的模型[5]-[8]相比，它充分利用图结构先验，以图注意力完成节点间的信息交互；与已有图网络方法[10] [11]相比，它显式引入边的几何信息并加入门控机制，从而缓解特征过度平滑与几何感知不足[12]。各部分功能为：

(1) 特征嵌入：把节点坐标与节点间通行代价分别映射为初始节点嵌入与初始边嵌入，并经批量归一化稳定特征分布；

(2) 图编码：先在邻域消息传递段结合几何边编码计算距离感知注意力，再由门控残差块过滤噪声、抑制特征趋同；随后在表征精化段经位置感知特征变换增强表达，并由全局图读出器生成全局图嵌入；

(3) 自回归解码：采用上下文－指针两级结构，先由多头上下文注意力刻画路径状态，再由指针式单头注意力给出节点兼容性与选择概率，并通过掩码保证访问合法性；

(4) 课程式训练：以课程式近端策略优化(C-PPO)在不同课程段动态调节探索强度与优化粒度，实现大规模解空间下的高效稳定训练。



Figure 1. Overall framework of DGAPN
图 1. DGAPN 总体框架

2.3. 几何边编码与初始嵌入

特征嵌入层先对节点坐标与边通行代价分别施加线性映射，得到初始节点嵌入 h_i^0 与初始边嵌入 e_{ij}^0，并通过批量归一化 BN(·)稳定特征尺度、加快收敛：

$$h_i^0 = \mathrm{BN}\left(W_x x_i + b_x\right) \tag{3}$$

$$e_{ij}^0 = \mathrm{BN}\left(W_d d_{ij} + b_d\right) \tag{4}$$

其中 $d_{ij}=\left\|x_i-x_j\right\|_2$ 为节点 i、j 间的原始通行代价，W_x、W_d 与 b_x、b_d 分别为可学习的权重矩阵与偏置。与仅使用节点坐标的做法不同，本文把通行代价作为独立的几何边特征保留下来，构成后续距离感知注意力的输入，即几何边编码(GEE)。

2.4. 邻域消息传递编码器

邻域消息传递段通过融合几何边编码的注意力实现节点间的局部信息交互。针对传统图注意力忽略边几何信息的缺陷[12]，本文在打分时显式引入初始边嵌入 e_{ij}^0，使其充当信息传播的“空间阻尼”：空间上越接近的节点越容易获得高权重，从而增强节点表示的几何敏感性。对第 ℓ 层，先拼接目标节点 i、邻居节点 j 的投影特征与边嵌入，经 LeakyReLU 得到原始得分 $s_{ij}^{(\ell)}$，再在邻域 $N(i)$内归一化为注意力系数 $\alpha_{ij}^{(\ell)}$：

$$\begin{aligned} s_{ij}^{(\ell)} &= \mathrm{LeakyReLU}\left(a^{\mathrm{T}}\left[W_\ell h_i \,\|\, W_\ell h_j \,\|\, e_{ij}^0\right]\right) \\ \alpha_{ij}^{(\ell)} &= \frac{\exp\left(s_{ij}^{(\ell)}\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(s_{ik}^{(\ell)}\right)} \end{aligned} \tag{5}$$

其中 $\|$ 表示向量拼接，a^{T} 为可学习的注意力向量，W_ℓ 为第 ℓ 层的投影矩阵。在此基础上，通过多头机制在不同子空间并行聚合邻域消息，形成候选表示 m_i：

$$m_i=\Big\|_{k=1}^{K}\left(\sum_{j\in\mathcal{N}(i)}\alpha_{ij}^{(\ell,k)}W_V^{(k)}h_j\right) \tag{6}$$

其中 K 为注意力头数，$W_V^{(k)}$ 为第 k 个头的值投影，$\alpha_{ij}^{(\ell,k)}$ 为该头对应的注意力系数。为抑制深层图网络的特征过度平滑，进一步引入门控残差块(GRB)，借助重置门与更新门自适应地调节跨层信息流，在保留有效特征的同时过滤噪声：

$$\begin{aligned} r_i &= \sigma\left(W_r m_i + U_r h_i\right) \\ z_i &= \sigma\left(W_z m_i + U_z h_i\right) \\ \tilde{h}_i &= \tanh\left(W m_i + U\left(r_i \odot h_i\right)\right) \\ h_i' &= \left(1-z_i\right)\odot h_i + z_i \odot \tilde{h}_i \end{aligned} \tag{7}$$

其中 r_i、z_i 分别为重置门与更新门，$\sigma(\cdot)$为 Sigmoid 函数，$\odot$ 为逐元素 Hadamard 乘积，W_r、U_r、W_z、U_z、W、U 均为可学习参数。该块把节点原有表示 h_i 与候选表示 m_i 自适应融合，使网络能够在层数加深时仍保持节点特征的区分度。

2.5. 表征精化与全局图读出

该段先由位置感知特征变换层(PFT)增强节点表示的非线性表达，再由全局图读出器(GGR)构建整体图嵌入。PFT 由两层线性变换与 ReLU 构成，通过非线性空间映射提升特征对复杂路径规划场景的适应能力：

$$h_i''=W_2\mathrm{ReLU}\left(W_3h_i'+b_3\right)+b_2 \tag{8}$$

其中 W_2、W_3 为权重矩阵，b_2、b_3 为偏置。为获得整体拓扑视野、避免短视决策造成的交叉路径，GGR 先对精化后的节点表示并行执行最大池化与平均池化，分别捕捉显著特征与整体分布：

$$\begin{aligned} h_{\max} &= \max_i\left(h_i''\right) \\ h_{\mathrm{mean}} &= \frac{1}{m}\sum_{i=1}^{m}h_i'' \end{aligned} \tag{9}$$

进而采用类 LSTM 的门控机制自适应融合两种池化结果，得到全局图嵌入 g：

$$\begin{aligned} f_t &= \sigma\left(W_f h_{\max} + U_f h_{\text{mean}}\right) \\ i_t &= \sigma\left(W_i h_{\max} + U_i h_{\text{mean}}\right) \\ o_t &= \sigma\left(W_o h_{\max} + U_o h_{\text{mean}}\right) \\ c_t &= f_t \odot h_{\text{mean}} + i_t \odot \tanh\left(W_c h_{\max}\right) \\ g &= o_t \odot \tanh\left(c_t\right) \end{aligned} \tag{10}$$

其中 f_t 为遗忘门、控制历史信息的保留，i_t 为输入门、控制当前特征的写入，o_t 为输出门、控制最终嵌入的输出，c_t 为中间记忆状态；W_f、W_i、W_c、W_o 与对应的 U_f、U_i、U_o 为各门控的可学习参数。全局图嵌入 g 与节点表示一同送入解码器。

2.6. 上下文－指针两级解码器

解码器采用上下文－指针两级结构(HCPD)，依据编码器给出的节点表示与全局图嵌入逐步选择下一节点。初始上下文由全局图嵌入、起始节点与上一访问节点拼接而成，在首个决策步尚无已访问节点时，相应位置使用可学习的占位向量 v_0：

$$c_0 = \begin{cases} \left[g \,\|\, h_{\pi_1} \,\|\, h_{\pi_{t-1}}\right], & t > 1 \\ \left[g \,\|\, v_0 \,\|\, v_0\right], & t = 1 \end{cases} \tag{11}$$

第一级为多头上下文注意力：以初始上下文 c_0 生成查询，以全部节点表示构成的矩阵 H 生成键与值，经多头注意力得到刻画当前路径状态的上下文向量 c_t：

$$\begin{aligned} Q_k &= W_Q^{(k)} c_0, K_k = W_K^{(k)} H, V_k = W_V^{(k)} H \\ c_t &= \|_{k=1}^{K} \operatorname{softmax}\left(\frac{Q_k K_k^{\mathsf{T}}}{\sqrt{d_k}}\right) V_k W_O \end{aligned} \tag{12}$$

其中 $W_Q^{(k)}$、$W_K^{(k)}$、$W_V^{(k)}$ 为第 k 个头的查询、键、值投影，d_k 为键向量维度，W_O 为输出投影。第二级为指针式单头注意力，负责把高维上下文映射为一维打分：它以带缩放的 tanh 计算每个未访问节点与上下文的兼容性 u_i，并对已访问节点置为负无穷以实现掩码：

$$u_i = \begin{cases} C \tanh\left(\dfrac{\left(W_{Q2} c_t\right)^{\mathsf{T}} \left(W_{K2} h_i\right)}{\sqrt{d_k}}\right), & i \notin \pi_{<t} \\ -\infty, & i \in \pi_{<t} \end{cases} \tag{13}$$

其中 W_{Q2}、W_{K2} 为可学习投影，C 为打分限幅常数，$\sqrt{d_k}$ 为缩放因子，用于缓解内积过大引起的梯度问题。最后对未访问节点的兼容性做 Softmax 归一化，得到节点选择概率，并据此自回归地生成完整回路：

$$p_\theta\left(\pi_t = i \mid \pi_{<t}, s\right) = \frac{\exp\left(u_i\right)}{\sum_{j \notin \pi_{<t}} \exp\left(u_j\right)} \tag{14}$$

2.7. 课程式近端策略优化

路径规划的可行解数量随节点数呈指数增长，给策略搜索带来困难。本文在近端策略优化(PPO)基础上提出课程式近端策略优化(C-PPO)，通过“探索－精修”两课程段与动态超参数调度，兼顾对解空间的广泛试探与精细利用。

2.7.1. 两课程段组织

训练按轮次顺序划分为两段：探索课程段(前 30 轮)引入高熵奖励，鼓励多样化动作以充分覆盖解空间，避免过早陷入局部最优；精修课程段(第 31 轮起)逐步衰减熵奖励，把重点转向策略的精细调整与累计回报最大化，推动策略稳定收敛。两段之间不往复切换，而是通过超参数的平滑调度实现自然过渡。

2.7.2. 限幅目标与价值估计

PPO 通过限制新旧策略的概率比率来约束单次更新步长。记概率比率为 $r_t(\theta)$，限幅代理目标为 $L^{\text{clip}}(\theta)$，其中 ε 为限幅范围，$\hat{A}_t$ 为优势估计：

$$\begin{aligned} r_t(\theta) &= \frac{p_\theta(\pi_t \mid s)}{p_{\theta_{\text{old}}}(\pi_t \mid s)} \\ L^{\text{clip}}(\theta) &= \mathbb{E}_t\left[\min\left(r_t(\theta)\hat{A}_t, \text{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_t\right)\right] \end{aligned} \tag{15}$$

价值网络以均方误差拟合回报，用于为策略提供基线、降低梯度方差：

$$L^V(\psi) = \mathbb{E}_t\left[\left(V_\psi(s_t) - R_t\right)^2\right] \tag{16}$$

其中 $V_\psi(s_t)$为参数 ψ 的价值网络对状态 s_t 的估计，R_t 为对应的回报。

2.7.3. 自适应优势与策略平滑

固定形式的优势估计难以同时满足两个课程段的需求。本文引入随训练进程衰减的优势调节因子 β_t，对优势尺度进行自适应调整：

$$\hat{A}_t = \beta_t\left(L(\pi) - V_\psi(s)\right) \tag{17}$$

探索课程段取较大的 β_t 以鼓励大胆试探，精修课程段逐步衰减以实现精细调整。为进一步抑制参数震荡，引入策略平滑更新，将当前策略与上一时刻策略按系数 λ 加权混合：

$$\bar{p}_\theta(\pi \mid s) = \frac{p_\theta(\pi \mid s) + \lambda p_{\theta_{\text{old}}}(\pi \mid s)}{1+\lambda} \tag{18}$$

其中 $\bar{p}_\theta$ 为平滑后的策略，λ 在探索段取值较低以保持响应速度，在精修段取值较高以增强收敛稳定性。

2.7.4. 联合优化目标

模型以联合目标进行端到端优化，由限幅代理目标、价值损失、策略熵项与熵奖励项组成：

$$\mathcal{L}(\theta, \psi) = L^{\text{clip}}(\theta) + w_v L^V(\psi) - w_e \mathcal{H}(\theta) + w_H L^{\text{bonus}}(\theta) \tag{19}$$

其中 $\mathcal{H}(\theta)$ 为策略熵、用于维持探索性，$L^{\text{bonus}}(\theta)$为熵奖励、主要在探索段驱动多样化动作，w_v、w_e、w_H 为各项权重，随课程段动态调整。核心超参数的调度规则如表 1 所示，参数的平滑过渡保证了由探索到精修的顺畅切换。

Table 1. Curriculum scheduling of the core hyperparameters in C-PPO
表 1. C-PPO 核心超参数的课程式调度

参数	含义	探索课程段	精修课程段
ε	限幅范围	0.2	0.2
w_v	价值损失系数	0.5	0.5
w_e	策略熵系数	0.01	0.01

续表

λ	策略平滑系数	0.5	0.8
β_t	优势调节因子	1.0	0.5 (随轮次线性衰减)
w_H	熵奖励系数	0.1	0.001 (指数衰减)

2.8. 计算复杂度分析

设节点数为 m、特征维度为 d、编码层数为 L、注意力头数为 K。邻域消息传递每层需计算节点对之间的注意力，复杂度为 $O(m^2d)$，编码阶段整体为 $O(Lm^2d)$；解码阶段顺序执行 m 步、每步对全部节点打分，复杂度为 $O(m^2d)$。因此模型单次构造路径的复杂度为 $O(m^2d)$，在节点数维度上呈二次增长，且训练完成后各实例可并行批量推理，避免了传统精确方法的指数级开销。

3. 实验与分析

3.1. 实验设置与评价指标

本文以 TSP 作为路径规划的标准验证载体，使用随机生成的欧氏实例与标准 TSPLIB 实例，以平均路径长度(Length)和优化间隙(Gap)作为评价指标。优化间隙衡量所得路径相对参考最优解的相对偏差，定义为：

$$\text{Gap} = \frac{1}{N}\sum_{i=1}^{N}\frac{L_i - L_i^*}{L_i^*}\times 100\% \tag{20}$$

其中 N 为测试实例数，L_i 为模型在第 i 个实例上得到的路径长度，L_i^* 为对应参考最优解。网络设置节点与边嵌入维度为 128，PFT 隐藏层维度为 512，编码器采用 8 个注意力头、4 层网络，解码器为 8 头上下文注意力接单头指针注意力。训练使用 Adam 优化器，共 100 轮，硬件为 Xeon Platinum 8175M CPU 与 4 块 RTX 3090 GPU。

3.2. 基准规模求解精度对比

在 20、50、100 节点的随机实例上，将 DGAPN 与精确求解器 Gurobi 以及 AM [7]、G-DGANet [13] 等深度强化学习方法对比，结果如表 2 所示。

Table 2. Comparison of average tour length and optimality gap across sizes
表 2. 不同规模下平均路径长度与优化间隙对比

模型	类型	20 Length	20 Gap	50 Length	50 Gap	100 Length	100 Gap
Gurobi	Solver	3.83	0.00%	5.67	0.00%	7.76	0.00%
AM [7]	RL + S	3.85	0.08%	5.73	0.52%	7.94	2.26%
G-DGANet [13]	RL + S	3.83	0.06%	5.72	0.40%	7.79	0.43%
DGAPN	RL + S	3.83	0.04%	5.69	0.11%	7.77	0.22%

结果表明，DGAPN 在各规模下均取得深度强化学习方法中的最佳结果，优化间隙明显低于基线且十分接近 Gurobi；同时其收敛速度更快、训练过程更平稳，能够高效搜索到优质回路。

3.3. 大规模泛化与可扩展性

为检验跨规模泛化能力，将在 100 节点上训练好的模型直接用于更大规模实例，结果如表 3 所示。

Table 3. Generalization results on large-scale instances
表 3. 大规模实例泛化结果

模型	250 节点	500 节点	1000 节点	2000 节点
GAT	12.54	16.32	27.65	—
AM [7]	14.65	18.18	28.16	—
G-DGANet [13]	12.78	17.19	27.21	—
DGAPN	11.05	15.13	24.32	33.85

DGAPN 在大规模下优势明显，2000 节点时仍保持良好的路径长度。实测显示推理时间随节点数按 $O(N^2)$增长，与图注意力的理论复杂度一致；路径长度随节点数呈 $O(\sqrt{N})$的亚线性增长，与 TSP 路径长度的理论尺度相吻合，说明模型具有良好的可扩展性。

3.4. 真实分布零样本泛化

为考察分布偏移下的泛化能力，直接使用在均匀随机实例上训练的模型，对 TSPLIB 中具有真实聚类特征的实例进行零样本测试，结果如表 4 所示。

Table 4. Cross-distribution generalization on TSPLIB instances
表 4. TSPLIB 实例跨分布泛化结果

实例	节点数	最优解	AM [7] Gap	G-DGANet [13] Gap	DGAPN Gap
ch150	150	6528	4.52%	2.30%	1.15%
a280	280	2579	6.74%	4.11%	2.56%
pcb442	442	50778	9.35%	5.86%	3.42%
pr1002	1002	259045	14.21%	8.45%	4.87%

当节点分布由均匀分布转为真实聚类分布时，DGAPN 的性能衰退明显小于对比方法，在 pr1002 (1002 节点)上间隙仍保持在 4.87%，表现出较强的跨分布泛化能力，可适应真实物流与调度等路径规划场景。

3.5. 消融实验

为检验各核心设计的有效性，在 20、50、100 节点实例上开展消融，分别移除门控残差块(-GRB)、全局图读出器(-GGR)、几何边编码(-GEE)与位置感知特征变换层(-PFT)，结果如表 5 所示。

Table 5. Multi-size ablation study in terms of optimality gap
表 5. 多规模消融实验结果(优化间隙)

模型	20 节点	50 节点	100 节点
DGAPN (完整)	0.04%	0.11%	0.22%
DGAPN (-GRB)	0.05%	0.35%	0.85%
DGAPN (-GGR)	0.26%	0.68%	1.45%
DGAPN (-PFT)	0.31%	0.85%	1.68%
DGAPN (-GEE)	0.34%	1.12%	2.56%

完整模型在各规模下均最优，移除任一设计都会造成性能下降。随规模增大，几何边编码的作用显著增强，说明显式几何先验是大规模路径规划的关键因素；全局图读出与门控残差的贡献也随规模增大而上升，验证了各设计之间的协同有效性。

3.6. 课程式训练策略的对照实验

为剥离训练策略本身的贡献，直接量化课程式近端策略优化(C-PPO)相对标准 PPO 的增益，在保持 DGAPN 网络结构、嵌入维度、训练数据与优化器设置完全一致的前提下，仅将训练算法替换为标准 PPO：全程采用固定超参数，不划分课程段，也不启用优势调节与策略平滑。两种算法分别在 TSP20、TSP50 与 TSP100 上各训练 100 轮，最终解质量、收敛速度与训练过程中的平均奖励、损失分别如表 6、图 2 与图 3 所示，其中收敛速度以累计回报完成总提升量 95%所对应的训练轮次(T95)衡量。

Table 6. Training performance comparison between C-PPO and standard PPO with the identical DGAPN architecture
表 6. DGAPN 架构下 C-PPO 与标准 PPO 的训练性能对比

算法	20 Length	20 Gap	50 Length	50 Gap	100 Length	100 Gap	T95	最终奖励
C-PPO	3.83	0.04%	5.69	0.11%	7.77	0.22%	62	−7.78
标准 PPO	3.84	0.16%	5.72	0.53%	7.84	1.05%	74	−7.85

由表 6 可见，在完全相同的网络架构下，C-PPO 在三个规模上的解质量均优于标准 PPO，其中 TSP100 优化间隙由 1.05%降至 0.22%，相对降幅约 79%，收敛轮次由 74 轮提前至 62 轮，表明两课程段组织与动态调度确实加快了有效策略的形成。图 2 中，C-PPO 的平均奖励在探索课程段快速上升，并在第 30 轮课程切换后平滑进入精修阶段，曲线波动随训练推进持续收窄；标准 PPO 的奖励上升缓慢，在第 45~70 轮出现明显的局部最优平台，最终奖励更低、随机波动更大。图 3 进一步显示，C-PPO 的价值损失在课程切换后保持稳定的单调下降，策略损失振荡幅度逐轮衰减，而标准 PPO 的两类损失在整个训练过程中振荡更剧烈、最终价值损失更高。上述结果直接证明，C-PPO 带来的增益来自训练策略本身而非网络结构，课程式调度、自适应优势与策略平滑共同提升了训练稳定性与收敛效率。

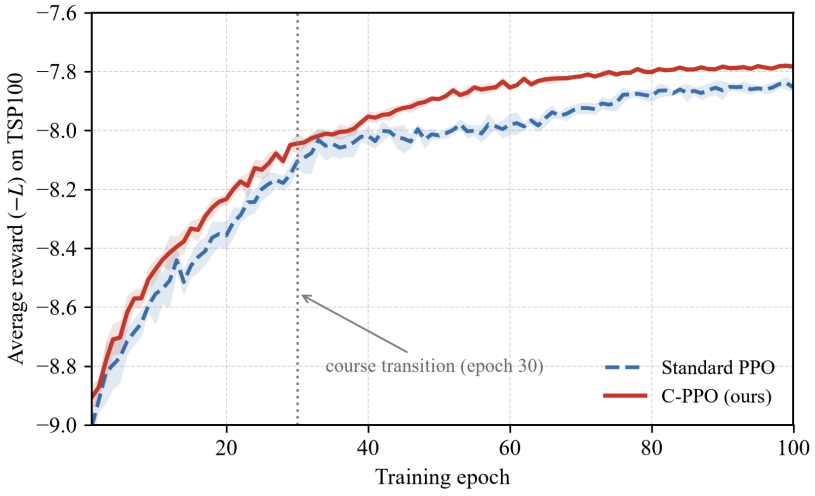


Figure 2. Average reward curves of C-PPO and standard PPO on TSP100
图 2. TSP100 上 C-PPO 与标准 PPO 的平均奖励曲线

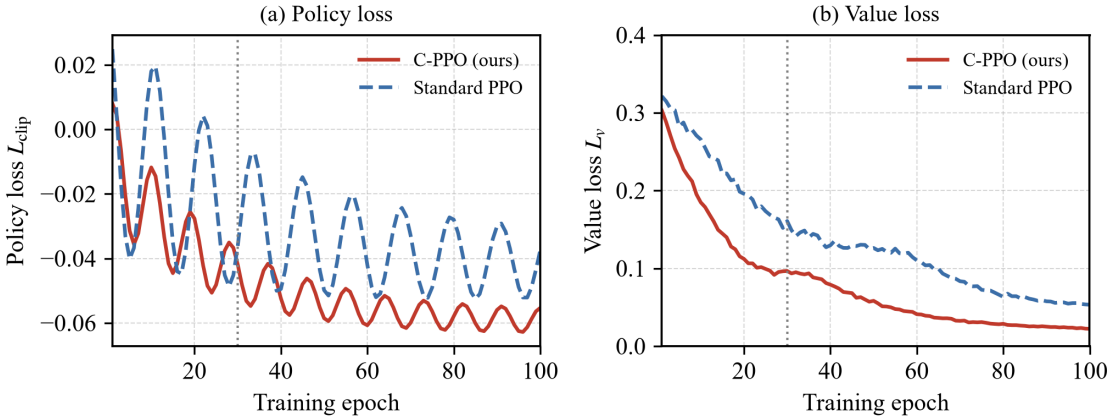


Figure 3. Training loss curves of C-PPO and standard PPO on TSP100
图 3. TSP100 上 C-PPO 与标准 PPO 的训练损失曲线

3.7. 关键超参数的敏感性分析

为检验 C-PPO 关键超参数取值的合理性，在 TSP100 上对限幅范围 ε、精修段策略平滑系数 λ、探索段熵奖励系数 w_H 与优势调节因子终值 β_t 进行单因素敏感性测试。每次实验仅改变一个参数、其余设置保持不变，以 100 节点优化间隙为评价指标，结果如表 7 所示。

Table 7. Sensitivity analysis of key C-PPO hyperparameters in terms of optimality gap on TSP100
表 7. C-PPO 关键超参数的敏感性测试(TSP100 优化间隙)

超参数	较小取值(Gap)	本文取值(Gap)	较大取值(Gap)
ε	0.1 (0.38%)	0.2 (0.22%)	0.3 (0.47%)
λ (精修段)	0.6 (0.41%)	0.8 (0.22%)	0.95 (0.53%)
w_H (探索段)	0.05 (0.45%)	0.1 (0.22%)	0.2 (0.36%)
β_t (精修段终值)	0.2 (0.39%)	0.5 (0.22%)	0.8 (0.31%)

结果表明，各参数在合理邻域内变化时，100 节点优化间隙的波动幅度不超过 0.31 个百分点，本文取值在四组参数下均取得最优结果，说明 C-PPO 对超参数不敏感、课程式调度具有较好的鲁棒性。具体而言，ε 过小会使更新步长受限、收敛变慢，过大则单次更新剧烈、策略易越过最优区域；λ 过小平滑不足、参数振荡明显，过大则策略更新迟滞、难以利用后期反馈；w_H 过小导致探索不充分、易陷入局部最优，过大则策略过于发散、精修段难以稳定收敛；β_t 过小会削弱优势信号，过大则使精修段仍保持激进试探。上述结果与两课程段的设计意图相符，也验证了表 1 所列取值的合理性。

3.8. 推理与训练效率对比

为评价算法的实用价值，在相同硬件(4 块 RTX 3090 GPU)、统一批大小与相同训练实例规模下，统计各模型在 TSP100 上完成 100 轮训练的总耗时，以及在 20、50、100、500 与 1000 节点测试集上的单实例平均推理时间(batch 大小为 1，取 1000 个实例的平均)，结果如表 8 所示。

训练耗时属于一次性成本，而推理时间随节点数近似按 $O(N^2)$增长，与 2.8 节的复杂度分析一致。DGAPN 因引入几何边编码与门控计算，推理时间较 AM 略有增加，1000 节点时单实例仅慢约 7 ms，仍

可在 0.3 s 内完成求解；与 G-DGANet 相比，DGAPN 的训练总耗时降低约 18%，各规模推理时间缩短约 19%~24%。结合表 2~5 的精度结果可知，DGAPN 以可接受的时间开销换取了显著的解质量与泛化性能提升，更好地满足路径规划“一次训练、批量复用”的实用需求。

Table 8. Comparison of total training time and average inference time between DGAPN and baselines
表 8. DGAPN 与基线模型的训练总耗时与平均推理时间对比

模型	训练总耗时/h	20 节点	50 节点	100 节点	500 节点	1000 节点
AM [7]	12.2	0.38	1.08	3.12	65.8	256
G-DGANet [13]	18.7	0.55	1.48	4.16	85.2	327
DGAPN	15.4	0.42	1.15	3.28	68.5	263

4. 讨论

实验结果显示，DGAPN 在节点遍历类路径规划上的优势来自三方面设计的协同。其一，几何边编码把通行代价作为显式空间先验引导注意力分配，相当于在搜索前对庞大解空间进行有效剪枝，这是模型在大规模与真实分布下仍保持良好泛化的根本原因。其二，门控残差块通过跨层门控缓解了深层图网络的特征趋同，配合位置感知特征变换的非线性增强，保证了高阶节点表示的区分度与几何敏感性。其三，课程式训练按课程段动态调节探索强度，改善了 PPO 在大规模解空间中易陷入局部最优或收敛不稳的问题，实现了高效的策略寻优。

同时，本文工作仍存在局限：一是目前主要面向二维欧氏距离下的无约束路径规划，对非对称代价与带约束情形的适应性尚需检验；二是在超大规模节点下，注意力机制的显存与计算开销仍然较高；三是现有结构尚未纳入时间窗、载重等复杂工程约束。

5. 结论

本文提出一种面向路径规划的距离感知图注意力策略网络 DGAPN，通过几何边编码与门控残差更新、双路池化全局图读出以及课程式近端策略优化三方面工作，有效提升了节点遍历类路径规划的求解精度与泛化能力。以 TSP 为载体的实验表明，该模型在随机实例与真实分布实例上均表现良好，在大规模场景下明显优于现有主流深度强化学习方法。

后续工作将围绕三方面展开：一是引入稀疏注意力实现模型轻量化，降低超大规模路径规划的资源消耗；二是扩展动态约束掩码，把模型迁移到带时间窗、载重限制的车辆路径规划等工程场景；三是结合真实工业物流与调度数据，进一步检验模型的实际应用价值。

特别地，针对上述通用性目标，后续将重点推动 DGAPN (尤其是几何边编码 GEE)由欧氏几何向非欧氏距离与任意成本矩阵推广。当前 GEE 中的边嵌入由节点坐标经欧氏距离计算得到，对球面距离、路网最短路距离等非欧氏度量以及时间、能耗、过路费等非对称、时变成本的刻画能力有限。为此，可将成本矩阵中的边权 d_{ij} 本身作为一种可学习的边特征直接输入：以边属性三元组(i, j, d_{ij})替代坐标对，在初始嵌入层将 d_{ij} 连同成本类型标识经线性映射与批量归一化得到边嵌入 e_{ij}^{0}，使注意力分配直接由真实通行代价而非坐标几何驱动；对非对称代价则以有向图建模，为边(i, j)与(j, i)分别维护边嵌入，并弱化对三角不等式等度量性质的依赖，由网络从数据中学习成本的结构规律。如此，GEE 不再依赖坐标与固定距离公式，DGAPN 可在仅有成本矩阵而缺少坐标的城际物流、多式联运与非欧空间路径规划等场景中直接复用，真正成为不依赖具体度量形式的通用路径规划框架。

致　谢

感谢所在单位与同行在研究过程中给予的支持与帮助，感谢相关科研项目的资助(南通市科技计划项目(MSZ2024106)、南通科技职业学院校青年基金(KY202404))。

参考文献

[1] Wang, Q., Lai, K.H. and Tang, C. (2023) Solving Combinatorial Optimization Problems over Graphs with BERT-Based Deep Reinforcement Learning. *Information Sciences*, **619**, 930-946. https://doi.org/10.1016/j.ins.2022.11.073

[2] 孔松涛, 刘池池, 史勇, 等. 深度强化学习在智能制造中的应用展望综述[J]. 计算机工程与应用, 2021, 57(2): 49-59.

[3] Applegate, D., Bixby, R., Chvatal, V., *et al*. (2006) The Traveling Salesman Problem: A Computational Study. Princeton University Press.

[4] Shi, X.H., Liang, Y.C., Lee, H.P., Lu, C. and Wang, Q.X. (2007) Particle Swarm Optimization-Based Algorithms for TSP and Generalized TSP. *Information Processing Letters*, **103**, 169-176. https://doi.org/10.1016/j.ipl.2007.03.010

[5] Vinyals, O., Fortunato, M. and Jaitly, N. (2015) Pointer Networks. https://arxiv.org/abs/1506.03134

[6] Bello, I., Pham, H., Le, Q.V., *et al*. (2016) Neural Combinatorial Optimization with Reinforcement Learning. https://arxiv.org/abs/1611.09940

[7] Kool, W., van Hoof, H. and Welling, M. (2018) Attention, Learn to Solve Routing Problems! https://arxiv.org/abs/1803.08475

[8] 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, Online, 6-12 December 2020, 21188-21198.

[9] Dai, H., Khalil, E.B., Zhang, Y., *et al*. (2017) Learning Combinatorial Optimization Algorithms over Graphs. https://arxiv.org/abs/1704.01665

[10] Joshi, C.K., Laurent, T. and Bresson, X. (2019) An Efficient Graph Convolutional Network Technique for the Travelling Salesman Problem. https://arxiv.org/abs/1906.01227

[11] Pan, X., Jin, Y., Ding, Y., Feng, M., Zhao, L., Song, L., *et al*. (2023) H-TSP: Hierarchically Solving the Large-Scale Traveling Salesman Problem. *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 9345-9353. https://doi.org/10.1609/aaai.v37i8.26120

[12] Fountoulakis, K., Levi, A., Yang, S., *et al*. (2022) Graph Attention Retrospective. https://arxiv.org/abs/2202.13060

[13] Fellek, G., Farid, A., Fujimura, S., Yoshie, O. and Gebreyesus, G. (2024) G-DGANet: Gated Deep Graph Attention Network with Reinforcement Learning for Solving Traveling Salesman Problem. *Neurocomputing*, **579**, Article 127392. https://doi.org/10.1016/j.neucom.2024.127392