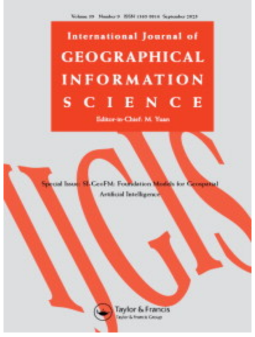


International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: www.tandfonline.com/journals/tgis20

Real-time multi-depot urban logistics optimization in megacities via transformer-based deep reinforcement learning

Qingfeng Guan, Yunpeng Fan, Yujia Wang, Lin Liang, Peng Luo & Yao Yao

To cite this article: Qingfeng Guan, Yunpeng Fan, Yujia Wang, Lin Liang, Peng Luo & Yao Yao (01 Sep 2025): Real-time multi-depot urban logistics optimization in megacities via transformer-based deep reinforcement learning, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2025.2551723

To link to this article: https://doi.org/10.1080/13658816.2025.2551723

Published online: 01 Sep 2025.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
https://doi.org/10.1080/13658816.2025.2551723

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Real-time multi-depot urban logistics optimization in megacities via transformer-based deep reinforcement learning

Qingfeng Guan[a,b], Yunpeng Fan[a], Yujia Wang[a], Lin Liang[a], Peng Luo[c] and Yao Yao[a,b]

[a]School of Geography and Information Engineering, China University of Geosciences, Wuhan, Hubei province, China; [b]National Engineering Research Center of Geographic Information System, China University of Geosciences, Wuhan, Hubei province, China; [c]MIT Senseable City Lab, Cambridge, MA, USA

ABSTRACT

Rising customer demands and the complexities of dynamic urban systems pose significant challenges for logistics distribution, especially since large-scale real-time dynamic traffic information is not always accessible. However, few studies have focused on optimizing logistics in the ever-changing traffic environments of megacities with multiple distribution centers. This study proposes two deep reinforcement learning models with Transformer architectures to optimize logistics distribution time costs across multiple depots in static and dynamic traffic scenarios, respectively. The first model (DTM-MDVRP) incorporates travel times between customers as edge information in the encoder to pre-plan delivery routes. The second model (DTM-DMDVRP) introduces a feature embedding module to extract real-time traffic information for dynamic route optimization. Wuhan city was selected for logistics optimization experiments. Results indicate that DTM-MDVRP surpasses heuristic methods and other deep reinforcement learning methods in optimization effectiveness and computation time. In dynamic urban traffic environments, DTM-DMDVRP further improves distribution efficiency. Compared to the traditional attention model, DTM-DMDVRP reduces time costs by 7.77, 3.51, and 3.58% across three problem scales and can optimize delivery routes for 100 customer points within 0.30 seconds. The proposed DTM-DMDVRP enables the real-time dynamic scheduling of logistics vehicles for logistics enterprises.

ARTICLE HISTORY
Received 9 November 2024
Accepted 19 August 2025

KEYWORDS
Multi-depot vehicle routing problem; dynamic logistics optimization; deep reinforcement learning; multi-head attention mechanism; complex road network

1. Introduction

Urban logistics distribution involves the allocation and transfer of goods within city areas through distribution centers or other logistics facilities to meet customer demands (Chang *et al.* 2020, Taniguchi *et al.* 2020). The process relies heavily on modern information technology and various transportation modes (He *et al.* 2022). The

CONTACT Yao Yao ✉ yaoy@cug.edu.cn

primary goal is to improve the efficiency and profitability of logistics companies operating in urban environments (Saha *et al.* 2023). Urban logistics is characterized by timeliness, wide coverage, high activity volume, short transportation distances, numerous logistics nodes, and constraints imposed by urban planning and control (Marcucci *et al.* 2018, Perboli *et al.* 2018, Dong *et al.* 2021). The rapid growth of urbanization and the expansion of e-commerce have significantly increased the importance of the logistics industry in urban economic development.

The increasing demand for distribution, combined with the highly complex and dynamic traffic in large cities, poses significant challenges for urban logistics. The demand has surged due to rapidly growing mobile e-commerce (Leng and Li 2022), resulting in more distribution tasks and delivery time pressures (Perera *et al.* 2020). Additionally, urban logistics distribution relies on a complex city road network (Yao *et al.* 2018), while dynamic events such as traffic congestion (Hammami 2020) and extreme weather conditions (Wu *et al.* 2020, Giordano *et al.* 2022) further reduce logistics efficiency. Effective distribution necessitates the comprehensive use of both static and dynamic information within the urban environment, coupled with rational route planning, to efficiently deliver goods to all customers across the region while minimizing logistics costs (Tang *et al.* 2023, Zou *et al.* 2024). Research in urban logistics distribution has become crucial for urban development (Strale 2019, He *et al.* 2022, Kaspi *et al.* 2022).

Urban logistics necessitates the transshipment and distribution of goods from multiple distribution centers due to the large size of cities and the relative dispersion of commercial centers (Zhou and Gao 2020). Each center serves customers within a specific region, catering to the needs of various locations. Consequently, logistics distribution is typically modeled as a multi-depot vehicle routing problem (MDVRP) (Cattaruzza *et al.* 2017, Dubey and Tanksale 2023). Multiple distribution centers handle customer deliveries, each equipped with several vehicles subject to the same capacity constraints (Vieira *et al.* 2021). Vehicles start from different logistics centers, deliver goods to customers according to predetermined routes, and return to the centers after completing all tasks. The optimization objective of the MDVRP is to minimize the total distribution cost through effective route planning and scheduling (Arishi and Krishnan 2023).

The MDVRP is a variant of the vehicle routing problem (VRP) (Konstantakopoulos *et al.* 2022) and has been shown to be NP-hard (Zou *et al.* 2024), making quick and effective solutions challenging. Traditional approaches include exact and heuristic algorithms. Exact algorithms, such as branch-and-bound (Laporte 1984, Bettinelli *et al.* 2011), can yield optimal solutions. However, the efficiency of exact methods is limited by the extensive computational resources and memory requirements as the problem size increases.

To address challenges in practical logistics, researchers employ heuristic algorithms to solve the VRP and various variants (Aliakbari *et al.* 2022, Lin *et al.* 2022, Hussain Ahmed and Yousefikhoshbakht 2023). Common approaches include genetic algorithms (GA), simulated annealing (SA), and ant colony optimization (ACO) (Abualigah *et al.* 2022). GA uses a non-deterministic evolutionary process that maintains a set of high-quality solutions evolving over time, thereby effectively avoiding local optima (Imani and Ghoreishi 2022). For example, Aliakbari *et al.* (2022) applied GA to solve a VRP involving multiple supply chains, time periods, and commodities. SA relies on local search and avoids local optima by accepting suboptimal solutions with a certain probability (Fontes *et al.* 2023). Fan et al. (2023) developed a multi-objective SA to minimize a logistics company's economic and environmental costs. ACO is inspired by the behavior of ants searching for food and employs pheromone-releasing and path-updating strategies to solve combinatorial optimization problems (Luo *et al.* 2020). For instance, Hou *et al.* (2024) proposed an adaptive ACO algorithm using real-time logistics features to improve instant delivery order scheduling efficiency. Single heuristic algorithms can effectively solve VRPs. However, as the scale of VRPs increases, the search space of heuristic algorithms rapidly expands, which significantly reduces optimization efficiency.

In recent years, researchers have explored the integration of various heuristics to develop hybrid models (Abdulkader *et al.* 2015). Hybrid heuristics combine properties from multiple algorithms to create more robust models by leveraging the strengths of different search strategies. For instance, Abdulkader *et al.* (2015) introduced a hybrid model for the MDVRP that integrates ACO with 2-opt improvement algorithms. Wang *et al.* (2019) proposed a hybrid heuristic algorithm based on ACO to solve the VRP involving customized service times. Similarly, Yao *et al.* (2023) developed a hybrid heuristic algorithm that merges Sparrow Search Algorithm (SSA) with SA, effectively solving the MDVRP under complex road networks and demonstrating suitability for large-scale urban logistics optimization. Additionally, Wang *et al.* (2024) proposed a hybrid heuristic algorithm combining spectral clustering, multi-objective ant colony optimization, and variable neighborhood search to solve the multi-depot vehicle routing problem with time windows. Although hybrid heuristic algorithms integrate multiple algorithmic advantages, the solution quality heavily depends on extensive parameter design and selection.

The rapid development of deep reinforcement learning (DRL) techniques has established a new technological foundation for solving VRPs. Vinyals *et al.* (2015) first applied DRL to solve the Travelling Salesman Problem (TSP), proposing a pointer network (PN) with an encoder-decoder structure using an attention mechanism to select elements from the input sequence. However, PN relies on supervised learning, and obtaining optimal solutions for large-scale VRP is difficult. Bello *et al.* (2016) extended the PN to reinforcement learning by setting the reward signal to the negative path length, eliminating the need for pre-collected optimal solutions. Nazari *et al.* (2018) further optimized the model by replacing the LSTM encoder in the PN with simple node embeddings, achieving results similar to traditional heuristics.

The Transformer model (Vaswani *et al.* 2017) has made significant breakthroughs in the field of natural language processing and has been applied to VRPs, demonstrating superior performance compared to traditional encoder-decoder models. Kool *et al.* (2018) introduced a DRL model based on the Transformer architecture and the policy gradient algorithm, demonstrating superiority over traditional baseline algorithms in solving various routing problems. Bdeir *et al.* (2021) developed a DRL model that combines the Transformer architecture with Q-Learning to tackle the capacity-constrained VRP and applied DRL to the MDVRP for the first time. Zou *et al.* (2024) proposed an enhanced Transformer model that incorporates a multi-attention mechanism and an attention-to-attention mechanism to address the low-carbon MDVRP. Li *et al.* (2024) proposed a multi-type attention mechanism that improves MDVRP solving efficiency and solution quality by separately encoding depot and customer features and using depot rotation augmentation. Moreover, DRL has been employed to address VRPs under large-scale (Li *et al.* 2024) and real-time constraints (Tu *et al.* 2024). Li *et al.* (2025) proposed a hierarchical DRL model with an improved Transformer to handle massive customer demand in dynamic logistics environments. While DRL algorithms can effectively solve VRPs, few studies have addressed the dynamic MDVRP in urban environments. Developing DRL algorithms capable of handling multi-depot and dynamic urban conditions is crucial for solving real-world logistics problems.

In conclusion, the reviewed literature highlights two key issues that remain insufficiently addressed. Firstly, exact algorithms and heuristic methods struggle to handle the multiple constraints and large-scale nature of urban multi-depot logistics, leading to limited computational efficiency and inadequate adaptability. Secondly, most existing DRL algorithms rely on Euclidean distance calculations or static traffic data, which do not fully capture the dynamic changes in urban road network speeds, restricting real-time optimization of multi-depot logistics routing.

Consequently, we propose two deep reinforcement learning-based logistics optimization models for complex urban multi-depot logistics, which incorporate encoders for processing environmental information. We first introduce the DTM-MDVRP (Delivery Time Minimization for Multi-Depot Vehicle Routing Problem) for scenarios lacking dynamic information or with stringent time constraints. In DTM-MDVRP, travel times between customers are incorporated as edge information in the encoder to preplan the delivery routes, aiming to improve computational efficiency. Furthermore, we present the DTM-DMDVRP (Dynamic Multi-Depot Vehicle Routing Problem with Delivery Time Minimization), which is designed for contexts with dynamic information. The DTM-DMDVRP enhances delivery efficiency by incorporating a dynamic feature embedding module that extracts real-time traffic information for route optimization under changing traffic conditions. We take Wuhan as the study area, comparing the proposed models with other logistics algorithms based on the city's road network and logistics data to verify the models' effectiveness.

2. Study area and data description

2.1. Study area

Wuhan is located in central China at the confluence of the Yangtze and Han Rivers, covering an area of approximately 8,500 square kilometers. By the end of 2022, the city had a permanent resident population exceeding 13 million. Wuhan features a well-developed transportation network structured around a "circular-radial" backbone road system, which includes urban expressways, main roads, and inner, middle, and outer ring roads (Yao *et al.* 2023). According to the Wuhan Highway Management Office (wuhan.gov.cn), the total road mileage has reached 16,000 kilometers, with a road density of 192.6 kilometers per 100 square kilometers. Wuhan's distinctive geography, featuring numerous bridges spanning the Yangtze River, contributes to the complexity of the road network. The logistics sector represents a vital component of Wuhan's tertiary industry. According to the Wuhan Statistics Bureau (https://tjj.wuhan.gov.cn/), the city's logistics industry handled 615.247 million tons of freight in 2022, with road transportation accounting for 449.774 million tons. Given Wuhan's intricate road network and substantial demand for goods transport, the city is an ideal area for investigating strategies to optimize urban logistics distribution routes.

2.2. Dataset

The primary datasets include logistics data, road network data, and taxi trajectory data. The logistics data, comprising customer points and logistics centers, were obtained from a logistics company in Wuhan via Amap (https://lbs.amap.com/). A random selection was made of 1,000 customer points and 4 logistics centers (Figure 1). Each customer's data includes latitude and longitude coordinates and delivery quantities (random values ranging from 1 to 100), as shown in Table 1.

The road network data (Figure 1) were sourced from OpenStreetMap (OSM) (http://www.openstreetmap.org), an open-source platform that provides freely accessible digital map data (Vargas-Munoz *et al.* 2021). The extracted road network for Wuhan includes 81,711 edges along with various associated attributes such as road latitude and

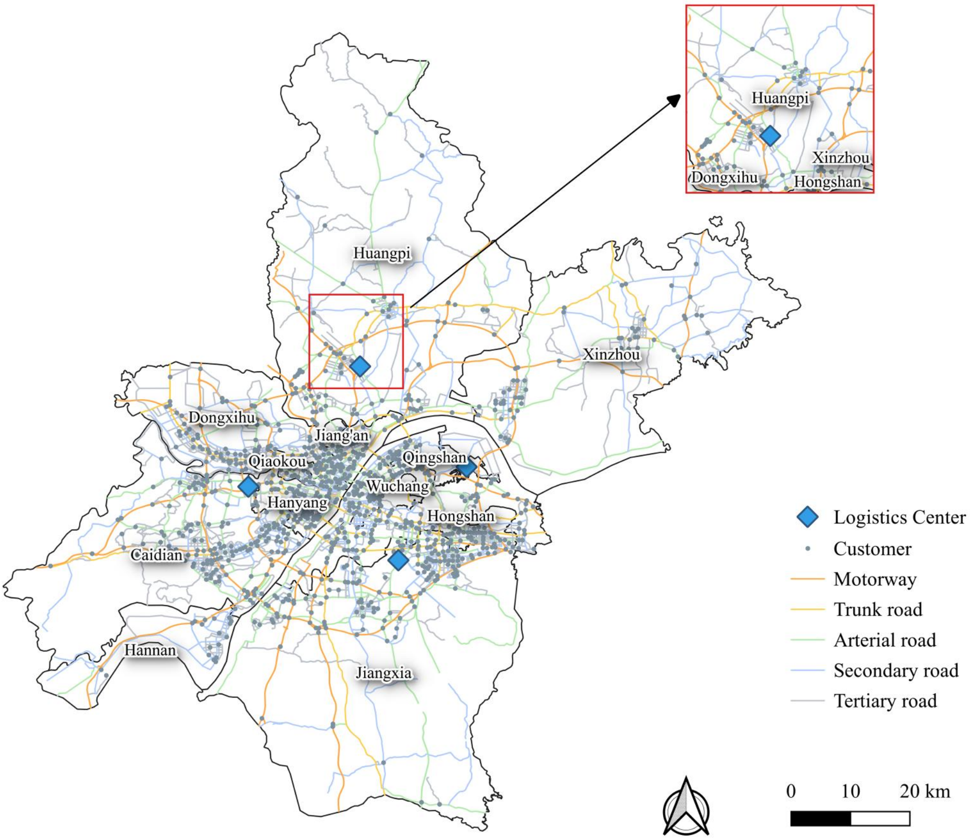


Figure 1. Logistics centers, customers and road network data in the study areas.

longitude, road type, length, and average travel speed. To accurately reflect actual traffic flow conditions, this study employed taxi trajectory data collected from a taxi company in Wuhan on 21 March 2013. The trajectory data contains extensive records of real-time location and speed information for taxis, which can be used to compute average travel speeds on various roads at different times. As demonstrated in previous studies, large-scale taxi trajectory data provides broad spatial and temporal coverage, enabling accurate characterization of urban traffic patterns (Wang *et al.* 2021, Chen *et al.* 2024)

3. Methodology

The following section describes the proposed DRL approach and the corresponding technical framework, as illustrated in Figure 2. We first present DTM-MDVRP, which incorporates edge information into the encoder and embeds the information into the Transformer's multi-head attention layer. We then introduce DTM-DMDVRP to optimize

Table 1. Customer point location coordinates and delivery requirements.

Sequence	Longitude	Latitude	Delivery weight/kg
1	114.06231	30.40521	6
2	114.13948	30.48460	3
3	114.38444	30.46892	5
…	…	…	…

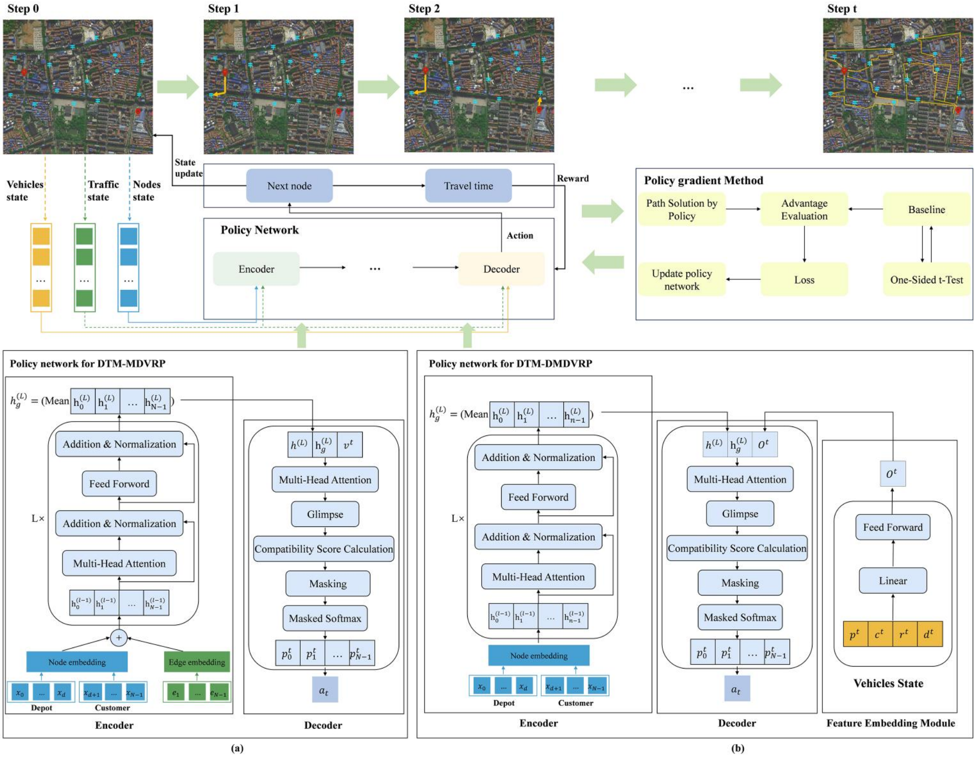


Figure 2. Diagram of DRL-based learning framework: (a) policy network for DTM-MDVRP; (b) policy network for DTM-DMDVRP.

routes within the context of dynamic urban environments. The DTM-DMDVRP accounts for temporal variations in edge weights and introduces a feature embedding module (Tang *et al.* 2023) to capture dynamic features of the MDVRP. In the decoder stage, a customized masking scheme ensures path feasibility and progressively generates logistics vehicle routes. Finally, the policy network is trained using the policy gradient algorithm.

3.1. Problem description

The MDVRP is modeled as a weighted graph $G(N, E, W)$, encompassing logistics centers, customer points, and road networks. The vertex set N consists of logistics centers $\{n_1, n_2, \ldots, n_d\}$ and customer points $\{n_{d+1}, n_{d+2}, \ldots, n_L\}$. The $E=\{e_{1\cdot1}, e_{1\cdot2}, \ldots, e_{i\cdot j}\}$ represents the set of roads. The $W=\{w_{1\cdot1}, w_{1\cdot2}, \ldots, w_{i\cdot j}\}$ represents the shortest weights between nodes. The average travel speed v' of a road is determined by the road classification. Travel time is calculated using the distance and speed of the road, yielding the inter-node weights $w_{i\cdot j}$.

Moreover, each customer j has a cargo demand q_j. Each logistics center d operates a fleet of homogeneous vehicles $V_d = \{v_1, v_2, \ldots, v_d\}$ with a capacity of Q. Vehicles depart from logistics centers to distribute goods throughout the urban road network. The solution to the MDVRP must adhere to the following constraints: each customer is served by exactly one vehicle; each vehicle must depart from and return to the same logistics center; the total customer demand cannot exceed the maximum load capacity; and each vehicle can complete the distribution service only once.

Given R as the solution to the MDVRP, the corresponding logistics distribution time can be calculated as follows:

$$T(R) = \sum_{d=1}^{D}\sum_{v=1}^{V}\sum_{n=1}^{N} T(x_{r_v^d(n)}, x_{r_v^d(n+1)}) \tag{1}$$

where $r_v^d(n)$ denotes the n-th customer point served by the v-th vehicle from logistics center d, and $T(x_{r_v^d(n)}, x_{r_v^d(n+1)})$ represents the delivery time between customer points.

3.2. Markov decision process of MDVRP

The construction of a solution for MDVRP is essentially a sequential decision-making process. At each step, the agent selects and adds a node to the current solution until a complete distribution path is established. In this study, the MDVRP is modeled as a Markov Decision Process (MDP), with the state S, action A, reward R, and state transition rule τ defined as follows.

The system's state includes both static and dynamic information. Static information includes the states of logistics centers and the customer points, denoted as $x_i = (p_i, q_i)$, where p_i represents the geographic coordinates of the node, and q_i represents the node's demand. Dynamic information covers vehicle status, customer accessibility, and traffic conditions. The action a_t determines which node the vehicle serves at time step t. Actions are selected based on the policy π and state s_t, i.e., $a_t \sim \pi(a|s_t)$. During the training phase, actions are randomly selected from the action space.

Various decoding strategies, such as greedy and sampling methods, are employed during the testing phase.

The system transitions from state s_t to state s_{t+1} based on the action currently being executed. While the node states remain static, the vehicle states, customer accessibility, and traffic conditions are updated concurrently with the execution of the action. The objective is to minimize the delivery time of logistics vehicles. Thus, the reward is defined as the negative value of the objective function, as shown in Equation (1).

3.3. Policy network

To minimize the delivery time of logistics vehicles, a Transformer-based policy network models the agent's stochastic policy $\pi_\theta(a_t|s_t)$ with parameter θ. Following the MDP framework, the model selects a vertex at each time step and generates a complete sequence solution. The formulation of the agent's probabilistic policy is presented as follows:

$$p_\theta(a|s) = \prod_{t=1}^{T} \pi_\theta(a_t|s_t, a_{1:t-1}) \tag{2}$$

where $\pi_\theta(a_t|s_t, a_{1:t-1})$ represents the probability of taking action a_t at time step t, given the state s_t and the history of previous actions $a_{1:t-1}$.

The policy network of DTM-MDVRP consists of an encoder and a decoder, as shown in Figure 2(a). For DTM-DMDVRP, the policy network additionally incorporates a feature embedding module, as illustrated in Figure 2(b).

3.3.1. Policy network for DTM-MDVRP

3.3.1.1. Encoder. The encoder embeds the original features of MDVRP instances into high-dimensional vectors. Leveraging a multi-layer attention mechanism, the encoder generates embeddings for individual nodes as well as a mean embedding that captures the overall graph structure. In contrast to previous studies (Kool *et al.* 2018, Arishi and Krishnan 2023), the proposed DTM-MDVRP integrates travel times between nodes as edge features into the initial embedding space.

The encoding process begins by embedding the node features into a high-dimensional vector $h_{n\_i}^{(0)}$ with dimension d_h=128. The initial embedding is computed using different parameter sets based on the node type (logistics center or customer), as specified in Equation (3).

$$h_{n\_i}^{(0)} = \begin{cases} w_d^x[x_i] + b_d^x, & if\ \ i \in D \\ w_c^x[x_i, q_i] + b_c^x, & if\ \ i \in C \end{cases} \tag{3}$$

where x_i denotes the geographic coordinates (latitude and longitude) of node i, and q_i represents the demand associated with customer nodes. The terms w_d^x, b_d^x, w_c^x, and b_c^x are the trainable parameters of the linear projection layers, corresponding to logistics center nodes (D) and customer nodes (C), respectively.

In parallel, the travel time from the current node to other nodes is encoded as an edge feature in the vector $h_{i\_e}^{(0)}$ of the same dimension d_h=128, as defined in Equation (4).

$$h^{(0)}_{e\_i} = w_{edge}[e_i] + b_{edge} \tag{4}$$

where e_i denotes the travel time from the current node i to other nodes, and w_{edge} and b_{edge} are the trainable parameters of the linear projection layer associated with the edge features.

The encoder then combines the node and edge features through a weighted fusion mechanism to generate the initial embedding $h^{(0)} = \{h^{(0)}_0, h^{(0)}_1, \ldots, h^{(0)}_{N-1}\}$ of the MDVRP instance.

Subsequently, the initial embedding $h^{(0)}$ is passed through L attention layers to generate the final node embedding $h^{(L)}$. Each layer consists of a multi-head attention (MHA) sublayer, a skip connection (SC) layer, a feed-forward (FF) layer, and a batch normalization (BN) layer, as detailed in Equations (5)–(7). The specifics of the MHA sublayer follow the model proposed by Kool *et al.* (2018).

$$h^{(l-1)} = \{h^{(l-1)}_0, h^{(l-1)}_1, \ldots, h^{(l-1)}_{N-1}\} \tag{5}$$

$$\tilde{h}^{(l)} = BN^{(l)}\left(h^{(l-1)} + MHA^{(l)}\left(h^{(l-1)}\right)\right) \tag{6}$$

$$h^{(l)} = BN^{(l)}\left(\tilde{h}^{(l)} + FF^{(l)}\left(\tilde{h}^{l}\right)\right) \tag{7}$$

where $h^{(l-1)}$ denotes the node embeddings from the previous layer, $MHA^{(l)}\left(h^{(l-1)}\right)$ is the output of the multi-head attention sublayer, $\tilde{h}^{(l)}$ is the intermediate embedding obtained by applying a skip connection and batch normalization to the attention output, and $h^{(l)}$ is the final node embedding produced by a feed-forward layer followed by batch normalization.

After passing through L attention layers, the encoder computes the mean of all node embeddings to obtain the graph embedding $h^{(L)}_g$:

$$h^{(L)}_g = \frac{1}{N}\sum_{i=1}^{N} h^{(L)}_i \tag{8}$$

3.3.1.2. Decoder. The decoder of the DTM-MDVRP combines the node embeddings $h^{(L)}$ and the graph embedding $h^{(L)}_g$ provided by the encoder, along with the current vehicle information V^t, to construct the decoder context H^c_t. Based on the constructed context, the decoder computes a probability distribution vector for selecting the next vertex. Details of the vehicle information and the decoder context are provided in Equations (9) and (10), respectively.

$$V^t = \left[r^t_v, p^t_v, c^t_v\right] \tag{9}$$

$$H^c_t = \left[h^{(L)}, h^{(L)}_g, V^t\right] \tag{10}$$

where r^t_v denotes the embedding of the nodes already visited by vehicle v, capturing the history of the vehicle's route; p^t_v indicates the current position of vehicle v within the graph; and c^t_v represents the remaining capacity of vehicle v, reflecting the available space for future deliveries.

At each decoding step t, the decoder generates a new context vector H'^c_t for the current state using MHA. H'^c_t is then linearly projected into a query vector q_t and a key vector k_t, which are used to compute the compatibility score u_t with all vertices, as shown in Equations (11)–(13). The compatibility score u_t is computed using a scaled dot product between the query and key vectors, followed by a non-linear transformation such as tanh, to ensure numerical stability.

$$H_t^{\prime c} = MHA(H_t^c) \tag{11}$$

$$q_t = W^Q H_t^{\prime c}, \quad k_t = W^k H_t^{\prime c} \tag{12}$$

$$u_t = \tanh \frac{q_t^T k_t}{\sqrt{d_k}} \tag{13}$$

where W^Q and W^k are learnable projection parameters.

To compute the probability vector p_t, the decoder applies masking rules to exclude inaccessible vertices due to various constraints, including vehicle capacity, customer access, logistics center access, and vehicle movement. The decoder then generates the probability vector p_t for vertex selection using the softmax function, as shown in Equation (14).

$$p_t = softmax\big(Mask_t(u_t)\big) \tag{14}$$

where $Mask_t$ denotes a masking operation that assigns large negative values to inaccessible vertices in u_t at step t.

3.3.2. Policy network for DTM-DMDVRP

The DTM-MDVRP operates using static weights when dynamic information is unavailable. To address dynamic scenarios, we propose the DTM-DMDVRP, which incorporates real-time traffic data into the policy network, enabling the model to effectively capture the dynamic nature of the MDVRP.

Compared to DTM-MDVRP, the node weights in DTM-DMDVRP are dynamically changing, and additional state variables must be considered. Accordingly, a dynamic feature embedding module (Tang *et al.* 2023) is introduced to capture vehicle and traffic state information, including the vehicle's position p^t, remaining capacity c^t, visited nodes r^t, and the current vehicle's elapsed time to reach each customer point d^t. d^t Reflects real-time traffic changes and is dynamically updated based on the continuously changing traffic conditions. The module extracts dynamic environment features through linear transformation and a feed-forward neural network (e.g., Equations (15) and (16)), which are used for vehicle routing decisions.

$$I^t = W_{linear}[p^t, c^t, r^t, d^t] + b_{linear} \tag{15}$$

$$O^t = ReLU(W_{ff} I_t + b_{ff}) \tag{16}$$

where W_{linear} and b_{linear} are the trainable parameters for the linear transformation, and W_{ff} and b_{ff} are the trainable parameters for the feed-forward network.

In the decoder, DTM-DMDVRP combines static node embeddings from the encoder with dynamic environmental features extracted by the feature embedding module. The fusion of information enables the decoder to compute a probability distribution over all candidate vertices for selection.

3.4. Policy gradient method

A standard policy gradient algorithm with a baseline is employed to reduce variance, following the methodology proposed by Kool *et al.* (2018). The algorithm primarily consists of two components: the policy network and the baseline estimator. The policy network generates a probability distribution over actions using an attention mechanism and selects actions through sampling. The baseline serves as a reference by evaluating the policy via a greedy rollout strategy, aiming to reduce the variance of the policy gradient and thereby stabilize the training process. The loss function is defined as follows:

$$\nabla_{\theta}J(\theta) = E_{p_{\theta}(\pi|s)}\left[\left(R(\pi_{\theta}) - b(s)\right)\nabla_{\theta}logp_{\theta}(\pi|s)\right] \tag{17}$$

where $R(\pi_\theta)$ is the reward generated by the policy network π_θ, $b(s)$ is the state-dependent baseline, and $\log p_\theta(\pi|s)$ is the log probability of selecting action a in state s, with π representing the policy that defines a probability distribution over actions. During training, a one-sided t-test is conducted to compare the performance of the policy network θ with the current best policy θ^* after each update. The significance level α is set to 0.05. If θ performs significantly better than θ^*, then θ^* is updated; otherwise, θ^* remains unchanged. The pseudo-code of the algorithm is provided in Table 2.

4. Results

4.1. Experiment setup

The performance of the proposed models was evaluated through extensive experiments on MDVRP instances with 20, 50, and 100 customer points. The training and testing datasets were obtained from logistics records in Wuhan. The detailed experimental setup is shown in Table 3.

Table 2. Policy gradient algorithm operations.

The policy gradient algorithm

```
Input: number of epochs $E$; number of batches $B$; significance level $\alpha$
Output: Optimized policy $\theta^*$
Initialize model parameters $\theta$, $\theta^* \leftarrow \theta$
for epoch = 1,2, … ,E do
    for batch = 1,2, … ,B do
        $x_i \leftarrow$ SampleBatch()
        for t = 1,2, … ,T do
            Calculate the output of the policy network at step $t$
        end for
        Compute the reward $L(\theta)$ for the solution $R$
        Compute the reward $L(\theta^*)$ for the baseline solution $R^*$ using greedy
                decoding
        Update $\theta$ by Adam according to $\nabla_\theta J(\theta)$
    end for
    if paired t-test $(L(\theta), L(\theta^*)) < \alpha$, then
        $\theta^* \leftarrow \theta$
    end if
end for
Return $\theta^*$
```

Table 3. The characteristics of randomly generated instance classes.

Instance class	Number of depots	Number of customers	Number of vehicles	Max capacity	Customer demand
C20-D2-V2	2	20	4	2	[0.2, 0.4]
C50-D2-V3	2	50	6	2	[0.1, 0.3]
C100-D3-V3	3	100	9	2	[0.1, 0.2]

Table 4. Instance size-based parameters for DTM-MDVRP.

Parameters	C20-D2-V2	C50-D2-V3	C100-D3-V3
Batch size	512	128	128
Batch steps	1000	4000	4000

Table 5. Instance size-based parameters for DTM-DMDVRP.

Parameters	C20-D2-V2	C50-D2-V3	C100-D3-V3
Batch size	128	64	64
Batch steps	2000	4000	4000

Table 6. Hyper parameter values.

Hyperparameters	value
Epoch	100
Seed	123
Optimizer	Adam
Tanh_clipping	10
Learning rate	1e-4
Encode layers	3
Warmup_beta	0.8
Embed_dim	128

Tables 4 and 5 present the parameters used in the study, which are determined by the instance size. The training instances were randomly generated, and the total number of test instances is 10,000, all derived from the same distribution. The hyperparameters used in the experiments are shown in Table 6.

4.2. Policy network training

All experiments were conducted on a computer equipped with an AMD EPYC 9754 CPU (2.1 GHz) and an NVIDIA GeForce RTX 3090. Figures 3 and 4 illustrate the learning progress of DTM-MDVRP and DTM-DMDVRP, depicting the negative average reward values for each epoch. As the training progresses, the distribution times gradually decrease and stabilize after approximately 20 epochs. Despite intermittent fluctuations, the training curves ultimately converge, indicating that both models have learned the stabilization strategy.

4.3. Comparative analysis

To evaluate the performance of the proposed models, we employed the following benchmark algorithms: ACO, SSA-SA (Yao *et al.* 2023), AM (Kool *et al.* 2018), and TAOA

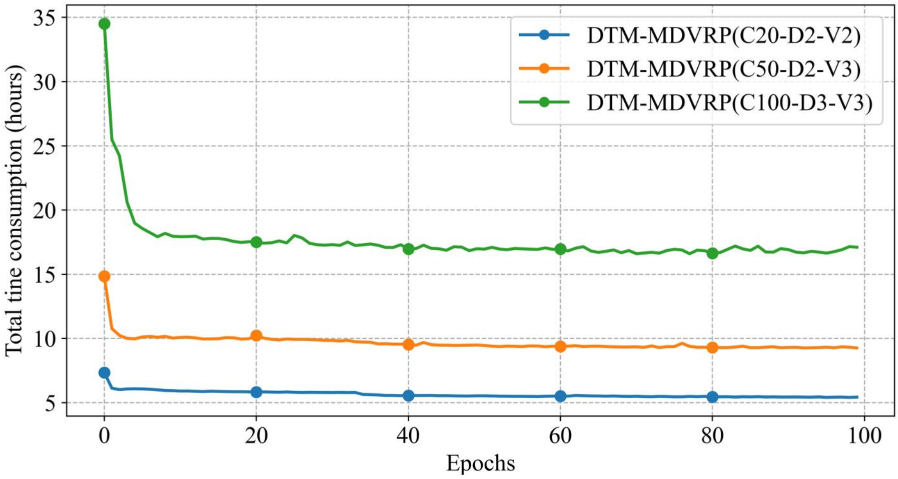


Figure 3. The total cost of the model over epochs for different DTM-MDVRP instances.

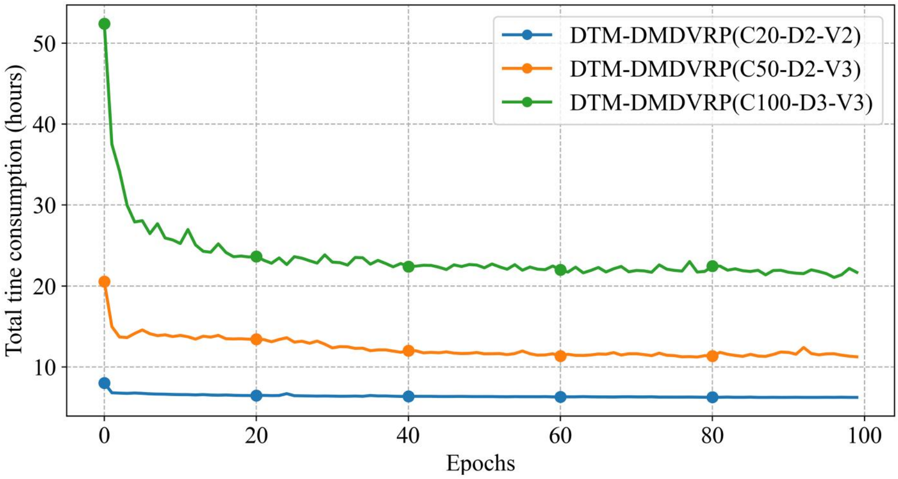


Figure 4. The total cost of the model over epochs for different DTM-DMDVRP instances.

(Zou *et al.* 2024). DTM-MDVRP was compared with all four algorithms, while DTM-DMDVRP was evaluated against AM, TAOA, and DTM-MDVRP.

Tables 7 and 8 show the test results of various algorithms across different instance sizes with performance metrics including Objective, Optimality Gap and Computation Time. The Optimality Gap is defined as the normalized distance between the Objective and the optimal Objective. In the tables, (G) denotes the greedy strategy, and (S) denotes the sampling strategy used for node selection during decoding. The results in Table 7 show that the DTM-MDVRP (S) achieves a lower cost than all other algorithms except ACO within an acceptable computation time. ACO performs best among all the algorithms due to the strong global search capability. DTM-MDVRP outperforms SSA-SA, TAOA, and AM in terms of path cost, proving the effectiveness of incorporating edge information between nodes. Despite the computational speed of the greedy decoding strategy, its performance remains inferior to that of the sampling strategy.

Table 7. Comparison of DTM-MDVRP algorithm efficiency.

	C20-D2-V2			C50-D2-V3			C100-D3-V3		
Method	Obj.(h)	Gap (%)	Time(s)	Obj.(h)	Gap (%)	Time(s)	Obj.(h)	Gap (%)	Time(s)
ACO	6.23	0	30.90	11.63	0	150.45	17.91	0	743.57
SSA-SA	6.64	6.58	34.61	13.28	14.19	55.30	21.92	22.39	82.16
AM (G)	7.05	13.16	0.04	13.22	13.67	0.09	21.73	21.33	0.21
AM (S)	6.79	8.99	0.06	12.93	11.18	0.12	21.31	18.98	0.24
TAOA(G)	7.04	13.00	0.04	13.01	11.87	0.10	20.03	11.84	0.24
TAOA(S)	6.69	7.38	0.08	12.24	5.25	0.16	19.61	9.49	0.29
Proposed DTM-MDVRP (G)	6.95	11.56	0.03	13.04	12.12	0.09	20.34	13.57	0.17
Proposed DTM-MDVRP (S)	6.52	4.65	0.07	12.10	4.04	0.13	18.99	6.03	0.27

Table 8. Comparison of DTM-DMDVRP algorithm efficiency.

	C20-D2-V2			C50-D2-V3			C100-D3-V3		
Method	Obj.(h)	Gap (%)	Time(s)	Obj.(h)	Gap (%)	Time(s)	Obj.(h)	Gap (%)	Time(s)
DTM-MDVRP (G)	8.59	13.18	0.03	18.16	24.98	0.09	26.55	17.48	0.17
DTM-MDVRP (S)	8.80	15.94	0.07	18.63	28.22	0.13	25.36	12.21	0.27
AM(G)	8.40	10.67	0.04	15.05	3.58	0.10	23.97	6.06	0.21
AM(S)	8.18	7.77	0.07	15.04	3.51	0.14	23.41	3.58	0.24
TAOA(G)	8.27	8.95	0.05	15.41	6.06	0.12	23.31	3.14	0.29
TAOA(S)	7.65	0.79	0.08	15.16	4.34	0.17	23.04	1.95	0.35
Proposed DTM-DMDVRP (G)	8.23	8.43	0.04	15.67	7.84	0.11	23.24	2.83	0.29
Proposed DTM-DMDVRP (S)	7.59	0	0.08	14.53	0	0.14	22.60	0	0.30

In small-scale scenarios, the performance of traditional baseline methods is comparable to that of DTM-MDVRP. However, as the problem scale increases, the expansion of the search space negatively impacts the computation time of heuristic methods, thereby reducing the efficiency of ACO and SSA-SA. The DTM-MDVRP demonstrates high computational efficiency, effectively balancing transportation costs and computation time. Consequently, the DTM-MDVRP is suitable for complex urban logistics systems.

The results in Table 8 indicate that the path cost of the greedy strategy is slightly higher than the sampling strategy in dynamic scenarios, consistent with the previous findings. Compared to the DTM-MDVRP, the path costs planned by the dynamic DTM-DMDVRP are reduced by 13.18, 24.98 and 12.21%, respectively. The reduction demonstrates the ability of DTM-DMDVRP to flexibly adjust the path planning of logistics vehicles according to dynamic traffic conditions, thereby improving the efficiency of logistics distribution. Additionally, in all three problem scales, the proposed DTM-DMDVRP outperforms both AM and TAOA, demonstrating the effectiveness of the dynamic feature embedding module in reducing path costs under the sampling strategy. Furthermore, the proposed DTM-DMDVRP model can optimize delivery routes for 100 customer points in 0.30 seconds, demonstrating good real-time response capability.

4.4. Route analysis

4.4.1. Optimized route of DTM-MDVRP

To evaluate the advantages of the DTM-MDVRP, three randomly generated problem instances of different sizes were optimized using DTM-MDVRP, ACO, SSA-SA, TAOA,

Table 9. Quantitative analysis comparison of optimization results.

Method	RAL (km)	RACA (km^2)	ADCL(km)	CNR	MSD(km)
DTM-MDVRP	156.66	316.20	27.67	1.28	82.55
ACO	168.13	319.47	28.14	1.40	81.86
SSA-SA	220.23	222.67	29.14	1.31	82.55
TAOA	169.36	381.20	35.20	1.42	82.55
AM	158.81	280.74	32.23	1.69	82.55

and AM. Each algorithm was executed 30 times, and the best result was retained for analysis. The optimization results were quantitatively analyzed using the following metrics: average length of optimized routes (RAL), average coverage area of optimized routes (RACA), average distance from customer points to logistics centers (ADCL), customer neighborhood ratio (CNR), and maximum service distance (MSD) of the logistics center (Table 9).

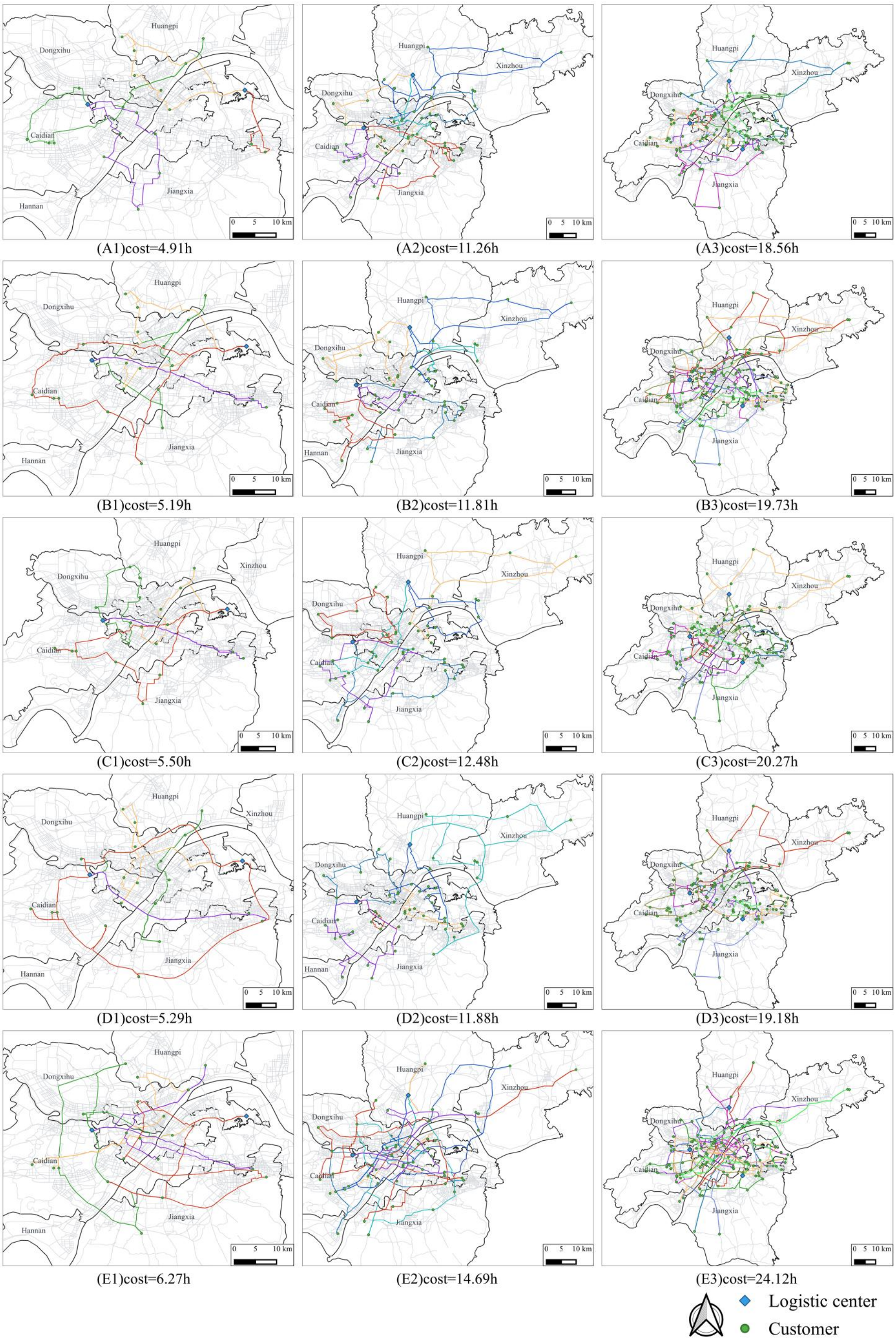
DTM-MDVRP performs best in both RAL and ADCL, with values of 156.66 km and 27.67 km, respectively. The optimized routes effectively minimize total travel length and the average distance from customer points to logistics centers, indicating a prioritization of customer points clustered around depots (Figure 5(A1, A2, A3)). The RACA of DTM-MDVRP is slightly higher than that of SSA-SA and AM, reflecting broader regional coverage while maintaining delivery efficiency. The CNR value is the lowest among all methods, indicating a high degree of spatial aggregation of customer points along the optimized routes. Additionally, the MSD of DTM-MDVRP is 82.55 km, which is comparable to other algorithms. DTM-MDVRP allows logistics centers to reasonably select service customer points and effectively partition large cities with complex demands into smaller areas served by different logistics centers. In contrast, distribution vehicles in other models may start from one logistics center but travel to customer points near another center, leading to increased distribution costs (Figure 5(B1, D1, E2, E3)).

4.4.2. Optimized route of DTM-DMDVRP

To further analyze the optimization capability of DTM-DMDVRP under dynamic traffic scenarios, 20 customer points were randomly selected for route optimization. Table 10 presents the distance and time distribution of routes passing through different speed sections. The results indicate that DTM-DMDVRP, TAOA and AM all achieve zero travel distance and time in the 0–10 km/h speed section, demonstrating that the optimized routes effectively avoid low-speed driving and improve delivery efficiency. DTM-DMDVRP exhibits a higher proportion of time spent in the 60–120 km/h speed sections than TAOA, AM, and DTM-MDVRP, contributing to improved overall performance. In conjunction with the results in Figure 6, DTM-DMDVRP can dynamically adjust vehicle routes to avoid highly congested road sections and optimize distribution efficiency.

4.4. Algorithm generalization

To evaluate the generalizability and robustness of the proposed models across diverse urban topologies, additional experiments were conducted involving 100 customer points across five districts of Wuhan: Jianghan, Hanyang, Wuchang, Hongshan, and Dongxihu. The selected districts feature a variety of road network structures, including



(A1)cost=4.91h (A2)cost=11.26h (A3)cost=18.56h

(B1)cost=5.19h (B2)cost=11.81h (B3)cost=19.73h

(C1)cost=5.50h (C2)cost=12.48h (C3)cost=20.27h

(D1)cost=5.29h (D2)cost=11.88h (D3)cost=19.18h

(E1)cost=6.27h (E2)cost=14.69h (E3)cost=24.12h

Figure 5. Different MDVRP instances (1: 20 customers, 2: 50 customers, 3: 100 customers) of the multi-depot logistic optimization results for five models: (A) Proposed DTM-MDVRP, (B) ACO, (C) AM, (D) TAOA, (E) SSA-SA.

Table 10. Percentage analysis of distance and time on different speed segments.

	0–10km/h		10–30km/h		30–60km/h		60–120km/h	
Method	Length	Time	Length (%)	Time (%)	Length (%)	Time (%)	Length (%)	Time (%)
DTM-DMDVRP	0	0	0.36	0.96	29.08	42.25	70.56	56.79
AM	0	0	1.00	2.44	27.15	42.88	71.85	54.68
TAOA	0	0	1.14	2.13	17.23	49.17	81.63	48.70
DTM-MDVRP	0.75%	1.67%	1.18	2.76	38.77	54.37	59.30	41.20

grid-based layouts, irregular historical cores, bridge-constrained zones, and wide-road industrial areas. The performance of the proposed models, including DTM-MDVRP and DTM-DMDVRP, was compared against benchmark algorithms. The results are shown in Tables 11 and 12.

Across all five districts, both DTM-MDVRP and DTM-DMDVRP achieved stable and competitive performance. DTM-MDVRP maintained low distribution time under static conditions, while DTM-DMDVRP consistently achieved the best results in dynamic traffic scenarios. The results indicate that both models adapt well to various road network structures and urban forms. The stable performance in different urban districts confirms the generalizability and operational applicability of the proposed methods in real-world logistics scenarios.

5. Discussion

5.1. Interpretation of findings

Highly complex urban road networks and rapidly increasing distribution demands pose significant challenges for modern logistics distribution optimization. However, most existing studies focus on theoretical innovations (Bdeir *et al.* 2021, Arishi and Krishnan 2023), primarily considering Euclidean distances between customer points and overlooking the structural complexity of urban road networks. Moreover, few studies address dynamic urban logistics distribution scenarios, failing to account for real-time changes in urban traffic, which limits practical applicability.

To address the problems, this study proposes the DTM-MDVRP for pre-planning logistics distribution routes in static scenarios and the DTM-DMDVRP for real-time route optimization in dynamic urban environments. Wuhan was selected as the study area, and the proposed models were trained at scales of 20, 50, and 100 customer points. Through training, the two models progressively optimized the travel time of logistics distribution routes. The training curve stabilized over time, ultimately yielding effective optimization strategies.

Regarding path cost optimization, DTM-MDVRP outperforms classical algorithms, including SA-SSA, TAOA, and AM. DTM-MDVRP reduces distribution costs by an average of 3.98, 6.42, and 10.89% compared to AM across three problem scales, demonstrating that the incorporation of inter-node edge features enhances the learning capabilities of DTM-MDVRP. Compared to heuristic algorithms, DTM-MDVRP achieves solutions within a few seconds. At a scale of 100 customer points, the average computation time of DTM-MDVRP is 0.27 seconds. DTM-MDVRP effectively addresses both transport costs and computation times in real logistics scenarios, better meeting the demands of complex urban logistics systems.

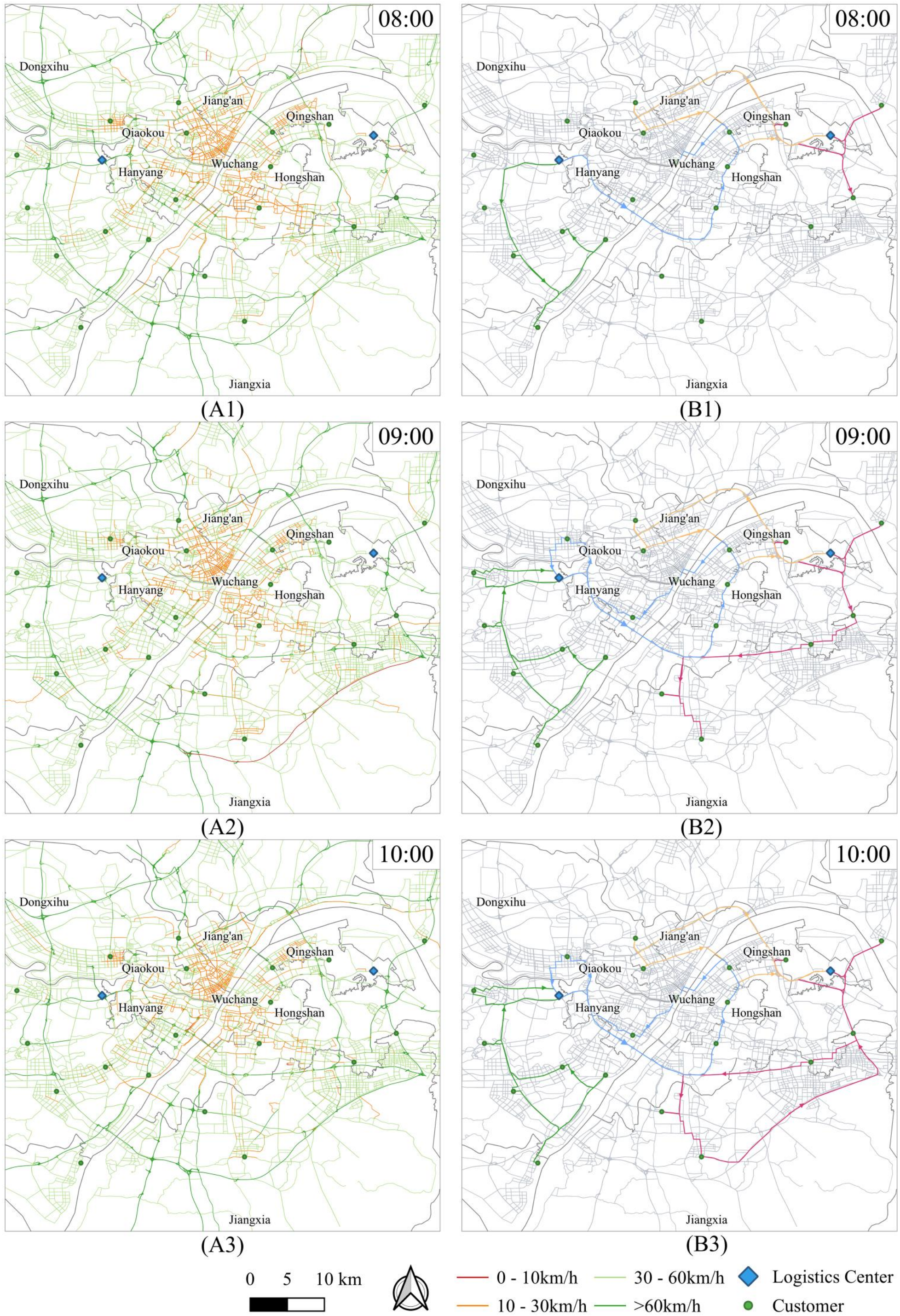


Figure 6. (A) Traffic speed changes from 8:00 to 10:00. (B) Details of delivery vehicle routes, including starting points, delivery stops, and endpoints.

Table 11. Generalizability of DTM-MDVRP across different urban districts in Wuhan.

	ACO	SSA-SA	AM	TAOA	DTM-MDVRP
Jianghan	9.25	12.72	12.32	10.45	10.41
Hanyang	11.97	16.68	12.81	12.20	12.10
Wuchang	7.65	10.87	10.11	8.81	8.46
Hongshan	11.28	16.17	14.11	12.22	12.15
Dongxihu	13.17	19.00	15.94	14.34	14.27

Table 12. Generalizability of DTM-DMDVRP across different urban districts in Wuhan.

	DTM-MDVRP	AM	TAOA	DTM-DMDVRP
Jianghan	15.80	14.44	14.30	14.19
Hanyang	16.76	15.01	14.83	14.65
Wuchang	12.86	12.67	12.53	12.37
Hongshan	18.63	14.65	14.57	14.41
Dongxihu	19.04	15.27	15.14	14.93

The DTM-MDVRP effectively optimizes distribution paths, generating reasonable routes that meet urban logistics needs. Compared to baseline algorithms, DTM-MDVRP provides zoned distribution services with an average route length of 155.66 km and an average distance of 27.67 km from customer points to the logistics center. Customer points in DTM-MDVRP exhibit high spatial agglomeration, with a nearest neighbor ratio of 1.28, which is lower than the baseline algorithms' average of 12.03%. Without additional optimization strategies, DTM-MDVRP enables logistics centers to select service customer points reasonably. The DTM-MDVRP effectively divides large areas into smaller zones served by different logistics centers, accommodating cities with expansive and complex demands.

In dynamic urban traffic environments, DTM-DMDVRP effectively reduces logistics distribution costs. Compared to DTM-MDVRP, DTM-DMDVRP reduces the time cost of dynamically optimized paths by 13.18, 24.98, and 12.21% across three problem instances. Furthermore, DTM-DMDVRP outperforms the baseline DRL methods AM and TAOA under the sampling decoding strategy, consistently producing lower path costs. An analysis of the distribution routes' passing distances and time shares in different speed sections shows that the optimized routes of DTM-DMDVRP are concentrated in non-congested areas, resulting in lower distribution costs. Validation in real traffic environments demonstrates that DTM-DMDVRP significantly improves urban logistics efficiency and reduces costs, providing essential technical support for the development of future intelligent logistics systems.

In addition to optimization performance, the proposed models demonstrate strong scalability across different problem configurations. Experimental results under varying numbers of depots, customer nodes, and vehicles indicate that DTM-MDVRP and DTM-DMDVRP consistently achieve high optimization quality and maintain low computational time. Generalizability was further evaluated through experiments conducted in five urban districts of Wuhan, each characterized by distinct road network structures. Stable and competitive performance across all districts confirms the adaptability of the proposed models to heterogeneous urban environments. Scalability across logistical configurations and generalizability across spatial structures together confirm the robustness and applicability of the proposed models in complex urban logistics systems.

5.2. The application of DTM-MDVRP and DTM-DMDVRP in a complex urban system

The primary scientific contribution of this study is the development of a rapid real-time algorithm for optimizing logistics distribution paths. The proposed algorithm is specifically tailored to urban dynamic logistics scenarios, with a strong emphasis on practical engineering implementation. To ensure applicability in real-world urban logistics, we consider multiple constraints, including the presence of multiple distribution centers, vehicle capacity limitations, and the complexity of urban road networks. Based on the constraints, this study introduces the DTM-MDVRP built on the Encoder-Decoder architecture of the Transformer. By incorporating inter-node edge features in the encoder, the DTM-MDVRP captures richer embedding information and enhances the optimization of urban logistics distribution. Furthermore, the DTM-DMDVRP is proposed for dynamic logistics optimization by utilizing real-time traffic data. The DTM-DMDVRP includes a feature embedding module in the policy network that enables the model to extract dynamic environmental features and select the most efficient paths.

While the current study focuses on Wuhan, the proposed models are theoretically applicable to urban logistics systems in other countries, provided that similar data inputs are available. Open-source platforms such as OSM provide comprehensive POI and road network data for cities in Europe and the United States. In addition, traffic datasets such as MeTS-10 (Neun *et al.* 2023) offer real-time traffic speed and volume information suitable for dynamic routing tasks. The structure and format of the mentioned datasets are compatible with the input requirements of the proposed model framework and can be integrated with minimal modification.

5.3. Limitations and future works

The study presents several limitations. One major limitation is that the current research focuses mainly on distribution efficiency. Future work will consider additional cost factors such as energy use and carbon emissions to address the optimization needs of various scenarios. Additionally, the framework can be extended to more complex MDVRP variants, including heterogeneous vehicle fleets and delivery time windows. Redefining the reward function and adjusting the masking mechanism will enable the integration of more intricate operational constraints. Another limitation arises from the heterogeneity of city sizes and structures. Although the model's generalizability has been validated across various regions of Wuhan, testing the proposed models across different city types is essential for fully validating the effectiveness of each model. The current models also struggle with larger customer numbers, such as 1,000 customer points. Future research will explore methods for optimizing city logistics on a larger scale with limited computational resources. Potential solutions may involve parallelizing the simulation of decision-making environments and distributing model training across multiple computing units, which can help accelerate experience collection, reduce memory bottlenecks, and enable the application of the models to more complex and large-scale urban logistics scenarios.

6. Conclusions

This study addresses the optimization of logistics distribution in complex urban road networks and dynamic traffic environments. We introduce an innovative end-to-end DRL model DTM-MDVRP to tackle the MDVRP in intricate road networks. Extensive experiments in large-scale urban logistics optimization demonstrate that DTM-MDVRP effectively generates high-quality vehicle routing solutions with consistent algorithmic performance. Furthermore, we present DTM-DMDVRP, designed to optimize urban logistics paths in real-time. Experimental results indicate that DTM-DMDVRP significantly reduces distribution costs in dynamic urban traffic conditions, confirming the applicability to real-world logistics distribution scenarios.

This study offers a feasible solution for dynamic urban logistics. By achieving real-time path optimization, the proposed DTM-DMDVRP can enhance logistics enterprises' efficiency, reduce operational costs, and provide valuable insights for the development of future intelligent logistics systems. Future research will apply the DTM-DMDVRP to complex urban scenarios, such as temporary road closures and urban flooding, to further validate real-time optimization capabilities in dynamic environments.

Author contributions

Qingfeng Guan: His main contributions are writing – original draft preparation, writing – review & editing, project administration, methodology. Yunpeng Fan: His main contributions are writing – original draft preparation, writing – review & editing, data curation, software, visualization. Yujia Wang: Her main contributions are writing – original draft preparation, writing – review & editing, validation, visualization. Lin Liang: Her main contributions are writing – review & editing, validation, visualization. Peng Luo: His main contributions are writing – review & editing. Yao Yao: His main contributions are writing – original draft preparation, writing – review & editing, methodology, project administration, supervision, funding acquisition.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the National Natural Science Foundation of China [42471491, 42171466]; CCF-DiDi GAIA Collaborative Research Funds [CCF-DiDi GAIA202516] and the "CUG Scholar" Scientific Research Funds at China University of Geosciences (Wuhan) [2022034].

Notes on contributors

Qingfeng Guan is a professor at China University of Geosciences (Wuhan). His research interests are high-performance spatial intelligence computation and urban computing.

Yunpeng Fan is a graduate student at China University of Geosciences (Wuhan). His research interests are reinforcement learning and logistics trajectory optimization.

Yujia Wang is a graduate student at China University of Geosciences (Wuhan). Her research interests are spatiotemporal data mining and location optimization.

Lin Liang is a graduate student at China University of Geosciences (Wuhan). Her research interests are trajectory data mining and resilient city assessment.

Peng Luo is a Postdoc Fellow at MIT Senseable City Lab. He obtained his Ph.D. degree at the Technical University of Munich, Germany. His research interests include urban analytics, spatial association modelling, social sensing, and applied artificial intelligence.

Yao Yao is a professor at China University of Geosciences (Wuhan). His research interests are geospatial big data mining, analysis, and computational urban science.

Data and codes availability statement

The codes and sample data to reproduce our work are publicly available at https://doi.org/10.6084/m9.figshare.26489536.

References

Abdulkader, M.M.S., Gajpal, Y., and ElMekkawy, T.Y., 2015. Hybridized ant colony algorithm for the multi compartment vehicle routing problem. *Applied Soft Computing*, 37, 196–203.

Abualigah, L., *et al.*, 2022. Meta-heuristic optimization algorithms for solving real-world mechanical engineering design problems: a comprehensive survey, applications, comparative analysis, and results. *Neural Computing and Applications*, 34 (6), 4081–4110.

Aliakbari, A., *et al.*, 2022. A new robust optimization model for relief logistics planning under uncertainty: a real-case study. *Soft Computing*, 26 (8), 3883–3901.

Arishi, A., and Krishnan, K., 2023. A multi-agent deep reinforcement learning approach for solving the multi-depot vehicle routing problem. *Journal of Management Analytics*, 10 (3), 493–515.

Bdeir, A., *et al.*, 2021. RP-DQN: an application of Q-learning to vehicle routing problems. *In*: S. Edelkamp, R. Möller, and E. Rueckert, eds. *German conference on artificial intelligence (Künstliche Intelligenz)*. Vol. 12873. Springer International Publishing, 3–16.

Bello, I., *et al.*, 2016. Neural combinatorial optimization with reinforcement learning. arXiv Preprint, arXiv:1611.09940.

Bettinelli, A., Ceselli, A., and Righini, G., 2011. A branch-and-cut-and-price algorithm for the multi-depot heterogeneous vehicle routing problem with time windows. *Transportation Research Part C: Emerging Technologies*, 19 (5), 723–740.

Cattaruzza, D., *et al.*, 2017. Vehicle routing problems for city logistics. *EURO Journal on Transportation and Logistics*, 6 (1), 51–79.

Chang, K.-C., *et al.*, 2020. Agent-based middleware framework using distributed CPS for improving resource utilization in smart city. *Future Generation Computer Systems*, 108, 445–453.

Chen, B.Y., *et al.*, 2024. Understanding user equilibrium states of road networks: evidence from two Chinese mega-cities using taxi trajectory mining. *Transportation Research Part A: Policy and Practice*, 180, 103976.

Dong, X., *et al.*, 2021. ITÖ algorithm with local search for large scale multiple balanced traveling salesmen problem. *Knowledge-Based Systems*, 229, 107330.

Dubey, N., and Tanksale, A., 2023. A multi-depot vehicle routing problem with time windows, split pickup and split delivery for surplus food recovery and redistribution. *Expert Systems with Applications*, 232, 120807.

Fan, L., Liu, C., and Zhang, W., 2023. Half-open time-dependent multi-depot electric vehicle routing problem considering battery recharging and swapping. *International Journal of Industrial Engineering Computations*, 14 (1), 129–146.

Fontes, D.B., Homayouni, S.M., and Gonçalves, J.F., 2023. A hybrid particle swarm optimization and simulated annealing algorithm for the job shop scheduling problem with transport resources. *European Journal of Operational Research*, 306 (3), 1140–1157.

Giordano, A., *et al.*, 2022. Impacts of topography and weather barriers on commercial cargo bicycle energy using urban delivery crowdsourced cycling data. *Sustainable Cities and Society*, 76, 103326.

Hammami, F., 2020. The impact of optimizing delivery areas on urban traffic congestion. *Research in Transportation Business & Management*, 37, 100569.

He, L., Liu, S., and Shen, Z.M., 2022. Smart urban transport and logistics: a business analytics perspective. *Production and Operations Management*, 31 (10), 3771–3787.

Hou, Y., *et al.*, 2024. Adaptive ant colony optimization algorithm based on real-time logistics features for instant delivery. *IEEE Transactions on Cybernetics*, 54 (11), 6358–6370.

Hussain Ahmed, Z., and Yousefikhoshbakht, M., 2023. An improved Tabu search algorithm for solving heterogeneous fixed fleet open vehicle routing problem with time windows. *Alexandria Engineering Journal*, 64, 349–363.

Imani, M., and Ghoreishi, S.F., 2022. Graph-based Bayesian optimization for large-scale objective-based experimental design. *IEEE Transactions on Neural Networks and Learning Systems*, 33 (10), 5913–5925.

Kaspi, M., Raviv, T., and Ulmer, M.W., 2022. Directions for future research on urban mobility and city logistics. *Networks*, 79 (3), 253–263.

Konstantakopoulos, G.D., Gayialis, S.P., and Kechagias, E.P., 2022. Vehicle routing problem and related algorithms for logistics distribution: a literature review and classification. *Operational Research*, 22 (3), 2033–2062.

Kool, W., Van Hoof, H., and Welling, M., 2018. Attention, learn to solve routing problems! arXiv Preprint, arXiv:1803.08475.

Laporte, G., 1984. Optimal solutions to capacitated multidepot vehicle routing problems. *Congressus Nemerantium*, 4, 283–292.

Leng, K., and Li, S., 2022. Distribution path optimization for intelligent logistics vehicles of urban rail transportation using VRP optimization model. *IEEE Transactions on Intelligent Transportation Systems*, 23 (2), 1661–1669.

Li, D., *et al.*, 2024. A reinforcement learning-based routing algorithm for large street networks. *International Journal of Geographical Information Science*, 38 (2), 183–215.

Li, J., *et al.*, 2024. Multi-type attention for solving multi-depot vehicle routing problems. *IEEE Transactions on Intelligent Transportation Systems*, 25 (11), 17831–17840.

Li, Y., *et al.*, 2025. A hierarchical deep reinforcement learning method for solving urban route planning problems under large-scale customers and real-time traffic conditions. *International Journal of Geographical Information Science*, 39 (1), 118–141.

Lin, B.-C., Liu, X.-F., and Mei, Y., 2022. Efficient extended ant colony optimization for capacitated electric vehicle routing. *In*: 2022 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 504–511.

Luo, Q., *et al.*, 2020. Research on path planning of mobile robot based on improved ant colony algorithm. *Neural Computing and Applications*, 32 (6), 1555–1566.

Marcucci, E., Gatta, V., and Le Pira, M., 2018. Gamification design to foster stakeholder engagement and behavior change: an application to urban freight transport. *Transportation Research Part A: Policy and Practice*, 118, 119–132.

Meduri, K., *et al.*, 2023. Developing a fog computing-based AI framework for real-time traffic management and optimization. *International Journal of Sustainable Development in Computing Science*, 5 (4), 1–24.

Nazari, M., *et al.*, 2018. Reinforcement learning for solving the vehicle routing problem. *In*: S. Bengio, *et al.*, eds. *Advances in neural information processing systems*. Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/9fb4651c05b2ed70f-ba5afe0b039a550-Paper.pdf

Neun, M., *et al.*, 2023. Metropolitan segment traffic speeds from massive floating car data in 10 cities. *IEEE Transactions on Intelligent Transportation Systems*, 24 (11), 12821–12830.

Perboli, G., *et al.*, 2018. Simulation–optimisation framework for city logistics: an application on multimodal last-mile delivery. *IET Intelligent Transport Systems*, 12 (4), 262–269.

Perera, S., *et al.*, 2020. Retail deliveries by drones: how will logistics networks change? Production and Operations Management, *29* (9), 2019–2034.

Saha, A., *et al.*, 2023. A dual hesitant fuzzy sets-based methodology for advantage prioritization of zero-emission last-mile delivery solutions for sustainable city logistics. *IEEE Transactions on Fuzzy Systems*, 31 (2), 407–420.

Strale, M., 2019. Sustainable urban logistics: what are we talking about? *Transportation Research Part A: Policy and Practice*, 130, 745–751.

Tang, M., *et al.*, 2023. Energy-optimal routing for electric vehicles using deep reinforcement learning with transformer. *Applied Energy*, 350, 121711.

Taniguchi, E., Thompson, R.G., and Qureshi, A.G., 2020. Modelling city logistics using recent innovative technologies. *Transportation Research Procedia*, 46, 3–12.

Tu, W., *et al.*, 2024. Deep online recommendations for connected E-taxis by coupling trajectory mining and reinforcement learning. *International Journal of Geographical Information Science*, 38 (2), 216–242.

Vargas-Munoz, J.E., *et al.*, 2021. OpenStreetMap: challenges and opportunities in machine learning and remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 9 (1), 184–199.

Vaswani, A., *et al.*, 2017. Attention is all you need. *In*: I. Guyon, *et al.*, eds. *Advances in neural information processing systems*. Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vieira, Y.E.M., De Mello Bandeira, R.A., and Silva Júnior, O.S.D., 2021. Multi-depot vehicle routing problem for large scale disaster relief in drought scenarios: The case of the Brazilian northeast region. *International Journal of Disaster Risk Reduction*, 58, 102193.

Vinyals, O., Fortunato, M., and Jaitly, N., 2015. Pointer networks. *In*: C. Cortes, *et al.*, eds. *Advances in neural information processing systems*. Vol. 28. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/29921001f2f04bd3baee84a12e98098f-Paper.pdf

Wang, P., *et al.*, 2021. Estimating traffic flow in large road networks based on multi-source traffic data. *IEEE Transactions on Intelligent Transportation Systems*, 22 (9), 5672–5683.

Wang, Y., *et al.*, 2019. A multi ant system based hybrid heuristic algorithm for vehicle routing problem with service time customization. *Swarm and Evolutionary Computation*, 50, 100563.

Wang, Y., *et al.*, 2024. Collaboration and resource sharing in the multidepot time-dependent vehicle routing problem with time windows. *Transportation Research Part E: Logistics and Transportation Review*, 192, 103798.

Wu, X., *et al.*, 2020. Finding of urban rainstorm and waterlogging disasters based on microblogging data and the location-routing problem model of urban emergency logistics. *Annals of Operations Research*, 290 (1–2), 865–896.

Yao, Y., *et al.*, 2018. Estimating the effects of "community opening" policy on alleviating traffic congestion in large Chinese cities by integrating ant colony optimization and complex network analyses. *Computers, Environment and Urban Systems*, 70, 163–174.

Yao, Y., *et al.*, 2023. Fast optimization for large scale logistics in complex urban systems using the hybrid sparrow search algorithm. *International Journal of Geographical Information Science*, 37 (6), 1420–1448.

Zhou, H., and Gao, H., 2020. The impact of urban morphology on urban transportation mode: a case study of Tokyo. *Case Studies on Transport Policy*, 8 (1), 197–205.

Zou, Y., *et al.*, 2024. An improved transformer model with multi-head attention and attention to attention for low-carbon multi-depot vehicle routing problem. *Annals of Operations Research*, 339 (1–2), 517–536.