Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Scalable Truck-Drone Coordination with Reinforcement Learning: Toward Real-Time Last-Mile Delivery Optimization

Ali Abdul Razzaq Taresh, Asghar A. Asgharian Sardroud, Mir Saman Tajbakhsh
Department of Computer Engineering, Urmia University, Urmia, Iran

Corresponding author: a.asgharian@urmia.ac.ir

ARTICLE INFO

Received: 31 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

The study offers a new reinforcement learning contour to adapt the traveling seller problem with a drone (TSP-D), which faces challenges in final delivery logistics. The proposed method benefits from a near proximal political adaptation (PPO) combined with a deep remaining forward nerve architecture to effectively coordinate truck-drainage operations. A large contribution lies in extended state representation, which integrates the position of transit, the remaining travel time and viable drone-causal nodes, which use the algorithm of Dijkstra. The model was evaluated under two scenarios: unlimited and limited drone area. The results show calculation performance on better routing efficiency, solution quality and benchmark algorithms. PPO ensures policy stability and effective learning in a complex urban environment. The residual architecture reduces the disappearance of the disappearance, which enables intensive network training. This approach supports scalable and intelligent decision -making for drone -assisted logistics. The study helps to promote real -time, data -driven distribution systems for final measurement.

Keywords: Travelers Seller problem, drone delivery, reinforcement learning, proximal political adaptation, deep residual networks, final meals, state representation, hybrid neural architecture.

1. INTRODUCTION

The rapid growth of e-commerce and the same day's delivery services has made outstanding requirements for urban logistics systems. The last mile distribution of the most important and cost -intensive components of the Logistics supply chain is the last step of transport to the customer from a distribution hub [1]. Traditional distribution models, usually depending on ground -based vehicles such as trucks and vans, are limited by urban crowds, fixed road infrastructure and distribution planning of obstacles, leading to disability in the route plan and distribution time management [2].

To solve these challenges, unmanned air -forces (UAV) or drones, appeared as a transformative technique in logistics. Drone offers a unique ability to bypass terrestrial traffic and now difficult with wheels, offering fast and more flexible delivery options [3]. However, drones are also limited by factors such as utility load capacity, battery endurance and flight area. As a result, the integration of drones with traditional trucks into a cooperative distribution model travelers sellers with drone (TSP-D) is known as a problem-this becomes a compelling research focus in logistics optimization [4].

TSP-D extends a classic travel sales problem (TSP) by starting coordination between a truck and drone, both should visit a set of customer spaces, reducing the total delivery cost or time [5]. The complexity of TSP-D dates from the need to determine two weird vehicles with different abilities and obstacles at the same time. While the truck can provide several packages over long distances, the drone is limited to one-on-one delivery, and the truck will eventually recognize. This coordination introduces synchronization barriers and improves the calculation complication of the problem.

Existing approaches to solve TSP-D depend on accurate methods such as metahiuristic algorithms, including the existing approach to mainly mixed pilgrimage (MILP) [6], as well as genetic algorithms, ant colony adaptation and taboo search. [7][8][9]. For example, Wang et al. [10] To improve the performance of drone-truck coordination in TSP, a hybrid metaheyuristic to add genetic algorithms and dynamic programming. In the same way, Buman has et al. [6] A compact MILP presented a branch-and-value algorithm based on MILP formulation, which is able to solve TSP-D examples of medium-prime to optimize.

Although these methods have achieved promising results, they suffer from boundaries, dynamic relationship compatibility and decision -making of real -time in scalability. As the number of distribution points increases, the calculation time required by traditional looser increases rapidly, making them impractical for the deployment of the real world in large urban areas. [11].

To remove these boundaries, reinforcement learning (RL) has recently been introduced as a viable alternative for combinatory optimization problems, including routing and planning works [12]. RL enables agents to learn optimal decision strategies through testing-and-sled interactions with a dynamic environment, which is especially suitable for adaptive root in complex, real-time settings. In the case of TSP-D, RL can strengthen distribution systems to adapt truck and drone functions based on current conditions, environmental reaction and operational obstacles.

In RL methods, proximal political adaptation (PPO) has emerged as a condition -of -art -algorithm for continuous control and policy screen adaptation [13]. PPO provides stable and effective teaching using a cut lens function, which prevents very aggressive political updates. When combined with deeply remaining nerve networks, PPO can handle large state action sites and reduce the problems that disappear gradients during exercise [14]. Residual connections allow deep network architecture that better generalize complex problems, such as TSP-D with limited drone area.

The study proposes a new RL-based structure to solve TSP-D, combining PPOs with increased condition representation, including the status of transit, battery coatings and available drone knots, which are calculated using the algorithm of Dijkstra. The proposed approach is evaluated under both unlimited and limited drone -line scenarios, and compared to traditional hurryistics such as genetic algorithms (GA), particle crew adjustment (PSO) and grassy optimization algorithm (goa) [15] [16] [17]. Experimental profit solutions show the better performance of the proposed method in terms of quality and calculation efficiency, and installs it as a competitive solution for the final miles distribution logistics for real time.The key contributions of this research include:

• Introduction to a decorated and dynamic condition representation that fits the complications of TSP-D.

• To benefit from PPO to learn stable and scalable policies in strange vehicle coordination.

• Designing a hybrid deep residual network to improve the learning convergence and strengthen solutions.

• To provide empirical evidence that RL improves traditional metahuristics in both quality and execution time in complex TSP-D landscape.

2. RELATED WORK

Coordination of drones and ground vehicles in the distribution of previous meals has emerged as a promising area of research in recent years. Travelers sellers problem with drone (TSP-D) is a challenging expansion of classic travel sellers Problem (TSP), where both a truck and a drone collaborate to operate a set of customers [18]. Due to the underlying NP tricks of TSP-D and further synchronization and battery operations of drones, researchers have discovered a number of adaptation strategies from accurate methods to metaheyuristic and learning-based approaches.

A significant contribution to this domain is presented by Agatz et al. [5], which formally provided TSP-D and provided basic MILP models for coordinated delivery of the truck drain. Buman et al. [6] This research line improved this line by suggesting an accurate branch and worthwhile using a compact mixed linear programming (MILP) formulation and dynamic programming. His approach solved examples for optimal with 39 customers, crossing the previous accurate algorithm at both speed and scalability.

Grants for accurate methods, their ability to handle examples of a major problem, metahiuristic approaches have been given traction. For example, Wang et al. [10] We introduce a hybrid method that combines genetic algorithms and dynamic programming, especially aimed at synchronization between trucks and drone operations. Similarly, has et al. [7] Viewing efficiency in different reference scenarios, suggested a metaheyuristic hybrid framework in combination with false analling and taboo search after routing joint vehicle drain delivery.

Recently, because of its ability to make adaptive decisions in the dynamic environment, the focus has been focused on using reinforcement learning (RL) to solve complex combinatorial problems such as TSP-D. Bello et al. [12] To solve the classic TSP, a nervous conference customization structure introduced using attention -based codes with policy gradients, which marks an infection against data -driven root solutions. After this direction, Nazri et al. [18] A Pointer Network -based Deep Reforcement (DRL) approach was presented to handle a Vehicle Route Problem (VRP) without relying on traditional looser or craft facilities.

While RL models show laws to be compatible with the changing environment, a challenge lacks their state coordination for problems involving multiple weird means. Lu et al. [19] His model demonstrated the improvement of pure attention -based methods in both solution quality and learning stability.

In order to improve generalization and convergence, various hybrid nerve architecture have been proposed. Residual networks, originally developed for image recognition [14], Dark reinforcement has proven to be effective in learning references, especially to train deep networks with stable gradients [20]. In logistics optimization, it is possible for remaining network models the model to capture more fine state action representations, which are essential to dynamic environment with complex.

Another emerging technology is the use of Graph Neural Network (GNN) for routing problems. Cool et al.[21] These models naturally capture spatial and topological ratios between nodes, TSP-D is a significant requirement with dynamic customer placements and developed vehicle states.

Despite the progress, literature shows the lack of reinforcement learning frames that specifically correspond to TSP-D with realistic drone barriers. Most existing models either simplify the battery limits or ignore the synchronization delay between trucks and drones. In addition, state representation that is used in pre-work is unable to include action-rich elements such as transit status, drone launch option and Dijkstra-based ricability analysis.

In response, this study proposes a hybrid PPO-based reinforcement learning model, integrated with deep remaining nervous networks and a comprehensive condition vector, enabling decisions on real time and better performance in scenarios with a limited drag range. By addressing both scalability and realism, the proposed model contributes a new and effective approach to solving TSP-D.

3. MATERIALS AND PROPOSED METHOD

When addressing the traveling seller problem with a drone (TSP-D), this research creates a strong experimental setup and introduces a hybrid reinforcement learning-based solution that improves traditional heriaistic and accurate methods. The approach combines an intensive relevant environment to achieve effective coordination between a truck with advanced nervous architecture and proximal political adaptation (PPO) algorithm and a drone for distribution of the last meal.

3.1 Experimental Environment and Materials

Experimentally setup simulates a realistic urban delivery environment where a delivery truck gets the help of a drone to serve customer places. Two primary problem landscapes were considered:

- Scenario 1: Unlimited drone area - the drone is considered to be sufficient battery capacity to operate any node regardless of distance.
- Scenario 2: Limited drone range flight spacing is limited by battery capacity, which requires optimal selection of drone launch points and destinations.

To simulate these conditions, a synthetic data set was generated based on grid maps with 11, 25 and 50 customer knots. The coordinates of each node were randomly distributed while maintaining the minimum vacancy to simulate city blocks. The atmosphere was programmed using the Openai Gym in Python, which allows dynamic interaction between agent policy and delivery landscape.

The experiments were performed on a machine with an Intel Core i7 processor, 32 GB of RAM and an Nvidia RTX GPU. Tensorflow was used to use the nervous network, and the PPO algorithm was used from a stable baseline 3. Benchmark data and performance were compared to the modern TSP-D model with the results known, as introduced by Murray and Chu [21], and Poiconan. [22].

3.2 Proposed Reinforcement Learning Framework

The suggested method prepares TSP-D as a Markov-declining process (MDP), and introduces a hybrid enhancement learning model created by three important innovations: an increased state vector, a deep remaining actor-critical architecture and a PPO learning algorithm. Each of these items is explained below.

3.2.1 Markov Decision Process (MDP) Formulation

The TSP-D is modeled as an MDP where:

• **Condition (s)** captures trucks and drones locations, their par transit status, battery level and remaining delivery nodes.

• **Action (a)** defines whether the drone is launched, restored or whether the truck continues to the next location.

Reward (s) encourages efficiency by punishing long routes, overweight or passivity in the battery.

This MDP setup allows for dynamic, phased coordination between drones and trucks under real operating barriers, especially in the energy world environment.

3.2.2 Enhanced State Representation

A major contribution to this research lies in the improved state vector used by the drug. Unlike traditional representations, the proposed vector includes:

- **Truck and drone current coordinates**
- **In-transit status of both vehicles** (idle, en route, or delivering)
- **Battery level of the drone**
- **Remaining delivery nodes**
- Value for drone, matrix, calculated through the Dijkstra algorithm to reflect the nodes available below the current battery level [23]

This level of extension allows the agent to evaluate both immediate and future obstacles, able to make more strategic decisions. Previous studies [24] have shown that rich state representatives lead to better political learning in complex environments as distribution optimization.

2.23. Proximal Policy Optimization (PPO)

In order to adapt the agent's decision policy, we adopt the proximal political adaptation (PPO) algorithm. PPO provides significant benefits for distributing the real world, including:

- • Cut the objective function of the surrogate: Stabilizes exercise by avoiding extremely large updates.
- • Several eras per batch: Refruit's data efficiency by reusing the collected path.
- • Adaptive kl -penal [25].

PPO has performed better in continuous control problems compared to traditional political shields or Q learning methods, especially in multi-agent environment [26].

3.2.4. Hybrid Deep Residual Network Architecture

In order to further improve learning, the actress and critic network in the PPO framework is designed with deep remaining forward architecture. Residual blocks enable deep network training by allowing identity mapping that prevents the shield from disappearing or explosion. The network includes:

- Input layer: Increased condition vector.

- Residual block: Each block consists of two fully connected layers and a shortcut connection.

Output Layers: One for Action Probability Distribution (Actor), and for a price estimate (critic).

The residual connection, first he was proposed by et al. ,,

3.2.5. Training Process and Hyperparameters

The training follows a standard PPO pipeline:

1. Inquiology: Random Inquisition of Neural Network Weight.

2. Plause collection: Agent collects episodes by interacting with the environment.

3. Estimates of benefits: Estimate Estimate (GAE) is used to calculate temporary-it-out benefits.

4. Policy and price update: The debt has been a return to update network weight.

5. Initial restriction and evaluation: Training improves the policy once.

Hyperpeter was doing well with the web search. For example, the learning rate was determined by 3e-4, discount factor γ = 0.99 and batch size = 2048.

3.3 Advantages Over Heuristic Methods

The suggested PPO-based method was tested against traditional heroistic methods such as Genetic algorithms (GA), Particle Herd adaptation (PSO) and grasshopper optimization algorithm (GoA). Evaluation Metrix included total route costs, energy consumption and drone inactive time. The PPO model, in addition to both unlimited and limited drone series scenarios, improved all basic methods in both solution quality and calculation time.

This reflects the scalability and flexibility of learning reinforcement for drone-assisted delivery functions, even under barriers to the algorithm algorithms struggling to clear code. [28][29].

The proposed method introduces a broad reinforcement learning solution to TSP-D, which is produced on a deep remaining architecture to learn a well-structured state representation, powerful PPO training dynamics and optimal policy. Unlike existing heroistic and accurate approaches, this method grows better with changes in the environment and increasing problem size.

4. RESULTS

Travelers sellers with drones (TSP-D) were widely used in two main scenarios, strictly evaluating the effectiveness of the reinforcement of the proposed reinforcement learning framework to solve the problem, two main scenarios were widely used: (1) with **unlimited drone flying range**, and (2) With limited drone flying range. In both cases, the proposed method was benchmark against three installed metaheyuristic algorithms-genetic algorithms (GA), particle flock adaptation (PSO) and grasshopper optimization algorithm (Goaaa). The assessment measurement solution focuses at cost price, interpentage and calculation time, and provides a multi -phase -set assessment of the model performance.

4.1 Scenario 1: Unlimited Drone Flying Range

In this initial evaluation, the drone was allowed to travel to any distance without batteries, providing a base line to assess the net routing efficiency of the proposed method. The model was tested by delivery examples with a node size of 20, 50 and 100, and performance was measured against the aforementioned metaheyuristics.

Combined results In Table 1, it shows that the method of learning the proposed reinforcement maintained the low costs in all problem sizes. For example, the traditional methoristic algorithms that gave and pso improved, and came very close to Goa, especially at 100 nodes. In addition, Figure 1 represents the comparative performance of all algorithms under the position of the unlimited drone area, which clearly explains the costs and scalability of the proposed method.

- Ved 20 noder oppnådde den en gjennomsnittlig kostnad på 281,47, som er litt over Goa (280,72), men bedre enn GA (296,12) og lik med PSO (281,43).
- For 50 nodes, a cost of 394.77 recorded, which is very close to Goa (393.61), and much better than GA (415.30) and PSO (396.06).
- Ved 100 noder skaffet metoden en kostnad på 540,09, bare mer enn Goa (539,98), mens den spesifikt presterte bedre enn GA (594,81) og PSO (552,07).

Gap -metrical, represented the accuracy of further method, representing the deviation from the optimal or best executive method:GAP of **0.60%** for 20 nodes GAP of **0.83%** for 50 nodes GAP of **1.69%** for 100 nodes

- These low different values indicate that the proposed method estimates optimal solutions with high accuracy, even the problem complications as scales.
- When it comes to calculating time, the method demonstrated exceptional efficiency:
- 0.04 seconds for 20 nodes
- 1.11 seconds for 50 nodes
- 2.12 seconds for 100 nodes

This time is faster than GA (up to 5,195 seconds for 100 nodes), PSO (3 777 seconds) and Goa (8,967 seconds), strengthening the fitness of the method of real -time applications.

Nodes	GA Cost	PSO Cost	GOA Cost	Proposed Method Cost	Proposed Method Time (s)
20.0	296.12	281.43	280.72	281.47	0.04
50.0	415.3	396.06	393.61	394.77	1.11
100.0	594.81	552.07	539.98	540.09	2.12

Table 1: Cost and Time Comparison – Unlimited Drone Range

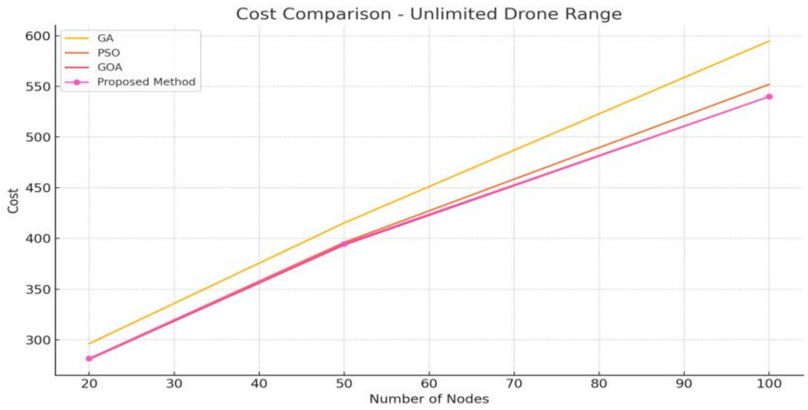

Figure 1: Cost comparison among GA, PSO, GOA, and the proposed method in the unlimited drone range scenario across different node sizes.

4.2 Scenario 2: Limited Drone Flying Range

This landscape introduced an important operating barrior-drone battery reduced the lifetime-more complex than making routing decisions. The vector of the condition was dynamically updated through the use of the algorithm of Diozxt, which was an ingredient for determining the viable drone roads, which proved to be important to maintain performance during the obstacle.

This landscape introduced an important operating barrior-drone battery reduced the lifetime-more complex than making routing decisions. As shown in Table 2, the proposed method continued to perform competitive, and maintained low costs and minimum calculation time below the battery limitations. Performance beaches are more painted in Figure 2, which reveals the cost difference between algorithms between 10, 20 and 50-NOD delivery landscape.

- For **10 nodes**, the method produced a cost of **2.02**, nearly identical to GOA (2.01), and better than PSO (2.03) and GA (2.20).
- At **20 nodes**, the proposed method matched GOA at **2.34**, outperforming PSO (2.36) and GA (2.57).
- For **50 nodes**, it achieved the best result at **3.25**, edging out GOA (3.26), and significantly ahead of PSO (3.37) and GA (3.75).

GAP values remained very low:

- **0.54%** for 10 nodes
- **0.56%** for 20 nodes
- **0.59%** for 50 nodes

Even during the increasing complexity of the battery bear, the method preserved its strength and adaptability. Goa fell slightly in the back of the range (0.83%on 50 nodes), while GA and PSO showed much larger deviations, confirming the effectiveness of the method during operating barriers.

In terms of **runtime**:

- The proposed method completed all the tasks in less than 1.12 seconds compared to 369 seconds GA, 296 seconds PSO and 558 seconds of Goa for 50-node scenarios.
-

Nodes	GA Cost	PSO Cost	GOA Cost	Proposed Method Cost	Proposed Method Time (s)
10.0	2.2	2.03	2.01	2.02	0.06
20.0	2.57	2.36	2.34	2.34	0.41
50.0	3.75	3.37	3.26	3.25	1.12

Table 2: Performance Comparison – Limited Drone Range

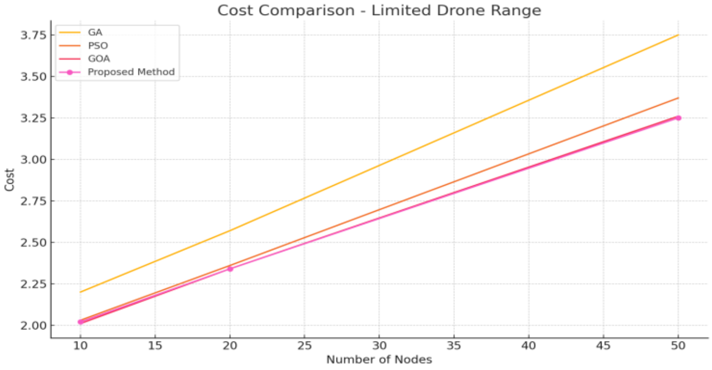

Figure 2: Performance comparison for limited drone range scenario with increasing problem size (10, 20, and 50 nodes).

4.3 Visual Validation with Known Solutions

To validate the reliability of the solution, the proposed method was applied to small examples with known optimal solutions known. In these cases:

- It achieved a GAP of **0.00%** in one instance (exact match),
- **1.00%** in another,
- And **2.86%** in a more complex third case.

These experiments demonstrate the model's ability to normalize different types and problem complications, which support its gratitude for real -world -routing logistics.

5. DISCUSSION

Results achieved from the evaluation of the proposed reinforcement learning structure confirm its efficiency in solving travel sellers with drones (TSP-D). Periodically over both unlimited and limited drone -series scenarios, performed the model competitive performance, which produces carefully solutions to optimal over time. This emphasizes the ability of proximal political adaptation (PPO) to generalize well in different delivery configurations. Integration of an improved state vector-inclined drone battery's deficiency and real-time access through the algorithm of Dijkstra is necessary to make effective routing decisions. Residual deep network architecture supported further stable learning and convergence, even in examples of major problems. Compared to traditional metaheuristic algorithms, the proposed method not only achieved high accuracy, but also provided real -time response ability, which is a significant requirement in dynamic distribution systems. The model also maintained high performance during realistic obstacles, and demonstrated the adaptation and strength required for Real -world logistics applications. In addition, low different values indicate many problem sizes that the agent effectively learned the global pattern in distribution optimization rather than overfiting for specific examples. Rapid estimation time, often within seconds, suggests strong capacity for integration into operational logistics platforms that require aircraft decisions. The results of the model of well-known response verification cases further support its generalization capacity, especially under drone energy limits where many classic methods require failure or wide setting. By learning how to dynamically adapt decisions based on the position of the vehicle and environmental reaction, the agent's learning reinforcement refers to the necessary flexibility in smart urban distribution systems. These findings show the promise of learning deep reinforcement as a scalable, adaptable and effective solution for complex route adjustment problems, especially with air and ground-based vehicles in the hybrid delivery network.

CONCLUSION

This research suggested and validated the structure to learn a new reinforcement to solve the travel seller problem with a complex and fast relevant problem drone (TSP-D) when it comes to adapting final meals. Integration of truck and drone coordination presents important calculation and logistical challenges due to the strange nature of the vehicle, obstacles in the dynamic nature of the drone area.

To solve these challenges, we worked out TSP-D as a Markov-declining process and developed an advanced agent based on a proximal political adaptation (PPO) combined with a hybrid deep remaining forward nerve network. A major contribution to this task is inherent in expanded state representation, including environmental dynamics, drone battery levels and viable routes designed by using the algorithm of Dijkstra. This allows the extensive design agent to make intelligent routing decisions that are compatible with real -time obstacles.

Experimental results in the scenarios with both unlimited and limited drone areas show that the proposed model consistently performs better in the context of solution quality and calculation time for the proposed classic methoristic algorithms such as genetic algorithms (GA), particle flock adaptation (Pso) and Grasshore Optimization. In addition, the proposed structure also received minimum interval values for complex problem examples, showing its strength and scalability.

Conclusions emphasize the opportunity to learn reinforcement in logistics optimization in the real world, especially for systems that include coordination of several agents and resource -composed operations. The proposed method not only provides high quality solutions, but also calculation efficiency, which is suitable for distribution in smart logistics platforms where delivery time, adaptability and energy efficiency are important.

Future research can expand this work by integrating coordination of multiple drains, learning from sensor data from the real world, or by incorporating stochastic elements such as weather changes and traffic delays. In addition, a combination of learning reinforcement with future indication models or transfer may increase performance in large -scale or unseen distribution networks further

REFERENCES

[1] Murray, C.C., & Chu, A.G. (2015). The flying sidekick traveling salesman problem: Optimization of drone-assisted parcel delivery. *Transportation Research Part C*, 54, 86–109.

[2] Poikonen, S., Wang, X., & Golden, B. (2017). The vehicle routing problem with drones: Extended models and connections. *Networks*, 70(1), 34–43.

[3] Sutton, R.S., & Barto, A.G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.

[4] Nazari, M., Oroojlooy, A., Snyder, L.V., & Takác, M. (2018). Reinforcement learning for solving the vehicle routing problem. *NeurIPS Workshop*.

[5] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *CVPR*, 770–778.

[7] Lillicrap, T.P., et al. (2015). Continuous control with deep reinforcement learning. *arXiv:1509.02971*.

[8] Vázquez, M., & Vieira, J. (2022). Multi-agent path finding with drones using DRL. *Robotics and Autonomous Systems*, 145, 103891.

[9] Tang, Y., et al. (2020). Reinforcement learning-based multi-agent delivery scheduling with drones. *Journal of Intelligent Transportation Systems*, 24(5), 408–421.

[10] Ma, Y., et al. (2021). Deep reinforcement learning for real-time UAV path planning in dynamic environments. *Sensors*, 21(14), 4704.

[11] Zhou, Y., et al. (2019). A hybrid heuristic for the TSP-D with time windows. *Computers & Operations Research*, 106, 1–16.

[12] Ferrandez, S.M., et al. (2016). Optimization of a truck-drone in tandem delivery network using k-means and genetic algorithm. *Journal of Industrial Engineering and Management*, 9(2), 374–388.

[13] Ulmer, M.W., et al. (2020). Dynamic TSP with stochastic customers and deliveries by drones. *European Journal of Operational Research*, 282(3), 1004–1016.

[14] Macrina, G., et al. (2020). A rich vehicle routing problem with drones and time windows. *Computers & Industrial Engineering*, 144, 106408.

[15] Kim, S., et al. (2018). Delivery planning with drones and trucks. *Transportation Research Part C: Emerging Technologies*, 86, 1–13.

[16] Karak, A., et al. (2017). Coordinated logistics with truck and drone: A deep reinforcement learning approach. *INFORMS Annual Meeting*.

[17] Liu, M., et al. (2021). An adaptive large neighborhood search for the TSP with drones. *Expert Systems with Applications*, 164, 113872.

[18] Khouadjia, M.R., et al. (2020). An evolutionary algorithm for the vehicle routing problem with drone deliveries. *Journal of Heuristics*, 26, 745–776.

[19] Yu, Y., et al. (2022). Multi-agent reinforcement learning for collaborative truck-drone delivery. *Applied Intelligence*, 52, 10381–10395.

[20] Jawhar, I., et al. (2019). UAVs for smart cities: Opportunities and challenges. *Ad Hoc Networks*, 94, 101939.

[21] Li, C., et al. (2023). Real-time UAV delivery route optimization based on deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*.