**Digitus : Journal of Computer Science Applications**
E-ISSN : 3031-3244
Volume. 3, Issue 2, April 2025
Page No: 92-104

# Generalizable and Energy Efficient Deep Reinforcement Learning for Urban Delivery Robot Navigation

**Samroh[1], Era Sari Munthe[2]**
**[1]STMIK Mercusuar, Indonesia**
**[2]Universitas Jayabaya, Indonesia**
Correspondent: samroh74@gmail.com[1]

Received : March 1, 2025
Accepted : April 15, 2025
Published : April 30, 2025

Citation: Samroh., Munthe, E, S. (2025). Generalizable and Energy Efficient Deep Reinforcement Learning for Urban Delivery Robot Navigation. Digitus : Journal of Computer Science Applications, 3 (2), 92-104.

**ABSTRACT:** The increasing demand for contactless urban logistics has driven the integration of autonomous delivery robots into real world operations. This study investigates the application of Deep Reinforcement Learning (DRL) to enhance robot navigation in complex urban environments, focusing on three advanced models: MODSRL, SOAR RL, and NavDP. MODSRL employs a multi objective framework to balance safety, efficiency, and success rate. SOAR RL is designed to handle high obstacle densities using anticipatory decision making. NavDP addresses the sim to real gap through domain adaptation and few shot learning. The models were trained and evaluated in simulation environments (CARLA, nuScenes, Argoverse) and validated using real world deployment data. Evaluation metrics included success rate, collision frequency, and energy efficiency. MODSRL achieved a 91.3% success rate with only 4.2% collision, outperforming baseline methods. SOAR RL showed robust performance in obstacle rich scenarios but highlighted a safety efficiency trade off. NavDP improved real world success rates from 50% to 80% with minimal adaptation data, demonstrating the feasibility of sim to real transfer. The results confirm the effectiveness of DRL in advancing autonomous delivery navigation. Integrating domain generalization, hybrid learning, and real time adaptation strategies will be essential to support large scale urban deployment. Future research should prioritize explainability, continual learning, and user centric navigation policies.

**Keywords:** Reinforcement Learning; Autonomous Delivery, Urban Navigation, Sim To Real Transfer, Multi Objective Learning, Domain Adaptation, Energy Efficiency.



## INTRODUCTION

The global expansion of autonomous delivery robots represents one of the most transformative trends in urban logistics. This acceleration is driven by growing consumer demand for contactless and efficient last mile delivery solutions, particularly in the wake of the COVID 19 pandemic. With

increasing congestion in urban centers and the exponential rise in online retail, the role of autonomous delivery technologies has become increasingly vital. Recent industry data highlights a projected market valuation of approximately $1.39 billion by 2028, supported by a compound annual growth rate (CAGR) exceeding 24% (S. Xu et al., 2023). This signals a pronounced shift in how urban goods are moved, particularly at the last mile stage, where efficiency, responsiveness, and scalability are paramount.

Companies such as Starship and Nuro have emerged as leaders in this domain. Starship, for instance, has partnered with local businesses and academic institutions to integrate small delivery robots capable of transporting food, groceries, and packages within neighborhoods and campuses (Xia & Mei, 2024). Nuro has deployed compact, driverless delivery vehicles in suburban and urban residential settings, emphasizing traffic law compliant navigation for food and grocery delivery (Xu et al., 2023). Their deployments underscore the viability of autonomous agents operating in real world environments and further demonstrate industry confidence in the technological readiness of these systems.

Despite this growth, urban navigation remains a significant hurdle. Cities are inherently unpredictable: high pedestrian density, erratic traffic patterns, cyclists, sudden road closures, and construction zones pose continual obstacles. The navigation task for delivery robots in such contexts is not merely about point to point movement but requires robust situational awareness and dynamic responsiveness. Traditional rule based planning methods, which depend on static maps and deterministic heuristics, offer partial utility in structured environments but have proven insufficient for complex urban settings. While effective under predictable traffic flows, these systems often falter when confronted with spontaneous or rare events, leading to delays, route failures, or safety concerns (Lee et al., 2024; Zhang et al., 2020).

A core limitation of conventional planning algorithms is their dependency on pre mapped environments, which do not account for evolving conditions. Static maps cannot adapt to sudden obstructions or atypical pedestrian behavior. Additionally, such systems typically incur substantial computational overhead, reducing their responsiveness and making them unsuitable for real time operations essential to last mile logistics (Zhang et al., 2020). Moreover, their rigidity precludes effective interaction with dynamic human agents, a key consideration in public urban settings.

These constraints have motivated a shift toward learning based navigation paradigms, most notably Reinforcement Learning (RL) and Deep Reinforcement Learning (DRL). Unlike rule based systems, RL enables agents to learn optimal policies through trial and error interactions with the environment, adapting behavior to maximize cumulative rewards. In urban delivery contexts, this allows for the development of autonomous systems that learn to balance competing objectives such as safety, speed, and pedestrian compliance based on real world data (Daranda & Dzemyda, 2022).

DRL, an extension of RL utilizing deep neural networks, further enhances this adaptability by enabling policy learning in high dimensional state spaces characteristic of real world environments. The integration of DRL with urban navigation has been shown to yield significant improvements

over traditional planners. Notably, DRL models can incorporate temporal dependencies and multi modal sensor inputs, allowing for more nuanced decision making in unpredictable settings (Karnan et al., 2022). Techniques such as experience replay and prioritized sampling improve learning efficiency, while policy regularization helps maintain safe and socially compliant behavior across diverse scenarios (Gao et al., 2024).

Moreover, recent advances in hierarchical DRL have introduced modular architectures capable of decomposing complex tasks into structured sub tasks, enhancing both interpretability and robustness. This hierarchical approach is particularly useful in navigation systems, as it allows agents to plan at multiple time scales short term obstacle avoidance and long term route optimization thereby achieving smoother and more reliable performance (Gao et al., 2024).

The empirical benchmarking of DRL systems against conventional models has further solidified their potential. Numerous studies have demonstrated that DRL based navigation consistently outperforms static planning methods in terms of task success, safety, and adaptability to unforeseen events (Yang et al., 2019). Importantly, DRL's capacity for continual learning enables long term improvements post deployment, a critical feature for systems operating in ever changing urban conditions. Benchmarking provides a rigorous foundation for comparative performance evaluation and supports the generalization of findings across different operational domains and datasets (Yang et al., 2019).

Nevertheless, despite the advancements in simulation and controlled testing, the transition from simulated performance to real world deployment remains a core challenge. This sim to real gap is driven by discrepancies in sensor noise, actuation models, environmental stochasticity, and visual domain shifts. Robots trained entirely in simulation often experience degraded performance when faced with real world complexities not captured during training. To address this, researchers have begun integrating domain adaptation strategies and few shot learning techniques to allow for fine tuning using minimal real world data (Daranda & Dzemyda, 2022). Such hybrid approaches mitigate the limitations of zero shot transfer and help bridge the performance gap in real deployments.

In conclusion, the landscape of autonomous delivery robotics is evolving rapidly, propelled by urbanization, technological maturity, and commercial momentum. Companies like Starship and Nuro exemplify how these technologies can be operationalized at scale, but significant scientific and engineering challenges persist. Among these, navigation in complex urban environments remains at the forefront. Traditional planning methods are increasingly being supplanted by reinforcement learning based approaches due to their adaptability, scalability, and data driven foundations. This study seeks to advance the field by evaluating cutting edge DRL architectures with an emphasis on safety, generalizability, and real world transfer. By benchmarking these systems across simulation and real world data, we aim to illuminate pathways for the reliable, scalable deployment of autonomous navigation in urban delivery robotics.

## METHOD

To evaluate the performance of reinforcement learning (RL) models for urban robot navigation, this study employs a set of standardized and realistic simulation environments. Among these, CARLA (Car Learning to Act) serves as a primary platform due to its ability to emulate diverse urban scenarios. CARLA provides high fidelity city maps, dynamic traffic actors, and realistic pedestrian models, thereby offering a safe yet comprehensive testbed for autonomous driving systems (Jiang et al., 2021). Similarly, the nuScenes dataset is used for its extensive real world urban sensor data, including annotated lidar, radar, and camera inputs. This choice is motivated by its representativeness of dense urban traffic conditions, making it particularly effective for testing perception and localization capabilities in complex cityscapes where accurate sensor fusion is critical (Jiang et al., 2021). Additionally, Argoverse is incorporated to validate performance on map centric urban mobility tasks.

These simulation platforms not only replicate physical conditions such as weather, road textures, and obstacle layouts but also allow the modeling of real time interactions between multiple agents. They offer an effective compromise between risk free experimentation and environmental realism, making them highly suitable for developing and testing deep reinforcement learning (DRL) algorithms in urban delivery contexts (Scheikl et al., 2023).

The study evaluates three key DRL architectures: MODSRL, SOAR RL, and NavDP. MODSRL is a multi objective framework that balances competing goals such as safety, speed, and path efficiency. SOAR RL is designed for dynamic obstacle rich scenarios and employs reactive decision making with predictive modeling for obstacle trajectories. NavDP incorporates domain adaptation strategies for transferring navigation policies from simulated to real environments using minimal in domain training data.

Each model is trained using episodic reinforcement learning with a reward function that integrates multiple metrics, including distance to goal, number of collisions, energy consumed, and adherence to social norms. The policies are optimized using proximal policy optimization (PPO) or soft actor critic (SAC), depending on the architecture's complexity. Hierarchical components are included in some models to separate high level goal planning from low level motor control.

A comprehensive set of performance metrics is applied to benchmark the navigation capabilities of each model. The success rate is the proportion of completed navigation tasks without failures or deviations. The collision rate quantifies the frequency of impacts with static or dynamic obstacles. Path efficiency measures the actual travel distance relative to the optimal path, while average task time assesses the timeliness of navigation execution (Elsken et al., 2019).

Additional metrics include energy consumption (kJ) calculated over each episode and user related indicators such as comfort or safety compliance in multi agent scenarios. These metrics are chosen to ensure holistic evaluation and are consistent with prior DRL benchmarking standards (Elsken et al., 2019).

A key component of this methodology is the transition from simulation trained policies to real world deployment. Given the inherent domain shift between simulated and physical environments stemming from sensor noise, lighting conditions, actuation discrepancies, and unmodeled events domain adaptation strategies are crucial.

This study employs several adaptation approaches. First, domain adversarial training is used to minimize the divergence between source (simulation) and target (real world) feature distributions. The objective is to learn domain invariant representations that generalize across different operational conditions (Hou et al., 2021; Tonioni et al., 2019). This is achieved through auxiliary discriminators that penalize domain specific cues in the latent feature space.

Second, few shot learning is integrated to enable the models to adapt quickly using minimal real world data samples. Meta learning algorithms are applied to leverage prior experience across tasks, thus enabling efficient transfer even in the presence of limited new domain exposure (Bonardi et al., 2020; Zhao et al., 2020). Such techniques are essential in urban navigation tasks, where exhaustive real world data collection is impractical.

Third, SimGAN and similar generative approaches are employed to augment simulation realism. By translating synthetic frames into photo realistic outputs, these techniques reduce the perceptual gap during inference and improve the robustness of perception modules (Jiang et al., 2021).

Training is conducted in simulated environments using curriculum learning to progressively introduce complexity, such as dynamic pedestrian flows, irregular vehicle patterns, and environmental noise. Episodes are capped at a maximum step count or until task completion/failure. Models are evaluated across three conditions: (1) in distribution simulation, (2) out of distribution simulation, and (3) real world settings (post adaptation).

For real world testing, limited field trials are conducted using pre mapped urban areas (e.g., campus zones and residential streets with controlled pedestrian traffic). Policy deployment is performed on embedded compute platforms, such as Jetson AGX, which support onboard inference and logging.

The models are implemented using Python and PyTorch. Training is distributed across multi GPU clusters, and simulation interfacing is handled through ROS and OpenAI Gym compatible wrappers. Hyperparameters and environment configurations are made publicly available for reproducibility.

In summary, this methodology combines high fidelity simulations, robust DRL architectures, standard benchmarking metrics, and advanced adaptation strategies to ensure both rigorous testing and real world applicability. The integration of domain adaptation and few shot learning significantly enhances the transferability of trained models, paving the way for scalable autonomous navigation in complex urban landscapes.

## RESULT AND DISCUSSION

### Performance of MODSRL Compared to Baseline Models

The Multi Objective Deep Reinforcement Learning (MODSRL) framework demonstrated significant improvements in navigation outcomes when benchmarked against three established baseline models: Collaborative Active Deep Reinforcement Learning (CADRL), Omnidirectional Multi Agent Reinforcement Learning (OM SARL), and Long Short Term Memory based Reinforcement Learning (LSTM RL). Unlike single objective architectures, MODSRL simultaneously optimizes for multiple goals, including task completion, collision avoidance, and energy efficiency (Xu et al., 2020). This multi objective approach allows MODSRL to dynamically adjust to competing environmental demands by prioritizing safety in dense scenarios and efficiency in sparse conditions.

**Table 1. Performance Comparison of MODSRL and Baseline Models**

| Model | Success Rate (%) | Collision Rate (%) | Notes |
|---|---|---|---|
| **CADRL** | 76.5 | 10.2 | Struggles with coordination in multi agent settings |
| **OM SARL** | 84.9 | – | Limited handling of dynamic agent distribution |
| **LSTM RL** | 82.1 | 8.7 | Computational inefficiency, limited generalization |
| **MODSRL** | **91.3** | **4.2** | Multi objective optimization ensures robustness in urban environments |

As illustrated in Table 1, MODSRL achieved a success rate of 91.3%, which exceeds CADRL by 14.8 percentage points, OM SARL by 6.4, and LSTM RL by 9.2. Furthermore, its collision rate was significantly lower at 4.2%, compared to 10.2% for CADRL and 8.7% for LSTM RL. These differences affirm the benefits of multi objective optimization in enhancing navigation robustness under complex urban conditions (Z. Xu et al., 2020).

The superiority of MODSRL was particularly evident in scenarios involving crowd navigation and dynamic agent distribution. Traditional models struggled with real time coordination, often reacting inadequately to sudden changes in pedestrian flow or traffic congestion. In contrast, MODSRL maintained high performance by adjusting its navigation strategy based on continuous environmental feedback.

### Obstacle Adaptability and Trade offs in SOAR RL

The SOAR RL architecture exhibited effective adaptability in obstacle rich environments, with success rates of 98.0% in sparse conditions and 87.1% in high density scenarios. As shown in Table 2, collision rates rose from 1.2% to 13.2% as the number of dynamic obstacles increased. These results highlight the scalability of SOAR RL but also reveal the cost of maintaining safety in increasingly constrained spaces.

**Table 2. SOAR RL Performance in Varying Obstacle Densities**

| Condition | Success Rate (%) | Collision Rate (%) | Notes |
|---|---|---|---|
| **Sparse Obstacles** | 98.0 | 1.2 | Highly effective in low density environments |
| **High Density Obstacles** | 87.1 | 13.2 | Demonstrates scalability but with higher collision trade offs |

SOAR RL's robustness stems from its combination of reactive and anticipatory learning components. The model incorporates trajectory prediction for surrounding agents, enabling proactive adjustments that mitigate the risk of collision. This dual layer approach allows the robot to navigate with foresight, a crucial advantage in environments with unpredictable obstacle dynamics.

However, SOAR RL also illustrates the fundamental trade off in reinforcement learning between safety and efficiency. For example, in high pedestrian density zones, conservative policies reduced collisions by 35% compared to aggressive strategies, but at the expense of a 20% increase in travel time and a 15% rise in energy use. Conversely, aggressive policies shortened routes but introduced higher collision likelihood, underscoring the need for balanced hybrid approaches (Jiang et al., 2021).

### Sim to Real Transfer and Domain Adaptation via NavDP

The NavDP model addresses one of the most persistent challenges in robotics: the sim to real transfer. Zero shot RL, while highly successful in simulation (99.8%), showed a sharp drop to 67.0% success in real world deployment. By contrast, NavDP, when combined with few shot fine tuning using minimal real world data, raised the real world success rate from 50.0% to 80.0% (see Table 3). This 30 percentage point gain underscores the effectiveness of domain adaptation techniques in real world applications.

**Table 3. Sim to Real Transfer Performance of NavDP**

| Model | Simulation Success (%) | Real World Success (%) | Notes |
|---|---|---|---|
| **Zero shot RL** | 99.8 | 67.0 | High simulation accuracy, but poor real world transferability |
| **Baseline RL** | – | 50.0 | Without adaptation, performance is limited |
| **NavDP + Few shot Fine tuning** | – | **80.0** | Domain randomization and adaptation improve transfer effectiveness |

NavDP leverages domain randomization and adversarial training to produce domain invariant representations, enabling more generalized navigation policies. Domain randomization introduces

controlled variability into the simulation altering lighting, textures, and motion dynamics while adversarial domain adaptation techniques ensure that learned features align with real world sensory inputs (Tonioni et al., 2019).

The integration of few shot learning allows the model to rapidly adapt to novel environments with limited data. Meta learning frameworks underpin this adaptability by drawing on prior simulation experience to guide new learning tasks (Bonardi et al., 2020). These strategies significantly reduce the deployment burden and data collection costs, thus enabling scalable deployment across diverse environments.

Datasets such as the Robotic Room Dataset (RRD) and environment specific variations of CARLA were used to evaluate these transfers. These datasets simulate key real world conditions, including dynamic interactions, occlusions, and path obstructions, providing a rigorous benchmark for assessing real world transferability.

### Energy Efficiency Across DRL Models and Environments

The final performance dimension assessed in this study is energy efficiency. MODSRL, SOAR RL, and NavDP were evaluated in three simulation environments nuScenes, Argoverse, and CARLA with energy consumption recorded in kilojoules per episode (Table 4).

**Table 4. Energy Efficiency Across DRL Models and Environments**

| Model | Environment | Avg. Energy Consumption (kJ) | Success Rate (%) | Collision Rate (%) | Notes |
|---|---|---|---|---|---|
| **MODSRL** | nuScenes | 11.8 | 92.5 | 2.5 | Multi objective optimization includes energy |
| **SOAR RL** | Argoverse | 13.0 | 89.0 | – | Slightly higher energy use |
| **NavDP** | CARLA | 12.4 | 91.3 | – | Balanced energy performance trade off |

MODSRL achieved the highest overall energy efficiency with an average consumption of 11.8 kJ in nuScenes while maintaining a success rate of 92.5% and collision rate of 2.5%. In contrast, SOAR RL, tested in Argoverse, consumed slightly more energy (13.0 kJ) with a success rate of 89.0%. NavDP, evaluated in CARLA, reported an average energy use of 12.4 kJ alongside a success rate of 91.3%.

These results confirm that DRL models can be tuned not only for navigation accuracy but also for energy optimization. MODSRL's multi objective design, which incorporates energy as an explicit optimization goal, is particularly advantageous for resource constrained robotic systems (Z. Xu et al., 2020). However, trade offs between energy use and performance remain. Energy efficient strategies may extend travel times and reduce task throughput, while performance optimized models might incur higher energy costs.

Simulation environments like CARLA offer embedded energy models that estimate consumption based on motion dynamics, actuator profiles, and battery simulation (Jiang et al., 2021). These tools allow for comprehensive evaluations of navigation strategies across both operational and energy efficiency metrics, aligning technical feasibility with sustainability considerations in urban robotics.

In summary, the results demonstrate that multi objective DRL models such as MODSRL consistently outperform traditional and single objective approaches across safety, success rate, and energy efficiency. SOAR RL offers scalable safety performance in dynamic environments but must manage safety efficiency trade offs. NavDP effectively bridges the sim to real gap, and energy benchmarking reinforces the relevance of efficient planning in deployment critical robotics applications. Together, these findings validate the efficacy of DRL in autonomous delivery and suggest clear pathways for real world deployment.

### Scalability Constraints in DRL for Real World Deployments

The application of Deep Reinforcement Learning (DRL) in urban delivery robotics holds substantial promise, particularly in enhancing autonomous navigation capabilities in complex environments. However, the results presented in this study also reveal a set of persistent limitations that must be addressed to facilitate broader and more reliable deployment. One of the most significant barriers to the real world scalability of DRL models lies in their high computational demands. Training these models involves extensive iteration across large state action spaces, requiring substantial computational resources that may not be readily accessible in practical deployment settings. This issue is particularly pronounced in urban environments, where the navigation system must adapt to rapidly changing variables, such as pedestrian density, traffic congestion, and environmental occlusions (Yingjie et al., 2025). The resulting discrepancy between simulated and real world dynamics can lead to degraded performance when models are deployed in unpredictable conditions.

### Performance vs. Generalization Trade offs

Moreover, while MODSRL and SOAR RL demonstrate strong performance in simulation, their applicability in live scenarios hinges on generalization. Many DRL models, when tuned for maximum performance in a specific domain, fall victim to overfitting. This effect, while yielding high success rates in controlled simulations, often results in poor transferability to new environments. The problem is exacerbated in urban settings, where human behavior, infrastructural layout, and environmental stimuli vary widely. This tension between performance optimization and generalization represents a critical challenge in current DRL design (Yingjie et al., 2025). Approaches such as regularization, ensemble learning, and dropout have been introduced to increase model robustness, while domain adaptation and transfer learning techniques have become standard for mitigating the negative effects of overfitting (Wang et al., 2024).

### Hybrid Learning Architectures for Obstacle Rich Environments

In terms of obstacle navigation and real time reactivity, SOAR RL offers clear advantages by blending reactive and predictive capabilities. Nevertheless, its performance degrades as obstacle

density increases, revealing a scalability limit when managing complex real time interactions. This limitation signals a need for further exploration into hybrid models that integrate both model based and model free learning. Such approaches may enable DRL systems to leverage the strengths of deterministic planning while maintaining the flexibility and adaptability of learning based techniques (Vieira et al., 2025). Additionally, hybrid architectures may be better suited to incorporate explicit environmental constraints and deliver more interpretable behavior, which is crucial for deployment in public spaces.

### Model Explainability and Public Trust

Interpretability itself is another critical concern. Despite their functional effectiveness, DRL models often function as black boxes, making their decision making processes difficult to audit or explain to stakeholders. In the context of urban delivery, this opacity may hinder regulatory approval and reduce public trust, particularly in cases involving near misses with pedestrians or unexpected detours. Therefore, explainability frameworks tailored to urban navigation such as trajectory heatmaps or causal policy explanations are essential for building transparency and accountability (Whittlestone et al., 2021). Enhancing explainability will be especially important as DRL systems are deployed in multi agent contexts, where coordination with humans and other robots must be seamless and predictable.

### Effective Transfer Learning with NavDP

The success of NavDP in bridging the sim to real gap further reinforces the value of domain adaptation and few shot learning strategies. Zero shot DRL, while efficient in simulation, cannot accommodate the nuanced discrepancies that arise in real world contexts. By contrast, NavDP's ability to adapt using limited real world data sets a precedent for scalable deployment across diverse environments. This also emphasizes the importance of robust benchmarking datasets that include realistic interactions and sensor noise, as these components are essential to train and evaluate models that are not only high performing but also resilient.

### Meta  and Continual Learning for Real Time Adaptation

From a forward looking perspective, several directions emerge as particularly promising. First, the integration of meta learning techniques into DRL pipelines could enhance the adaptability of these systems in real time. Meta learning allows models to learn how to learn, thereby reducing the time and data required to acclimate to new tasks or environments (Yingjie et al., 2025). This capability is critical in urban logistics, where delivery routes, obstacles, and customer preferences can change dynamically. Second, continual learning frameworks can ensure that deployed DRL systems remain up to date by incorporating new information over time without catastrophic forgetting. When paired with online learning strategies, these models can evolve alongside their operating environments, maintaining relevance without exhaustive retraining cycles (Bridgelall, 2024).

### User Centric and Collaborative Learning Strategies

Moreover, future DRL integration efforts should emphasize user centric navigation strategies. Learning systems that can adapt to customer preferences such as delivery time windows, noise sensitivity, or accessibility constraints stand to improve not only operational efficiency but also user satisfaction. Incorporating consumer behavioral data into the reward function may help in

aligning navigation policies with end user expectations, thereby enhancing the commercial viability of autonomous delivery systems.

Finally, collaborative learning techniques represent a compelling avenue for improving generalization and robustness. By enabling multiple delivery robots to share experiences and policy updates, these methods facilitate distributed learning and accelerated adaptation. Shared learning environments can enable broader pattern recognition across deployments, ultimately leading to more generalized navigation policies that are effective across varied geographies and user demographics.

## CONCLUSION

This study demonstrates that Deep Reinforcement Learning (DRL) architectures—MODSRL, SOAR RL, and NavDP—offer substantial advantages for autonomous urban delivery navigation. Compared with conventional methods, these models deliver higher success rates, improved safety, and greater energy efficiency. MODSRL proved robust in multi-agent environments, SOAR RL highlighted the trade-off between safety and efficiency in dense obstacle scenarios, and NavDP effectively bridged the sim-to-real gap through domain adaptation and few-shot learning.

The findings underscore both the promise and challenges of DRL for large-scale deployment. Persistent issues such as sim-to-real transfer gaps, model interpretability, and real-time scalability require further exploration through hybrid learning, meta-learning, and collaborative strategies. Future research should prioritize explainable and user-centric DRL approaches to build public trust and regulatory acceptance, ensuring delivery robots evolve into reliable, safe, and sustainable actors within the urban logistics ecosystem.

## REFERENCE

Bonardi, A., James, S., & Davison, A. J. (2020). Learning One-Shot Imitation From Humans Without Humans. Ieee Robotics and Automation Letters, 5(2), 3533–3539. https://doi.org/10.1109/lra.2020.2977835

Bridgelall, R. (2024). Locating Electrified Aircraft Service to Reduce Urban Congestion. Information, 15(4), 186. https://doi.org/10.3390/info15040186

Daranda, A., & Dzemyda, G. (2022). Reinforcement Learning Strategies for Vessel Navigation. Integrated Computer-Aided Engineering, 30(1), 53–66. https://doi.org/10.3233/ica-220688

Elsken, T., Staffler, B., Metzen, J. H., & Hutter, F. (2019). Meta-Learning of Neural Architectures for Few-Shot Learning. https://doi.org/10.48550/arxiv.1911.11090

Gao, Y., Wu, J., Yang, X., & Ji, Z. (2024). Efficient Hierarchical Reinforcement Learning for Mapless Navigation With Predictive Neighbouring Space Scoring. Ieee Transactions on Automation Science and Engineering, 21(4), 5457–5472. https://doi.org/10.1109/tase.2023.3312237

Hou, Y., Lai, Y., Chen, C., Che, W., & Liu, T. (2021). Learning to Bridge Metric Spaces: Few-Shot Joint Learning of Intent Detection and Slot Filling. 3190–3200. https://doi.org/10.18653/v1/2021.findings-acl.282

Jiang, Y., Zhang, T., Ho, D. E., Bai, Y., Liu, C. K., Levine, S., & Tan, J. (2021). SimGAN: Hybrid Simulator Identification for Domain Adaptation via Adversarial Reinforcement Learning. 2884–2890. https://doi.org/10.1109/icra48506.2021.9561731

Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J. L., Biswas, J., & Stone, P. (2022). Socially Compliant Navigation Dataset (SCAND): A Large-Scale Dataset of Demonstrations for Social Navigation. https://doi.org/10.48550/arxiv.2203.15041

Lee, J., Bjelonic, M., Reske, A., Wellhausen, L., Miki, T., & Hutter, M. (2024). Learning Robust Autonomous Navigation and Locomotion for Wheeled-Legged Robots. Science Robotics, 9(89). https://doi.org/10.1126/scirobotics.adi9641

Scheikl, P. M., Tagliabue, E., Gyenes, B., Wagner, M., Dall'Alba, D., Fiorini, P., & Mathis-Ullrich, F. (2023). Sim-to-Real Transfer for Visual Reinforcement Learning of Deformable Object Manipulation for Robot-Assisted Surgery. Ieee Robotics and Automation Letters, 8(2), 560–567. https://doi.org/10.1109/lra.2022.3227873

Tonioni, A., Rahnama, O., Joy, T., Stefano, L. D., Ajanthan, T., & Torr, P. H. S. (2019). Learning to Adapt for Stereo. https://doi.org/10.1109/cvpr.2019.00989

Vieira, M., Vieira, M. A., Galvão, G., Louro, P., Fantoni, A., Vieira, P., & Véstias, M. (2025). Enhancing Airport Traffic Flow: Intelligent System Based on VLC, Rerouting Techniques, and Adaptive Reward Learning. Sensors, 25(9), 2842. https://doi.org/10.3390/s25092842

Wang, C., Chen, J., & Yu, X. (2024). Research on Benefit Allocation Based on Multi-Weight H-Shapley Value: A Case Study of Express Logistics Sharing. Plos One, 19(7), e0305656. https://doi.org/10.1371/journal.pone.0305656

Whittlestone, J., Arulkumaran, K., & Crosby, M. (2021). The Societal Implications of Deep Reinforcement Learning. Journal of Artificial Intelligence Research, 70. https://doi.org/10.1613/jair.1.12360

Xia, Y., & Mei, C. (2024). Integrating UAS for 3D Terrain Mapping and Autonomous Navigation: A Review of Multi-Camera and Reinforcement Learning. https://doi.org/10.31219/osf.io/z7mdv

Xu, S., Hao, J., Xue-mei, C., & Hu, Y. (2023). Navigating Autonomous Vehicles in Uncertain Environments With Distributional Reinforcement Learning. Proceedings of the Institution of Mechanical Engineers Part D Journal of Automobile Engineering, 238(12), 3653–3663. https://doi.org/10.1177/09544070231186841

Xu, Z., Wu, K., Che, Z., Tang, J., & Ye, J. (2020). Knowledge Transfer in Multi-Task Deep Reinforcement Learning for Continuous Control. https://doi.org/10.48550/arxiv.2010.07494

Yang, Y., Bevan, M. A., & Li, B. (2019). Efficient Navigation of Colloidal Robots in an Unknown Environment via Deep Reinforcement Learning. Advanced Intelligent Systems, 2(1). https://doi.org/10.1002/aisy.201900106

Yingjie, Z., Hasan, W. Z. W., Ramli, H. R., Norsahperi, N. M. H., Kassim, M. S. M., & Yao, Y. (2025). Deep Reinforcement Learning of Mobile Robot Navigation in Dynamic Environment: A Review. Sensors, 25(11), 3394. https://doi.org/10.3390/s25113394

Zhang, H., Song, M., & He, H. (2020). Achieving the Success of Sustainability Development Projects Through Big Data Analytics and Artificial Intelligence Capability. Sustainability, 12(3), 949. https://doi.org/10.3390/su12030949

Zhao, P., Ram, P., Lu, S., Yao, Y., Bouneffouf, D., Lin, X., & Liu, S. (2020). Learning to Generate Image Source-Agnostic Universal Adversarial Perturbations. https://doi.org/10.48550/arxiv.2009.13714