Advances in Science and Technology Research Journal

Advances in Science and Technology Research Journal, 2026, 20(5), 372–385
https://doi.org/10.12913/22998624/216568
ISSN 2299-8624, License CC-BY 4.0

Received: 2025.11.24
Accepted: 2026.03.15
Published: 2026.04.01

Adaptive traffic signal control using proximal policy optimization transformer and radio frequency identification-based vehicle detection in the simulation of urban mobility environment

Daniel Kleczyński[1*], Jakub Drzał[2], Bartosz Pawłowicz[3]

[1] The Faculty of Electrical and Computer Engineering, Rzeszów University of Technology, Aleja Powstańców Warszawy 12, Rzeszów, Poland
[2] Department of Metrology and Diagnostic Systems, Rzeszów University of Technology, Aleja Powstańców Warszawy 12, Rzeszów, Poland
[3] Department of Electronic and Telecommunications Systems, Rzeszów University of Technology, Aleja Powstańców Warszawy 12, Rzeszów, Poland
* Corresponding author's e-mail: danielkleczynski@gmail.com

ABSTRACT

This research paper proposed an adaptive traffic signal control method based on proximal policy optimization (PPO) integrated with a transformer architecture, utilizing exclusively radio frequency identification (RFID)-based vehicle detection and type recognition (car, bus, ambulance) within the SUMO simulation environment. RFID readers function as vehicle detectors, generating a stream of aggregated counts in short time windows. This stream is sequentially modeled by the transformer, enabling the PPO policy to capture inflow variability and determine phase maintenance or switching decisions while adhering to safety constraints. The reward function is designed to minimize global travel time and queue lengths. The proposed method was experimentally compared against a classic sequential algorithm and the adaptive Miller algorithm, using identical input data and phase constraints. Results indicate that the PPO-transformer achieves a reduction in average delay by 28.6% and queue lengths by 36.0% compared to the fixed-time baseline. Furthermore, the model outperforms the adaptive Miller algorithm, reducing delays by 9.1% and the number of stops by 9.5%, while simultaneously increasing total intersection throughput by 12% (relative to the fixed-time baseline). Sensitivity analysis demonstrates the robustness of the PPO-transformer to partial RFID read losses and bursty traffic inflows.

Keywords: adaptive traffic signal control, DRL, ITS, PPO, RFID, SUMO, transformer.

INTRODUCTION

Modern traffic management systems face unprecedented challenges resulting from the rapid growth of traffic volume, the increasing heterogeneity of vehicle fleets, and increasingly complex urban mobility patterns. Traditional traffic signal control algorithms, based on fixed time sequences or simple actuation mechanisms, prove insufficient in the face of dynamically changing traffic conditions [1].

This article presents a comparison of a classic control system with an advanced solution using deep reinforcement learning (DRL). The concept of the considered control problem is illustrated in Figure 1. It depicts the feedback loop present within the system: the simulator generates traffic congestion, which is captured by the detection system (Sensor System), and based on this data, the model makes decisions regarding signal actuation.

Addressing the challenges visualized in Figure 1, DRL emerges as a natural evolution of traffic control systems. This approach, combined with precise detection, offers key advantages in terms of scalability and adaptability. A primary asset of this solution is automatic feature extraction. Unlike rule-based systems, DRL agents automatically learn state representations and

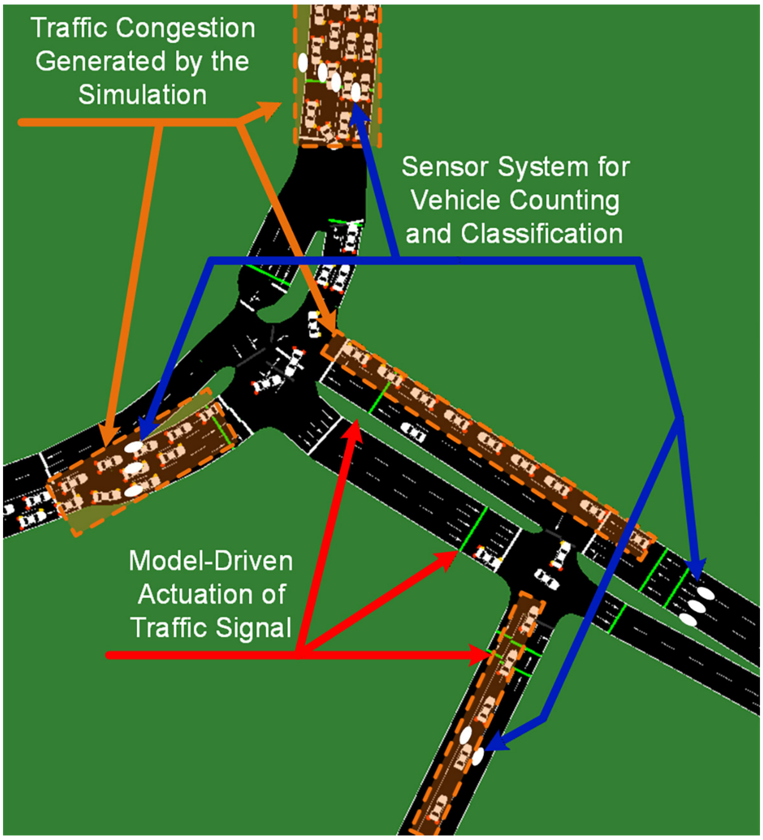
Traffic Congestion Generated by the Simulation
Sensor System for Vehicle Counting and Classification
Model-Driven Actuation of Traffic Signal

Figure 1. Concept of the closed-loop control system – traffic congestion generated by the simulation is monitored by the vehicle classification and counting system (sensor system), which serves as input for the model-driven actuation of the traffic signal

optimal action policies directly from the data provided by the sensor system. This eliminates the need for manual coding of all possible scenarios and allows for the discovery of complex, non-linear dependencies in traffic data. Moreover, DRL enables global optimization. Algorithms such as PPO with a transformer architecture can model long-term dependencies and coordinate actions across the entire network. In simulation tests within the SUMO (Simulation of Urban MObility) environment, the PPO-transformer approach achieved a reduction in average delay by 28.6% relative to the fixed-time algorithm, a 12% increase in network throughput, and superior adaptation to variable traffic patterns.

Radio frequency identification (RFID) technology is a key component of the proposed system, providing a robust and privacy-preserving data source for DRL algorithms. Unlike vision-based detection, RFID is resistant to adverse weather (rain, fog, snow), improving the reliability of intelligent transportation systems (ITS) [2]. By aggregating only vehicle presence and type, without images or detailed trajectories, it avoids typical GDPR-related issues while enabling near-100% accuracy in vehicle-type classification (passenger cars, heavy goods vehicles, emergency vehicles). Passive transponders further make RFID a cost-effective alternative to camera systems and induction loops [3].

Aggregated RFID measurements, processed in short time windows, are well suited to transformer-based sequential models, which capture spatiotemporal traffic patterns and determine signal actuation. However, deploying DRL-based control in real urban environments faces key challenges: safety (supervisory mechanisms enforcing minimum and maximum signal times), fault tolerance (operation under partial sensor data loss), and computational complexity (scaling to city-wide networks). As RFID is increasingly deployed in Smart City applications such as logistics, public transport, fleet management, electronic toll collection (ETC), parking management, and delivery monitoring, it creates a natural foundation for ML- and RL-based traffic control. The aim of this work was to build on these developments and propose DRL-driven traffic control algorithms that exploit the growing presence of RFID in urban infrastructure, showing that only the combination of RFID with Deep Reinforcement Learning yields a scalable solution for managing complex transport networks.

IDENTIFICATION, PRIORITIZATION, AND FLEET HETEROGENEITY

The dynamic development of urban areas and the increasing complexity of vehicle fleets impose new, fundamental requirements on traffic control systems. Contemporary systems must manage not only traffic volume but also its diverse composition and changing priorities.

The first key challenge is the need for precise real-time vehicle identification, driven by the expansion of low emission zones (LEZ), known in Poland as clean transport zones (SCT). Their implementation, exemplified by the zone launched in Warsaw on July 1st, 2024, relies on progressively tightened EURO emission standards [4]. This compels traffic control systems to integrate with law enforcement mechanisms (e.g., ANPR cameras or vehicle databases) to not only detect the presence of a vehicle but also verify its access rights.

The second challenge is the growing fleet heterogeneity [5]. Urban traffic is currently a mix of conventional, electric (EV), and in the testing phase autonomous vehicles (AV), alongside a wide spectrum of commercial vehicles [6]. Each of these classes possesses distinct dynamics and requirements. The rising, albeit uneven across Europe, adoption of electric vehicles necessitates the integration of traffic management systems with charging infrastructure and the application of eco-driving algorithms capable of significantly reducing energy consumption [7]. Simultaneously, simulations of the impact of AV indicate the potential for drastic reductions in delays (by as much as 86–91%) and increases in network throughput (by 16–25%) at high penetration levels [8].

The third critical functional requirement is the capability for dynamic prioritization of specific vehicle types. Systems must distinguish between standard traffic signal priority (TSP), applied to public transit or freight transport to extend the green phase, and the overriding emergency vehicle pre-emption (EVP) mode [9]. EVP, by assuming full control over the signal cycle, is crucial for safety and reducing the number of accidents involving emergency vehicles. The effective implementation of these mechanisms requires a transition from point detection (induction loops) to area tracking (GPS, V2X) [10].

In summary, the regulatory needs of these trends for identification (LEZ), dynamic complexity (EV/AV), and operational requirements (EVP/TSP) render the systems based on a single sensor type insufficient. A modern traffic control system must be capable of processing a multi-dimensional state vector, answering not only the question of “if” a vehicle is present, but also “what” kind of vehicle it is, “what” permissions it holds, and “what” its priority is.

In this context, hybrid solutions appear to be the most universal and future-proof. While vision systems based on deep learning (CNN, YOLO) excel at presence detection, queue estimation, and general type classification, RFID technology is indispensable for ensuring identity assurance [6]. It allows for error-free, weather independent distinction of emergency vehicles (EVP) and public transport (TSP), or automatic verification of emission standards (LEZ) for a specific, tagged vehicle capabilities not guaranteed by image analysis alone. It is only the fusion of data from these complementary sources (e.g., vision for queue detection and RFID for precise identification) that creates a complete and reliable state vector. Such a vector can then effectively feed adaptive algorithms (such as DRL), enabling true multi-criteria optimization (emissions, safety, throughput) in such a complex environment [11].

RELATED WORKS

ITS have evolved from simple fixed-time solutions to advanced algorithms based on machine learning and deep learning (Figure 2). This section presents a synthetic review of recent research (2020–2025) on traffic signal control algorithms, with particular emphasis on input data, decision-making mechanisms, and achieved results in the context of specific performance metrics.

Traffic control algorithms can be categorized into several main groups. These include Fixed-time algorithms, based on preset schedules and incapable of adaptation; Adaptive algorithms (Vehicle actuated, SCOOT, SCATS), which utilize detector data to dynamically adjust phases; and Max-Pressure algorithms, which optimize flow by minimizing the differences in vehicle counts

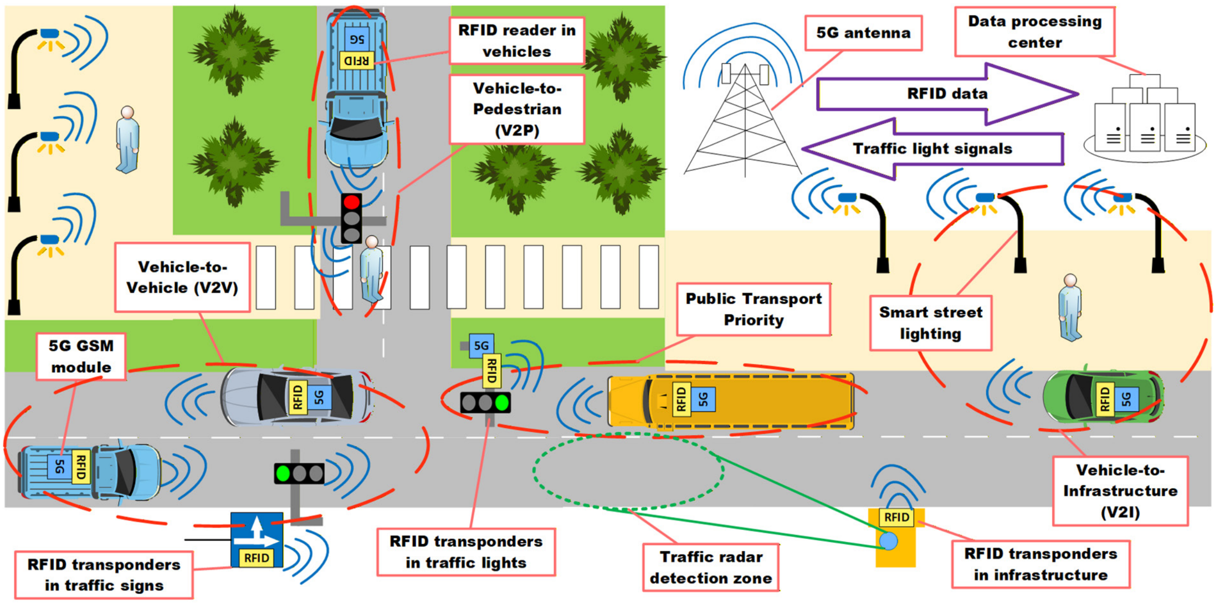
RFID reader in vehicles
Vehicle-to-Pedestrian (V2P)
5G antenna
Data processing center
RFID data
Traffic light signals
Vehicle-to-Vehicle (V2V)
Public Transport Priority
Smart street lighting
5G GSM module
RFID
5G
Vehicle-to-Infrastructure (V2I)
RFID transponders in traffic signs
RFID transponders in traffic lights
Traffic radar detection zone
RFID transponders in infrastructure

Figure 2. Implementation of an ITS system using the RFID technology in a Smart City environment, based on the example of C-V2X (cellular vehicle to everything)

between upstream and downstream approaches. Among newer approaches, reinforcement learning (RL) stands out, learning optimal strategies through interaction with the environment, as well as Evolutionary Algorithms (Genetic Algorithms, Genetic Programming), which seek solutions through selection and mutation mechanisms. These are complemented by Fuzzy Logic algorithms, which handle data uncertainty, and Hybrid algorithms that combine the strengths of multiple approaches.

Table 1 summarizes the input data and control mechanisms for selected algorithms, illustrating the diversity of approaches regarding environmental perception and decision-making. Table 2 presents a detailed comparison of the results achieved by individual algorithms, indicating the evaluation metrics used.

On the basis of the literature review, several key observations can be drawn. Foremost is the dominance of RL and DRL methods (particularly DQN and PPO), which achieve the best results in terms of waiting time reduction (ranging from 27–68%) and demonstrate high adaptability to variable traffic conditions [12]. Simultaneously, Max-Pressure algorithms prove effective, with delay reductions at the level of 15–30%, while their modifications incorporating additional road users enhance their practical utility [13].

In the context of scalability, federated learning (FL) emerges as a promising solution, enabling model training across multiple intersections without data centralization and achieving waiting time reductions exceeding 60% [14]. It is also worth noting the role of evolutionary algorithms (GA, GP), such as GPLight or NSGA2, which offer explainable solutions capable of multi-objective optimization [15]. Conversely, Fuzzy Logic remains a valuable method robust to data uncertainty and changing environmental conditions, reducing waiting time by 16% [16]. Ultimately, hybrid approaches, combining the strengths of RL and Max-Pressure, create a synergy reflected in higher throughput and improved fairness across the network [17].

These observations lead to the formulation of recommendations for future research and deployment. Focus should be placed on utilizing Multi-Agent RL to coordinate entire intersection networks and on integrating data from diverse sources such as cameras, induction sensors, as well as V2X communication to increase the accuracy of environmental state representation. The development of Explainable AI (XAI) methods is also becoming critical to improve trust in autonomous systems. Further validation of algorithms in real-world urban conditions, rather than solely in simulations, is essential. Finally, research should increasingly account for multi-objective optimization, encompassing not only traffic fluidity, but also CO_2 emissions, fuel consumption, and safety.

Recent studies (2022–2025) show that adaptive traffic-signal control methods, like RL/DRL

Table 1. Characteristics of input and output data for traffic control algorithms

Algorithm	Input data	Control (Action)
Deep RL (DQN, PPO)	Queue length, vehicle speed, phase history, optional camera images	Selection of the active phase, dynamic determination of phase duration
Max-Pressure	Vehicle counts on incoming and outgoing lanes of each intersection arm	Phase switching based on calculated ”pressure”, optional adaptive phase duration
Federated RL	Queue length, waiting time, vehicle count, position and speed, delay	Phase selection at specific time intervals (ΔT)
Genetic Programming	Traffic features on lanes (e.g., vehicle count, speed, density)	Next phase selection based on a ”phase urgency” function
Genetic Algorithm (NSGA2)	Vehicle flow, saturation flow, demand	Optimization of green times for all phases within an inter-section network
Fuzzy Logic	Traffic density, queue length, weather conditions (rain, fog)	Dynamic adjustment of green light duration based on fuzzy rules
SARSA(λ)	Intersection saturation, vehicle speed	Control mode selection (maintain, extend, shorten, terminate phase)
Q-learning	Vehicle presence detection, flow	Phase change, cycle length optimization
Hybrid RL + MaxPressure	Combination of RL data (historical state) and MaxPressure (queue lengths)	Both phase selection and adaptive determination of its duration

Table 2. Traffic control algorithm results – delay reduction and throughput improvement

Algorithm	Main results	Evaluation metrics
Deep RL (DQN, PPO)	Reduction in waiting time by 27–68% compared to fixed-time methods	Average waiting time, queue length, number of stops
Max-Pressure	Reduction of average delay by 15–30%	Average delay, queue length
Federated RL	Reduction: 39.95% (halting vehicles), 55.65% (first vehicle waiting time), 64.48% (cumulative waiting time)	Number of halted vehicles, first vehicle waiting time, total waiting time
Genetic Programming	Outperforms MPLight and heuristic methods in most scenarios	Average travel time, throughput
Genetic algorithm (NSGA2)	Average delay reduction of 39% vs. current BHTrans plan; 3% improvement vs. mono-objective GA	Average delay, number of stops per cycle
Fuzzy logic	Waiting time reduction from 26s to 22.1s (approx. 16% improvement)	Average waiting time
SARSA(λ)	City center: 51.2s delay (online) vs. 59.1s (offline); New districts: 21.6s vs. 39.5s	Average delay, number of waiting vehicles
Q-learning	Flow improvement of approx. 30% compared to fixed-time	Queue waiting time, flow
Hybrid RL + MaxPressure	Increased throughput and fair-ness across the intersection network	Throughput, fairness, travel time

(including federated variants) and advanced heuristics (e.g., Max-Pressure, Genetic Programming), generally outperform classic fixed-time control. Their edge comes from real-time data and the ability to adjust both phase order and green durations on the fly. Looking ahead, ITS development is moving toward networked, multi-agent control supported by XAI and V2X integration. Building on prior work [18], an RFID-based intersection control demonstrator was developed and deployed. It used passive vehicle tags and inlet readers for detection, a microcontroller for data aggregation and control logic, signal controllers for actuation, and a communication layer for component coordination. Using a physical scale model (Figure 3), a traditional sequential strategy was compared with the RFID-driven approach. The RFID-based Miller algorithm performed better, reducing average waiting time by 15–20% and queue lengths by 10–18%, while improving priority-vehicle handling thanks to faster response.

The obtained results confirmed that the application of RFID enables adaptive and flexible signal control, increasing intersection throughput relative to the traditional fixed-time algorithm. Despite successful implementation and positive results at the single intersection level, the RFID system encountered fundamental limitations when attempting to scale to the level of an urban network.

The first challenge is the explosion of decision-making complexity. The contemporary urban traffic environment is characterized by unprecedented diversity in vehicle types and their characteristics. This encompasses both powertrain heterogeneity

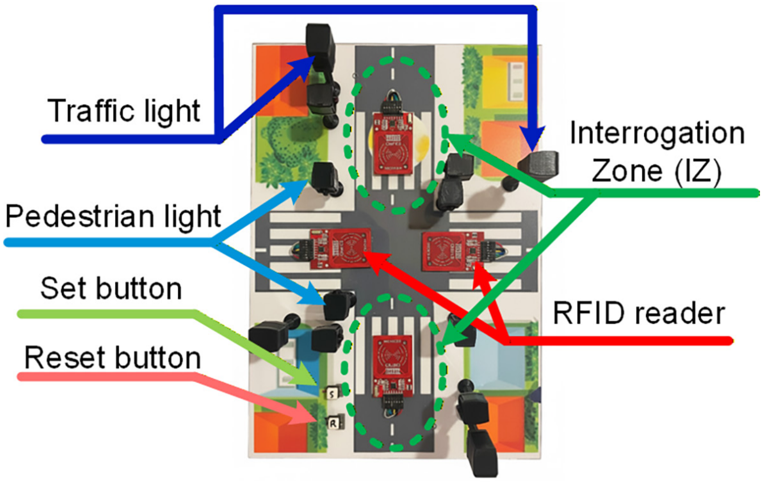

Figure 3. Intersection scale model with the RFID system – general view and reader placement

where electric vehicles (characterized by faster acceleration under congested conditions), hybrid, hydrogen, and traditional internal combustion vehicles exhibit distinct movement dynamics and functional diversity, including passenger cars, public transport, special purpose vehicles (emergency services, construction), and delivery drones.

Additionally, the system must handle dynamic priorities, such as the necessity to facilitate ambulances or prioritize public transport during peak hours. The state and action space in such a system grows exponentially with the number of considered parameters, rendering traditional algorithmic approaches impractical. Manual encoding of responses for $O(n^m)$ possible state combinations (Figure 4), where n is the number of possible states of a single element and m is the number of elements in the system, becomes computationally intractable.

Secondly, the analysis of the RFID system revealed fundamental limits of the rule-based approach. Primarily, a lack of long-term adaptation was diagnosed, as the system reacted only to the current state without the capability to learn from historical traffic patterns. A tendency towards local optimization was also evident, resulting from a lack of coordination mechanisms between intersections, which led to suboptimal solutions on a network scale. Finally, the rigidity of priorities and the difficulty in dynamically adjusting them to changing conditions proved to be problematic.

Beyond the aforementioned issues, the nature of the scale model itself as a research platform proved to be a significant limitation. While the intersection model allowed for practical concept verification, it did not offer the possibility of rapid prototyping of a larger intersection network or testing highly complex traffic patterns (e.g., dynamic intensities during peak hours, traffic waves, or simulations of special corridors for priority vehicles). Every modification of the physical installation was time-consuming and burdened with spatial constraints expanding the model with additional elements would involve enormous labor and cost expenditures. In practice, this meant that beyond a certain complexity threshold, the physical platform ceased to be useful for research purposes.

Consequently, the natural research progression was the transition to the SUMO simulation environment, which enables flexible definition of road network topologies, generation of diverse traffic scenarios, and rapid verification of the efficiency of various control algorithms.

METHODOLOGY

The research was conducted on a model of an isolated, four-way intersection (Figure 5). This architecture is widely used as a fundamental test scenario for new traffic control algorithms [1]. The system operates on two main, non-conflicting phases: Phase 0 (granting priority to the North-South direction, N↕S) and Phase 1 (granting priority to the East-West direction, W↔E). To ensure safety and traffic fluidity, hard timing constraints were introduced: a minimum green phase duration ($t_{green}^{min} = 10\ st$), a maximum green phase duration ($t_{green}^{max} = 60\ s$), as well as a fixed yellow time (3 s) and an inter-green interval.

All experiments were conducted using the microscopic simulator SUMO, version 1.19 [19].

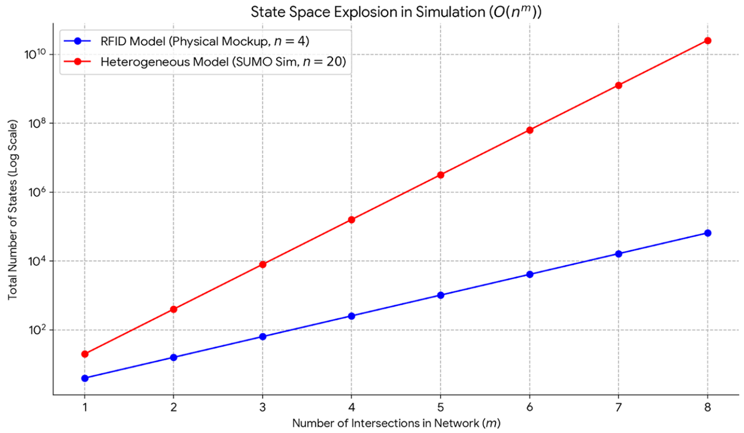

Figure 4. Graph illustrating the exponential explosion of the state space

SUMO was selected as the research platform due to its wide acceptance within the scientific community, high fidelity in modeling vehicle dynamics, and flexible API (including sumo-rl), which enables integration with deep learning libraries. The SUMO model was calibrated according to FHWA (Federal Highway Administration) guidelines to ensure realistic traffic dynamics [20].

This process involved three key stages: first, setting car-following parameters (τ = 1.5 s for driver reaction time and σ = 0.6 for acceleration randomness), which are typical for urban conditions; second, matching traffic volumes and directions using the routeSampler tool (based on Integer Linear Programming, ILP) [21]; and third, speed validation by maintaining the Mean Absolute Percentage Error (MAPE) below the acceptable threshold of 15% required by FHWA standards [22]. Validation using real trajectory data [23] further confirms the high fidelity of the microsimulation model, allowing for the precise reproduction of traffic conditions and the collection of detailed metrics.

Reinforcement learning problem definition (MDP)

The traffic control model was formulated as a Markov Decision Process (MDP), wherein the DRL agent learns the optimal control policy $\pi(a_t|s_t)$ through interaction with the SUMO environment. A key element of the methodology is the use of the data from RFID transponders as the sole source of information regarding the traffic state. This choice is motivated by numerous practical advantages confirmed in the literature [24].

Firstly, this technology guarantees environmental resilience, operating reliably under rain, fog, and low-light conditions, unlike vision-based systems. Secondly, it ensures privacy protection, as the system does not record images, thereby eliminating GDPR-related legal and social issues

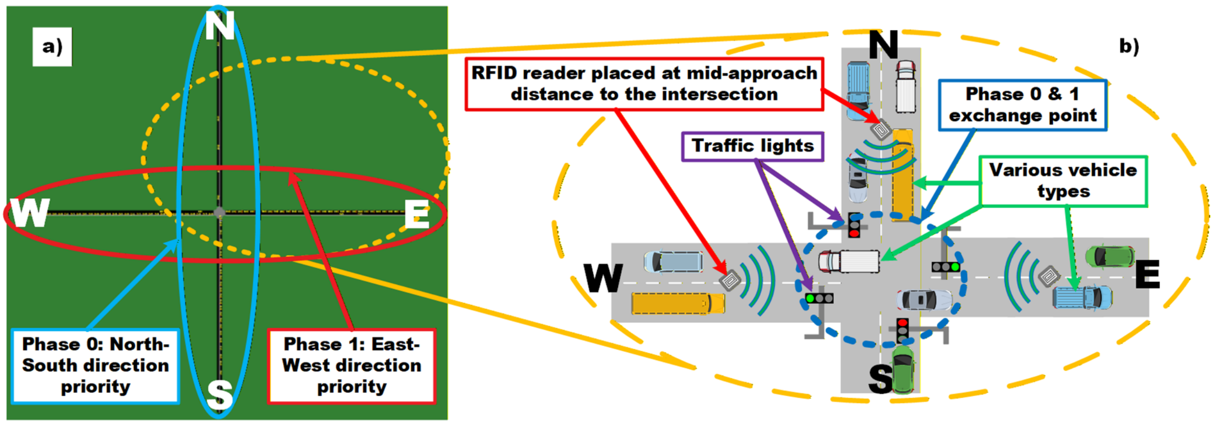
a)
N
W
E
S
Phase 0: North-South direction priority
Phase 1: East-West direction priority
b)
N
W
E
S
RFID reader placed at mid-approach distance to the intersection
Traffic lights
Phase 0 & 1 exchange point
Various vehicle types

Figure 5. Visualization of the analyzed intersection: a) Intersection model in the SUMO simulator; b) Conceptual model illustrating the placement of RFID readers on approaches

[3]. Thirdly, it allows for precise identification of vehicle types (passenger cars, buses, emergency vehicles) with an accuracy exceeding 95% [25], which, depends on appropriate protocol calibration and vehicle speed. Finally, it is characterized by low implementation costs due to inexpensive, passive transponders [24].

RFID readers (simulated in SUMO as detectors of the Induction Loop type) placed at each inlet $j \in \{1,2,3,4\}$ count vehicles, distinguishing between three predefined classes: passenger cars, buses, and ambulances. For each time step t, the state of approach j is represented by a count vector:

$$c_t^{(j)} = \left[n_t^{car,(j)}, n_t^{bus,(j)}, n_t^{amb,(j)}\right] \quad (1)$$

where: $n_t^{type,(j)}$ denotes the number of vehicles of a given type detected at approach j. The full observation o_t is a concatenation of counts from all four approaches (c_t), information about the current phase (p_t), and the normalized duration of this phase (τ_t):

$$o_t = \left[c_t^{(1)} \parallel c_t^{(2)} \parallel c_t^{(3)} \parallel c_t^{(4)} \parallel p_t \parallel \tau_t\right] \in \mathbb{R}^{12+K+|\tau|} \quad (2)$$

where: $p_t \in \{0,1\}^K$ is the one-hot encoding of the current phase $K = 2$, and $\tau_t = \frac{g_t^{elapsed}}{g^{max}}$ is the normalized time elapsed since the beginning of the phase.

The agent operates in a discrete action space, deciding on the selection of the next signal phase. Decisions are made at each time step; however, a safety supervisor system implements them only if timing constraints are met. The action space A comprises two possibilities: Action 0 (request to activate Phase 0, N↕S) and Action 1 (request to activate Phase 1, W↔E). If the agent selects an action corresponding to the currently active phase, the phase continues (provided t < t_{green}^{max}); if it chooses a phase change action, the change occurs only if t ≥ t_{green}^{min}.

The agent's objective is to maximize the cumulative reward. The reward function r_t was designed as a weighted sum of negative penalties, promoting traffic fluidity and minimizing negative indicators, which is consistent with literature regarding the traffic signal control (TSC) evaluation:

$$r_t = -\alpha \cdot \mathrm{delay}_t - \\ -\beta \cdot queue_t - \gamma \cdot stops_t \quad (3)$$

where: $delay_t$, $queue_t$, and $stops_t$ represent the sum of delays, total queue length, and sum of stops at the intersection, respectively [24]. Including the average waiting time (delay) as the primary metric aligns with the highway capacity manual (HCM) and directly reflects user experience [26]. Although more complex metrics, such as delay entropy for assessing fairness between traffic streams, were considered, this work focused on aggregated metrics to simplify optimization. Studies show that the reward functions incorporating balance (entropy) can reduce average delay by 7.4–15%. The weights α, β, γ were determined empirically during hyperparameter optimization to normalize the contribution of individual components [27].

Agent architecture: PPO-transformer

To model the control policy, a hybrid architecture combining the PPO algorithm with a transformer encoder was employed. The choice of the PPO algorithm was driven by its high stability, sample efficiency, and lower sensitivity to hyperparameter tuning compared to other policy gradient methods [28]. A comparison of the algorithms in Table 3 indicates that PPO provides the best trade-off between training stability and computational efficiency, which is essential for safety-critical applications such as traffic signal control (TSC) [29]. Table 3 summarizes DQN, PPO, and A2C in terms of performance, scalability, and sample efficiency. PPO is selected for subsequent experiments because its clipped policy-gradient updates ensure stable optimization, support multiple minibatch epochs, and integrate efficiently with parallel environment rollouts [30].

Literature surveys indicate that policy-based algorithms enable fine-grained control of continuous phase durations, with proximal policy optimization (PPO) offering a strong balance between stability and efficiency. PPO stabilizes training by constraining policy updates through a clipped objective, simplifying the trust-region approach of TRPO while reducing computational complexity [31]. It is widely used in multi-intersection control, effectively balancing exploration and exploitation [32], and provides more stable learning as well as easier tuning than Q-learning and off-policy continuous methods, such as DDPG, TD3, or SAC.

Table 3. Comparison of DQN, PPO, and A2C

Algorithm	Mean return (Ms. Pac-Man, 10M steps)	Speed scaling (×, 32 envs)	Sample efficiency (steps to convergence, qualitative)
DQN	12.133 ± 0.013	1.0×	High sample efficiency reported for discrete-action settings (i.e., fewer environment interactions required to reach a stable performance regime).
PPO	11.837 ± 0.012	4.2×	Moderate sample efficiency; described as stable in continuous-control settings (i.e., convergence typically requires more interactions than highly sample-efficient baselines, while exhibiting reliable optimization dynamics).
A2C	11.445 ± 0.016	5.1×	Low sample efficiency; described as supporting rapid exploration (i.e., comparatively more interactions are required before convergence, despite faster early-stage state-action space coverage).

An integral component of the architecture is the transformer module, used for efficient modelling of temporal dependencies in sequential RFID transponders data [33]. The self-attention mechanism enables the weighting of the importance of historical traffic states, allowing for the detection of patterns (e.g., vehicle platoons) and predictive decision-making.

The input data preparation process, visualized in Figure 6, involves flattening lane-level information and concatenating it. The input sequence X_t fed into the encoder comprises a window of the L most recent observations:

$$X_t = [o_{t-L+1}, \dots, o_t] \in \mathbb{R}^{L \times D} \tag{4}$$

In the implementation, an empirically selected time window of L = 48 steps was adopted, which, with Δt = 5 s, corresponds to 4 minutes of history. This allows for the effective identification of cyclic patterns characteristic of urban traffic [25].

The applied decision-making structure of the agent, presented in Figure 7, differs from standard solutions due to the presence of two specialized MLP output heads operating on the hidden state representation generated by the transformer. The lower module, designated as MLP phase, functions as the phase selection head and decides the priority traffic direction (N↕S or W↔E) using a Softmax activation function. In parallel, the upper module (Timing adjustment actions) is responsible for correcting the green phase duration by selecting one of three discrete actions: shortening (−5 s), no change (+0 s), or extension (+5 s). This decomposition of the action space allows the agent to simultaneously react to current vehicle presence and smoothly adapt the cycle length.

Policy optimization π_θ is realized by maximizing the clipped objective function of the PPO algorithm, defined as:

$$L^{\text{CLIP}}(\theta) = \\ = \mathbb{E}_t\left[\min\begin{pmatrix} r_t(\theta)\hat{A}_t, \\ clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t \end{pmatrix}\right] \tag{5}$$

where: $r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the probability ratio of the new and old policies, $\hat{A}_t$ is the advantage estimator, old and ϵ is the clipping parameter, ensuring monotonic policy improvement without drastic updates during training.

The performance of the proposed PPO-transformer method is compared with two standard baseline algorithms that utilize identical input data from transponders, ensuring fair comparison conditions. The first baseline is the Sequential Fixed-time algorithm, a traditional, widely used method based on a fixed cycle (C = 120 s), where green times are predefined and immutable. The second is the Miller-style algorithm, an advanced adaptive algorithm similar to MOVA logic, which uses the RFID data to estimate queues and dynamically minimize delays within the same safety constraints as the PPO agent.

Algorithm evaluation relies on a set of standard metrics reported directly by SUMO (modules tripinfo and queueoutput), guaranteeing consistency and reproducibility of measurements. The primary metric adopted is the average waiting time [s] (total time spent stopped). Secondary metrics included average delay per vehicle [s] (the difference between actual travel time and theoretical free flow travel time), average queue length [veh], number of stops [veh], and total throughput [veh/h].

The DRL agent training process utilized 16 parallel instances of the SUMO environment to accelerate experience collection and stabilize gradients.

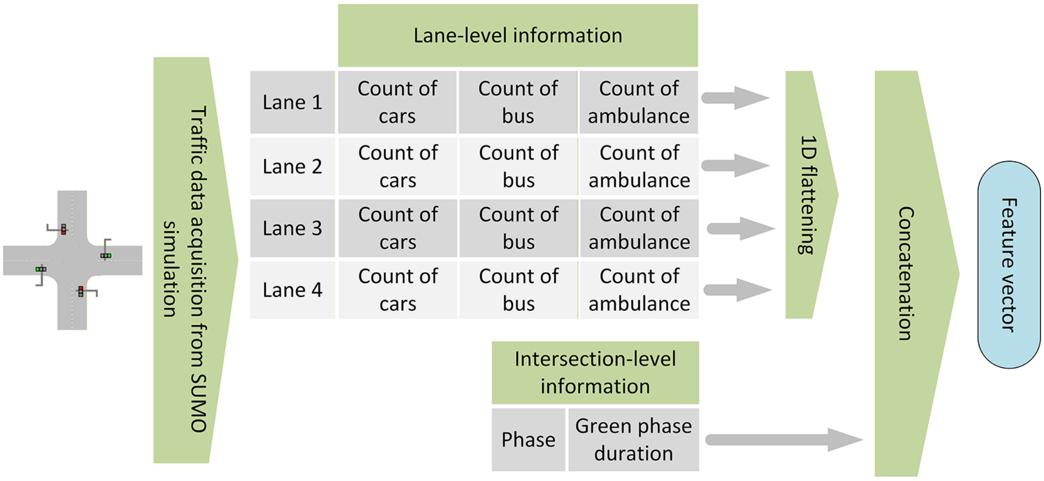

Figure 6. Feature vector creation scheme: lane-level information from RFID transponders is flattened and concatenated into a one-dimensional input vector

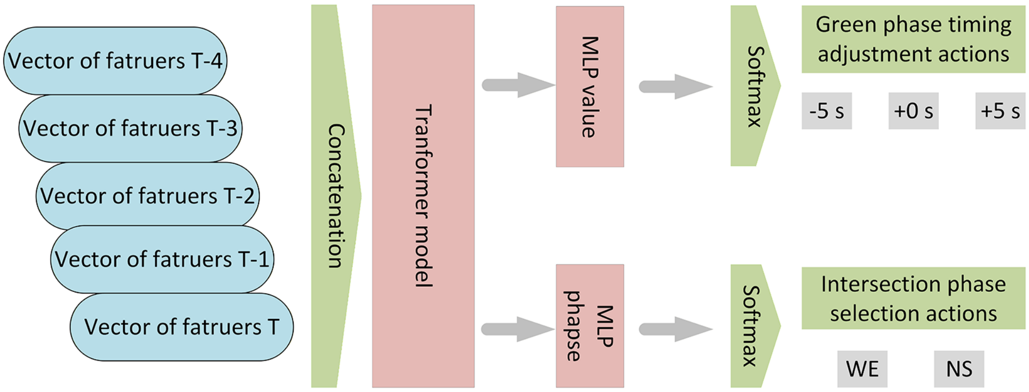

Figure 7. Detailed agent architecture. After processing feature vectors via the transformer, the network splits into two MLP decision modules: the lower one responsible for phase selection (NS/WE) and the upper one for correcting phase duration

Optimal hyperparameters for the PPO algorithm and the transformer architecture were determined using the Optuna library integrated with Ray Tune, employing Bayesian optimization. Key configuration parameters are presented in Table 4.

RESULTS

Table 5 presents a summary comparison of the performance of the three control algorithms based on average values from 10 independent simulation runs. The proposed PPO-transformer method achieves the best results across all key metrics, outperforming both the Fixed-time algorithm and the adaptive Miller algorithm.

Analysis of the key achievements of PPO-transformer reveals significant improvement relative to the baseline methods. A reduction in average delay of 28.6% was achieved compared to the Fixed-time algorithm and 9.1% compared to the Miller algorithm. A distinct shortening of queues is also observed, by 36.0% relative to Fixed-time and 7.7% relative to Miller, respectively. The number of stops was reduced by 29.6% and 9.5%, respectively. Furthermore, the proposed method translated into a 12% increase in total intersection throughput compared to the Fixed-time baseline.

Figure 8 depicts the learning curves for the three best training runs of the PPO-transformer agent, which were selected based on the hyperparameter optimization process using Optuna. The chart also indicates the performance levels (measured as cumulative reward) achieved by the baseline algorithms. All three PPO-transformer training runs exhibit stable convergence and achieve

Table 4. Hyperparameters and training configuration for PPO-transformer

Category	Parameter	Symbol	Value
PPO parameters	Learning rate Discount factor GAE parameter Clipping parameter Rollout size	lr γ λ_{GAE} ϵ -	3×10^{-4} 0.99 0.95 0.2 1000 steps × 16 enviroments
Transformer architecture	Time window Model dimension Number of attention heads Number of layers	L d_{model} - -	48 steps 64 4 2
Reward function weights	Delay weight Queue weight Stops weight	α β γ	0.55 0.30 0.15

Table 5. Hyperparameters and training configuration for PPO-transformer

Method	Delay [s]	Queue [veh]	Stops	Throughput
Fixed-time	42.0 ± 6.0	7.5 ± 2.0	2.7 ± 0.4	Baseline
Miller	33.0 ± 5.0	5.2 ± 1.8	2.1 ± 0.35	+7%
PPO-Transformer	30.0 ± 4.5	4.8 ± 1.6	1.9 ± 0.3	+12%

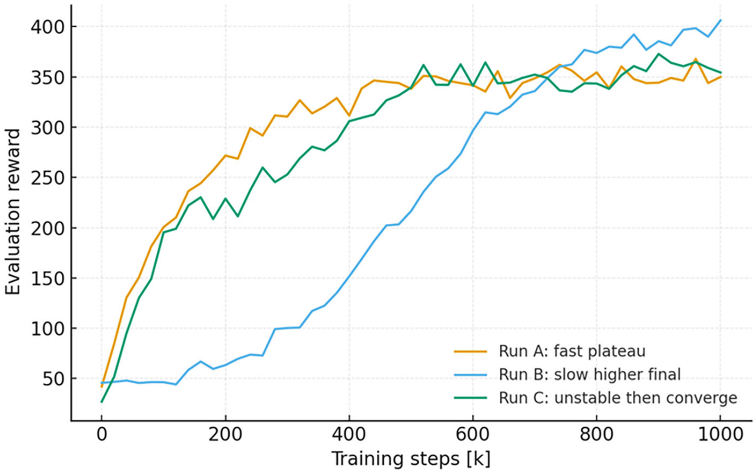

Figure 8. Learning curves for the three best PPO-transformer training runs compared with baseline algorithms

performance levels surpassing both baseline algorithms after approximately 1000 training steps.

Figure 9 shows box-plot distributions of key performance metrics for all three control algorithms. The results indicate that the PPO-transformer agent achieves superior typical performance, evidenced by lower medians for delay, queue length, and stops, as well as higher

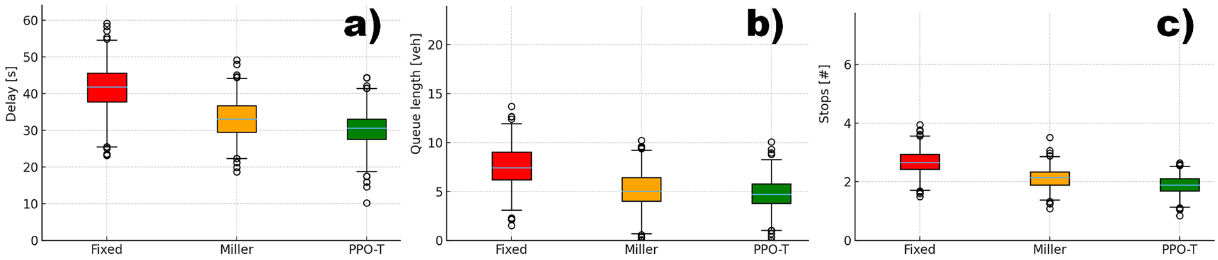

Figure 9. Comparison of performance metric distributions for the three control algorithms: a) Delay [s]; b) Queue length [veh]; c) Number of stops

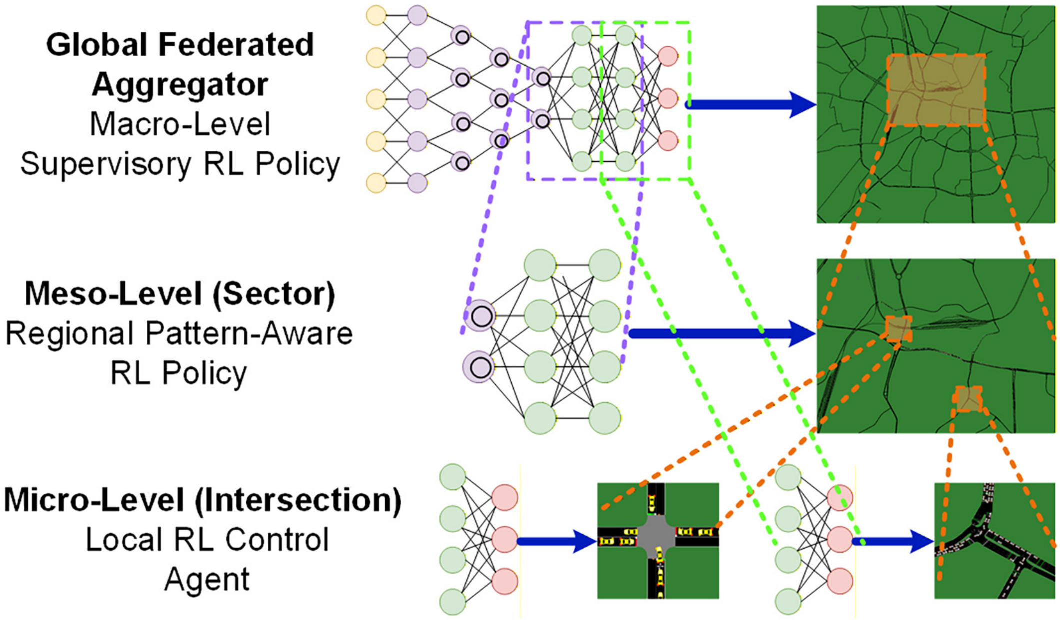
Global Federated Aggregator
Macro-Level Supervisory RL Policy
Meso-Level (Sector)
Regional Pattern-Aware RL Policy
Micro-Level (Intersection)
Local RL Control Agent

Figure 10. Concept of a hierarchical (layered) urban network control architecture

stability reflected in narrower interquartile ranges and reduced spread compared with the fixed-time algorithm. Observed outliers, which are retained in the analysis, correspond to the rare episodes of exceptionally high demand or atypical phase sequences and have minimal influence on the medians and quartiles. This confirms that the performance gains result from improvements under typical operating conditions, rather than isolated extreme cases.

CONCLUSIONS

This study demonstrated the evolution of a traffic control system, transitioning from a physical demonstration model based on RFID to an advanced PPO-transformer agent within the SUMO simulation environment. It was shown that at the level of a single, isolated intersection a fundamental test scenario, the proposed approach, powered by data from transponders, achieves significant improvements in key metrics. In experiments, the PPO-transformer model reduced average delay by 28.6% compared to the fixed-time algorithm and by 9.1% compared to the adaptive Miller algorithm, simultaneously increasing intersection throughput by 12%.

A key challenge identified in this work remains the issue of scalability and the tendency towards local optimization, which is insufficient on a city-wide network scale. This complexity is compounded by increasing fleet heterogeneity, encompassing electric and autonomous vehicles, as well as the necessity to manage dynamic priorities (TSP for public transport, EVP for emergency vehicles) and regulatory requirements such as low emission zones. The study indicates that future systems must rely on data fusion from complementary sources, where vision systems estimate queues, while RFID technology ensures identity assurance for specific vehicles.

On the basis of these conclusions, future research will concentrate on the transition from a single-agent model to a multi-level, layered management architecture for the entire urban network, in accordance with the concept presented in Figure 10. This approach, aligning with recommendations regarding Multi-Agent RL (MARL) formulated in the literature review, assumes the decomposition of the problem into hierarchical levels. At the lowest level, the intersection level, autonomous agents (such as the PPO-transformer) will operate, optimizing local flow based on data fusion. The intermediate level, the city sector, will be managed by supervisory models whose task is not to control signals directly, but to detect traffic patterns on a district scale (e.g., traffic waves) and set goals for subordinate local agents. This addresses the problem of purely local optimization and enables coordination, consistent with the promising results for FL.

At the highest level, the city level, a master control model will be responsible for global optimization, balancing flow between sectors and realizing multi-objective optimization goals, encompassing not only traffic fluidity, but also CO_2 emissions and safety. Further research will also cover the development of XAI methods to improve trust in the system, integration with V2X communication, and algorithm validation in real-world urban conditions, which has been identified as a crucial research direction.

REFERENCES

1. Qadri SSSM, Gökçe MA, Öner E. State-of-art review of traffic signal control methods: challenges and opportunities. Eur Transp Res Rev. 2020;12(1):55. https://doi.org/10.1186/s12544-020-00439-1
2. Li W, Wang W, Wang H. Vehicle occupant detection based on MM-wave radar. Sensors. 2024;24(11):3334. https://doi.org/10.3390/s24113334
3. Kerekes R. A. et al., Vehicle Classification and Identification Using Multi-Modal Sensing and Signal Learning, 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), Sydney, NSW, Australia, 2017; 1–5, https://doi.org/10.1109/VTCSpring.2017.8108568
4. Solaris Bus Coach. What are the Low Emission Zones? Aug. 2024. (Accessed: 18.11.2025). https://ecity.solarisbus.com/en/knowledge-base/low-emission-zones
5. Dengbo He, Huan Yu, and Xiaotong Sun. Sharing the Road: Mixed Traffic with Con-nected Autonomous Vehicles. CRC Press, 2025. (Accessed: 18.11.2025). https://personal.hkust-gz.edu.cn
6. Ji A, Ma X. Vehicle detection and classification for traffic management and autonomous systems using YOLOv10. Alexandria Engineering Journal. 2025 Jul 5;127:804–16. https://doi.org/10.1016/j.aej.2025.06.049
7. Hesami S, De Cauwer C, Vafaeipour M, Rombaut E, Vanhaverbeke L, Coosemans T. Bi-layer eco-driving control design of autonomous electric vehicles in presence of signalized intersections and preceding vehicles. Journal of Intelligent Transportation Systems. 2025 Mar 21;1–18. https://doi.org/10.1080/15472450.2025.2478637
8. Themoonlight.io. Heterogeneous Mixed Traffic Control and Coordination – Literature Review. Mar. 2025. https://doi.org/10.48550/arXiv.2409.12330
9. Washington State DOT. Traffic signal priority & preemption. 2024. (Accessed: 18.11.2025). https://tsmowa.org
10. FHWA Operations. Traffic Signal Timing Manual: Chapter 9 – Preemption and Priority. Tech. rep. Federal Highway Administration, Jan. 2017. (Accessed: 18.11.2025) https://ops.fhwa.dot.gov/publications/fhwahop08024/chapter9.htm
11. Christopher Carey. Smart traffic management could save cities US$277 billion by 2025. Mar. 2021. (Accessed: 18.11.2025). https://cities-today.com
12. Matvey Gerasyov, Kiselev D, Maxim Beketov, Makarov I. VIAAI: Reliable Deep Reinforcement Learning for Traffic Signal Control. 2024 Dec 9;887–90. https://doi.org/10.1109/ICDMW65004.2024.00121
13. Liu H, Gayah VV, Levin MW. A max pressure algorithm for traffic signals considering pedestrian queues. Transportation Research Part C Emerging Technologies. 2024 Sep 23;169:104865–5. https://doi.org/10.1016/j.trc.2024.104865
14. Bao J, Wu C, Lin Y, Zhong L, Chen X, Yin R. A scalable approach to optimize traffic signal control with federated reinforcement learning. Scientific Report. 2023 Nov 6;13(1). https://doi.org/10.1038/s41598-023-46074-3
15. Liao XC, Mei Y, Zhang M. Learning Traffic Signal Control via Genetic Programming. Proceedings of the Genetic and Evolutionary Computation Conference. 2024 Jul 8. https://doi.org/10.1145/3638529.3654037
16. Wu Y. Enhancing urban traffic flow through fuzzy logic-based signal light control optimization. International Journal of e-Collaboration. 2024 Nov 9;20(1):1–13. https://doi.org/10.4018/IJeC.358746
17. Xiao F, Lu J, Li L, Tu W, Li C. Advances in reinforcement learning for traffic signal control: a review of recent progress. Intelligent Transportation Infrastructure. 2025 May 5.
18. https://doi.org/10.1093/iti/liaf009
19. Stęchły A, Pawłowicz B, Siwiec K, Drzał J, Kosior A. Top-level smart city traffic management system with radio frequency identification – based road event detection. Advances in Science and Technology Research Journal. 2025;19(11):429–441. https://doi.org/10.12913/22998624/209674
20. Lopez PA, Behrisch M, Bieker-Walz L, Erdmann J, Flötteröd Y, Hilbrich R, et al. Microscopic traffic simulation using SUMO. IEEE Xplore. 2018; 2575–82. https://doi.org/10.1109/ITSC.2018.8569938
21. Federal Highway Administration. Traffic Analysis Tools Program: Traffic Analysis Toolbox, Volume II: Decision Support Methodology for Selecting Traffic Analysis Tools. U.S. Department of Transportation, 2004. (Accessed: 18.11.2025). https://ops.fhwa.dot.gov
22. SUMO Documentation. routeSampler.py - SUMO Documentation. 2023. (Accessed: 18.11.2025). url: https://sumo.dlr.de/docs/Tools/Routes.html
23. Federal Highway Administration. Traffic Analysis Tools Program: Guidebook on Traffic Signal Control

Algorithms and Applications. U.S. Department of Transporta-tion, 2010. (Accessed: 18.11.2025). https://ops.fhwa.dot.gov

24. Yavuz MN, Özen H. Calibration of microscopic traffic simulation of urban road network including mini-roundabouts and unsignalized intersection using open-source simulation tool. Sci J Silesian Univ Technol Ser Transp. 2024;122:305–18. https://doi.org/10.20858/sjsutst.2024.122.17
25. Pawłowicz B, Trybus B, Salach M, Jankowski-Mihułowicz P. Dynamic RFID Identification in urban traffic management systems. Sensors. 2020;20(15):4225. https://doi.org/10.3390/s20154225
26. Islam I, Li W, Li S, Heaslip K. Heterogeneous Mixed Traffic Control and Coordination. arXiv (Cornell University). 2024 Sep 18; https://doi.org/10.48550/arXiv.2409.12330
27. Transportation Research Board. Highway Capacity Manual (HCM). Washington,D. C.: Transportation Research Board, 2010. (Accessed: 18.11.2025). https://onlinepubs.trb.org/Onlinepubs/sr/sr209/209.pdf
28. Dion F, Rakha H, Kang Y-S. Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. Transp Res Part B Methodol. 2004;38(2):99–122. https://doi.org/10.1016/S0191-2615(03)00003-1
29. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal Policy Optimization Algorithms. In: arXiv preprint arXiv:1707.06347 (2017). https://doi.org/10.48550/arXiv.1707.06347
30. Huang L, Qu X. Improving traffic signal control operations using proximal policy optimization. IET Intell Transp Syst. 2023;17(3). https://doi.org/10.1049/itr2.12286
31. ZiRui Wang, Yue Deng, Junfeng Long, and Yin Zhang. Parallelizing model-based reinforcement learning over the sequence length. In Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24), 2024; 37. Curran Associates Inc., Red Hook, NY, USA, Article 4176, 131398–131433.
32. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. Proximal Policy Optimization Algorithms. 2017. https://doi.org/10.48550/arXiv.1707.06347
33. Michailidis, P., Michailidis, I., Lazaridis, C. R., Kosmatopoulos, E. Traffic Signal Control via Reinforcement Learning: A Review on Applications and Innovations. Infrastructures, 2025; 10(5), 114. https://doi.org/10.3390/infrastructures10050114
34. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017; 5998–6008. https://doi.org/10.48550/arXiv.1706.03762