TransferLight: Zero-Shot Traffic Signal Control on any Road-Network

Johann Schmidt*[1], Frank Dreyer*, Sayed Abid Hashimi, Sebastian Stober

Artificial Intelligence Lab
Otto-von-Guericke University
Magdeburg, Germany
[1] johann.schmidt@ovgu.de

Abstract

Traffic signal control plays a crucial role in urban mobility. However, existing methods often struggle to generalize beyond their training environments to unseen scenarios with varying traffic dynamics. We present TransferLight, a novel framework designed for robust generalization across road-networks, diverse traffic conditions and intersection geometries. At its core, we propose a log-distance reward function, offering spatially-aware signal prioritization while remaining adaptable to varied lane configurations—overcoming the limitations of traditional pressure-based rewards. Our hierarchical, heterogeneous, and directed graph neural network architecture effectively captures granular traffic dynamics, enabling transferability to arbitrary intersection layouts. Using a decentralized multi-agent approach, global rewards, and novel state transition priors, we develop a single, weight-tied policy that scales zero-shot to any road network without retraining. Through domain randomization during training, we additionally enhance generalization capabilities. Experimental results validate TransferLight's superior performance in unseen scenarios, advancing practical, generalizable intelligent transportation systems to meet evolving urban traffic demands.

1 Introduction

Coordinating traffic at intersections is a major challenge for urban planning. Due to the high and ever-increasing volume of traffic in city centres, intersections can quickly become a bottleneck if traffic is not properly coordinated, which can lead to severe traffic congestion. To avoid congested roads, signalized intersection are used to safely and efficiently coordinate traffic flows. Traffic Signal Control (TSC) aims to optimise the traffic flow and related measures (Wang, Abdulhai, and Sanner 2023).

A common solution for TSC is to view it as an optimization problem by designing a mathematical model of the traffic environment using conventional traffic engineering theories and finding a closed-form solution based on that model. Provided that the assumptions inherent to the underlying traffic models are satisfied, such solutions produce good results in theory. However, assumptions such as uniform traffic (Webster 1958; Little, Kelson, and Gartner 1981; Roess, Prassas, and McShane 2004) or unlimited vehicle storage capacity of lanes (Varaiya 2013) are difficult or even impossible to observe in reality, which is why such solutions are not optimal in practice, especially when traffic demand is high and fluctuates significantly. Hence, the field pivoted towards adaptive signal control policies, which are learned from data through deep reinforcement learning (RL) (Wei et al. 2021). Yet, most existing works still struggle to effectively transfer their learned policies to changing traffic conditions.

*These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

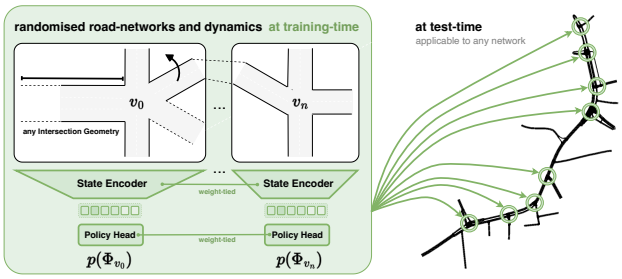

Figure 1: Our proposed traffic signal controller learns a general policy for flexible phase prediction during training. Due to the weight-tied models, we can apply the learned model to any road-network during inference.

Rigid State and Action Spaces The majority of RL-based approaches employ overly rigid data structures to encode the mapping from states to actions. Numerous studies simply encode states and actions as fixed-size vectors or spatial matrices (Zheng et al. 2019; Wang et al. 2024). This approach inherently constrains the learned policy to a specific intersection geometry, which is defined by the structural arrangement of lanes, movements, and phases. Consequently, the reusability of such models is limited to networks of homogeneous intersections with identical geometries (Wei et al. 2018, 2019a). In an attempt to increase flexibility, states and actions are (zero)-padded (Zheng et al. 2019; Chen et al. 2020), introducing upper bounds to the system's diversity. However, due to the combinatorial explosion of possible intersection layouts, the required number of paddings grows exponentially (Chu et al. 2019), potentially compromising training efficiency and generalization ability of the model.

Rigid Traffic Environments Another significant limitation in current RL-based approaches is the insufficient consideration of traffic dynamics' variability during training. The majority of methods employ identical spatio-temporal traffic patterns across all training episodes (Wei et al. 2018, 2019a). While these models may exhibit impressive performance within this constrained setting, they typically suffer from substantial performance degradation when confronted with real-world variability (Zheng et al. 2019; Yoon et al. 2021). This performance decline can be attributed to overfitting and the drastically constrained exploration space during training (Jiang, Kolter, and Raileanu 2024). The limited exposure to diverse traffic scenarios during the learning phase results in models that lack robustness and adaptability to the complex and dynamic nature of real-world traffic conditions (Korecki, Dailisan, and Helbing 2023). Consequently, these models struggle to generalize effectively to the multifaceted and often unpredictable traffic patterns encountered in practical applications, highlighting a critical gap between laboratory performance and real-world efficacy.

Degenerated Reward Functions Reinforcement Learning is driven by the choice of reward to be optimised. As long-term objectives, like travel time, depend on a sequence of actions, credit assignment is difficult and might impact the training efficiency drastically. Hence, short-term objectives are used instead, like waiting time or queue length (Zheng et al. 2019; Devailly, Larocque, and Charlin 2022, 2024), or weighted combinations of them (Wei et al. 2018; Yoon et al. 2021; Wu, Kim, and Ma 2023). Unfortunately, these rewards do not correlate, leading to different optima (Wei et al. 2019a). As a solution, Wei et al. (2019a) showed that max-pressure control policy stabilise the traffic system over time, which lets queue length and travel time settle in a local optima. Based on these guarantees, pressure-based rewards are frequently used in recent works (Oroojlooy et al. 2020). However, as pressure is computed as a mean, it is invariant to various transformations of the input signal. Different spatial locations of heavy traffic loads along the lane do not influence the indicator, leading to misjudgments of states.

Towards General Control Policies The limitations of existing traffic signal control (TSC) approaches, particularly their inability to generalize across intersection and traffic conditions, necessitate a more robust and flexible solution.[1] We present *TransferLight*, a novel model that addresses these challenges by leveraging graph-structured representations and advanced training techniques. Our contributions include

- We introduce a novel log-distance reward function that provides a continuous, spatially-aware signal prioritizing near-intersection vehicles while remaining bounded and adaptable to diverse lane configurations, addressing key limitations of traditional pressure-based rewards (Wei et al. 2019a).
- Building upon prior research (Yoon et al. 2021; Devailly, Larocque, and Charlin 2022, 2024), we propose a heterogeneous graph neural network (Kipf and Welling 2017) architecture for state encoding. This approach captures fine-grained traffic dynamics and enhances generalization, enabling universal applicability of the learned policy to varied intersection and road network geometries.
- We utilize domain randomization to vary both static and dynamic features of the traffic environment during the training process similar to Devailly, Larocque, and Charlin (2022, 2024). This approach enhances the model's generalization capabilities to novel scenarios.
- We use a decentralised multi-agent approach with a global reward and novel state transition priors to foster proactive decisions. This allows us to learn a single shared (weight-tied) general policy that can be zero-shot scaled to any road-network during test-time without retraining.

By combining these elements, *TransferLight* overcomes the limitations of previous approaches, offering a unified framework for learning robust and adaptive traffic signal control policies. Our experimental results demonstrate that *TransferLight* achieves good performances on novel (unseen) scenarios, making a significant step towards practical, generalizable and intelligent transportation systems.

2 Priliminaries

Traffic Signal Control We define a road network as a graph $G = (\mathcal{V}, \mathcal{I} \cup \mathcal{O})$, where $\mathcal{V} = \{v_k \mid k \in [1, 2 \dots V]\}$ is the set of V signalised junctions. This geometric structure defines the environment for an agent to act on. For notational convenience, we differentiate between incoming lanes $\mathcal{I}_v$ and outgoing lanes $\mathcal{O}_v$ for each intersection $v \in \mathcal{V}$.[2] For situations, where we do not need to differentiate between incoming and outgoing lanes, we use $\ell \in \mathcal{I} \cup \mathcal{O}$ to denote an arbitrary lane. Each lane ℓ defines a finite one-dimensional coordinate space $\ell \subset \mathbb{R}_+ \setminus \{0\}$ with its origin at the intersection's centre.

As in (Wei et al. 2019c; Urbanik et al. 2015), we define $m_v = (i, o)$ to be a movement from $i \in \mathcal{I}_v$ to $o \in \mathcal{O}_v$ with $m_v \in \mathcal{M}_v \subset \mathcal{I}_v \times \mathcal{O}_v$. A movement can be either *permitted*, *prohibited*, or *protected*. A movement is *protected* if the associated road users have priority and do not have to give way to other movements. A movement is *prohibited* if the signal is red, and it's *permitted* if the associated road users must yield the right-of-way to the colliding traffic before they are allowed to cross the intersection. A phase ϕ describes a timing procedure associated with the simultaneous operation of one or more traffic movements (Urbanik et al. 2015) with a green interval, a yellow change interval and an optional red clearance interval. Let $\phi_v \in \Phi_v$ be a phase at intersection v, and $\mathcal{M}_{\phi_v} \subset \mathcal{M}$ be the associated right-of-way movements. The phase set Φ_v defines the discrete action space for an agent acting on v.

This defines the static part of the environment. The dynamics are given by a set of moving vehicles $\mathcal{C} = \{c_k \mid$

[1] The ideal solution would be a model capable of maintaining consistent, high-quality performance across a wide spectrum of road-networks and traffic dynamics. Once the general control policy is obtained, it can be applied to any (urban) environment.

[2] Such that, $\mathcal{I} := \bigcup_{v \in \mathcal{V}} \mathcal{I}_v$ and $\mathcal{O} := \bigcup_{v \in \mathcal{V}} \mathcal{O}_v$.

$k \in [1, 2 \dots C]\}$. These are modelled as points on the one-dimensional coordinate space ℓ. We define a state $\mathcal{C}_t$ by the vehicle positions at a time point t. Hence, each vehicle's motion is captured by $c(t)$, which is evaluated at an a priori defined sampling frequency of the sensor (or the simulation).

Cooperative Markov Games In a multi-intersection road network, agent coordination is crucial for efficient traffic flow. This scenario extends the Markov Decision Process to a Markov Game (Littman 1994). At each time step t, every agent $v \in \mathcal{V}$ observes the environment state $\mathcal{C}^t \in \mathcal{C}$ and selects an action $\phi_v^t \in \Phi_v$ using its policy $\pi_v(\phi_v^t \mid \mathcal{C}^t) : \Phi_i \times \mathcal{C} \mapsto \mathbb{R}_+$. The environment then transitions to $\mathcal{C}^{t+1}$ according to $T(\mathcal{C}^{t+1} \mid \phi^t, \mathcal{C}^t) : \mathcal{C} \times \Phi \times \mathcal{C} \mapsto \mathbb{R}_+$, where $\Phi = \bigcup_{v \in \mathcal{V}} \Phi_v$ is the joint action space. Each agent receives a reward r_v^{t+1} based on $R_v(\mathcal{C}^t, \phi^t, \mathcal{C}^{t+1}) : \mathcal{C} \times \Phi \times \mathcal{C} \mapsto \mathbb{R}$, denoted as R_v^t for brevity.

In fully cooperative Markov Games, the global reward is equivalent to individual rewards ($R^t = R_v^t, \forall v \in \mathcal{V}$) or a team average ($R^t = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} R_v^t$). While the former entails aligned goals for individual agents, the latter allows agents to pursue distinct objectives that contribute to the overall team benefit. The objective is then to find a joint policy $\pi = \{\pi_v \mid v \in \mathcal{V}\}$ that maximizes the expected discounted sum of global future rewards:

$$\pi^* = \operatorname*{argmax}_{\pi} \mathbb{E}_{\mathcal{C}^t \sim \mu} \mathbb{E}\pi \left[\sum_{k=0}^{\infty} \gamma^k R^{t+k} \mid \mathcal{C}^t \right], \qquad (1)$$

where $\mu(\mathcal{C}^t \mid \pi)$ is the stationary distribution of the Markov chain under joint policy π.[3]

3 Lifting Pressure-based Rewards

Under mild assumptions[4], a max-pressure control policy *stabilises the traffic system* over time (Wei et al. 2019a). This means, that measures like queue length, throughput, and travel time settle in local optima. We build upon these theoretical results by eliminating a remaining shortcoming of pressure-based systems.

Degeneracies of Pressure We prove that the pressure of a movement suffers from several degeneracies introducing plateaus to the reward surface, which prohibit convergence to superior extrema. The pressure $\rho(m)$ of a movement $m \in \mathcal{M}$ (Wei et al. 2019a) is defined by the difference between the incoming and outgoing vehicle densities, such that

$$\frac{C_i}{|i|} - \frac{C_o}{|o|} \quad \text{with} \quad m = (i, o), \qquad (2)$$

where $|i|, |o| \in \mathbb{R}_+ \setminus \{0\}$ are the length of the lanes. Densities are computed by the arithmetic mean over vehicles [5],

[3]Note that Eq. (1) is permutation invariant with respect to $\mathcal{V}$. The geometric structure of the road network needs to be induced in the state representation.

[4]That is, no physical queue expansion for non-arterial environments and admissible average demand.

[5]We can interpret the traffic density as $\frac{1}{|\ell|} \sum_{p \in \ell} \mathbb{1}_p$, where $\mathbb{1}_p \in \{0, 1\}$ is an indicator returning 1 if there is a vehicle at the spatial p on ℓ.

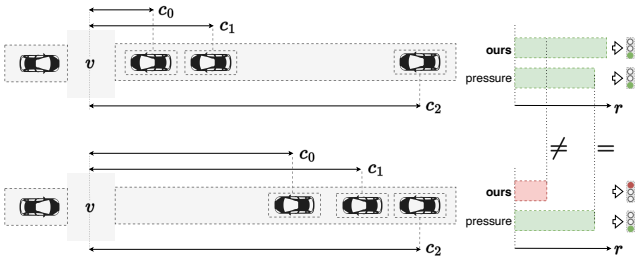

Figure 2: Pressure (see Eq. (2)) is symmetric to vehicle position translations within the lane's coordinate space. Our more expressive measure breaks this symmetry.

which comes with the following fundamental properties arsing from the linearity of the operation:

- *permutation invariance*, $\text{mean}(\mathbf{x}) = \text{mean}(\pi\mathbf{x})$,
- *translation equivariance*, $\text{mean}(\mathbf{x} + b) = \text{mean}(\mathbf{x}) + b$,
- *scale equivariance*, $\text{mean}(b\mathbf{x}) = b\,\text{mean}(\mathbf{x})$,

for any sequence $\mathbf{x} \in \mathbb{R}^n$, permutation matrix $\pi \in \{0, 1\}^{n \times n}$, and $b \in \mathbb{R}$. These symmetries also apply locally, such that $\text{mean}(\{x_1 + b, x_2 - b, \dots, x_n\}) = \text{mean}(\{x_1, x_2, \dots, x_n\})$ for instance. By these relations, *equivalence classes* are formed, i.e., subsets with constant outputs under these transformed inputs. Hence, the pressure stays constant, when permuting the positions of vehicles, shifting vehicles along the lane or scaling the distribution of vehicles. The latter two are of specific interest, as the first one would not change the state $\mathcal{C}_t$.

Modelling the reward function by (pure) pressure maps these equivalence classes on the reward surface and with that on the loss surface. As gradients on these plateaus are exactly zero, gradient-based optimisation will fail leaving these regions. This might be mitigated by a drastically increased momentum term (Kingma and Ba 2014), allowing the model to jump over these regions. However, the model can extract valuable information from these regions *iff* these degeneracies are lifted.

Lifting the Degeneracies We argue, that the degeneracies of Eq. (2) can get lifted by inducing spatial information. As stated in Section 2, every vehicle c can be interpreted as a point on the lane's one-dimensional coordinate space ℓ. Using the Euclidean distances does not lift the degeneracies[6] Instead, we use log-distance, defined by $\log(c + \epsilon) \in [\log \epsilon \approx -\infty, \log(1 + \epsilon) \approx 0]$, where $\epsilon \approx 0$. The farther away a vehicle c from v, the larger the log-distance (closing in on 0). This can be computed for an entire lane $\ell \in \mathcal{I} \cup \mathcal{O}$ by $\{\log c + \epsilon \mid c \in \mathcal{C}_\ell\} = \log(\mathcal{C}_\ell + \epsilon)$.

We interpret the cumulated log-distances as the negative energy of the system. Analogously to a simplified potential energy of a system of particles, where the energy

[6]For example, a configuration with a single vehicle at a large distance from the intersection's centre would yield the same metric value as a configuration with multiple vehicles positioned closer to the centre, provided the sum of their distances is equal to that of the single distant vehicle.

increases with distance between particles, as leveraged in (Schmidt, Köhler, and Borstell 2024). *The goal is to minimise this energy, i.e., push the densities away from the intersection's centre.* We interpret the total log-distance as the energy $E_\ell \in \mathbb{R}_+$ of the lane,

$$\hat{E}_\ell = \sum_{c \in \mathcal{C}_\ell} \log(c + \epsilon). \tag{3}$$

This breaks both the translation and scale equivariance.[7] Therefore, $\hat{E}$ lifts the degeneracies of E using the symmetry-breaking log-distance formulation. We define the cumulated vehicle positions on a lane to be its energy, $E_\ell := C_\ell \in \mathbb{R}_+$. With this, we can formulate the average log-pressure by the cumulated and normalised log-distances,

$$\sum_{(i,o) \in \mathcal{M}_v} \frac{1}{|i|} \hat{E}_i - \frac{1}{|o|} \hat{E}_o. \tag{4}$$

With this we define the reward

$$r_i = -\left| \sum_{(i,o) \in \mathcal{M}_v} \frac{1}{|i|} \hat{E}_i - \frac{1}{|o|} \hat{E}_o \right|. \tag{5}$$

We focus on cooperative Markov Games (Littman 1994), where agents have an incentive to work together to achieve a team goal, which can be expressed by a global reward function $R(t)$. In such Multi-Agent settings, sharing information among agents is key, as the other agents induce otherwise unpredictable dynamics (non-stationary environments), which limits cooperation (Zhang, Yang, and Zha 2020). This can be done by joint state and action spaces, which, however, require supportive mechanisms to cope with the exponentially growing joint spaces (Choudhury et al. 2021). Hence, action and state spaces are often disjoint and agents are trained by a global reward function to encourage cooperation (Wei et al. 2019a; Chen et al. 2020; van der Pol et al. 2022). In the following, we propose our state encoding to cope with these challenges.

4 Graph-Structured State Encoding

Following Devailly, Larocque, and Charlin (2022, 2024), we utilize a graph neural network on a heterogeneous graph to encode both static and dynamic state characteristics of individual intersections. *This allows us to encode any intersection geometry* regardless of the length of lanes and the number of lanes, approaches, movements and phases. By sharing the parameters across all intersections in the network, *the model is encouraged to converge to a policy that generalises various intersection configurations and traffic conditions*. We contextualise encodings by state transition priors to *allow for proactive decisions* (which enable green waves). We provide an illustration of our state encoding in Fig. 3. We will discuss its core elements in the following.

[7] This follows from $\log(c + \epsilon + b) \neq b + \log(c + \epsilon)$ for $b \neq 0$ and $\log(bc + \epsilon) \neq b \log(c + \epsilon)$.

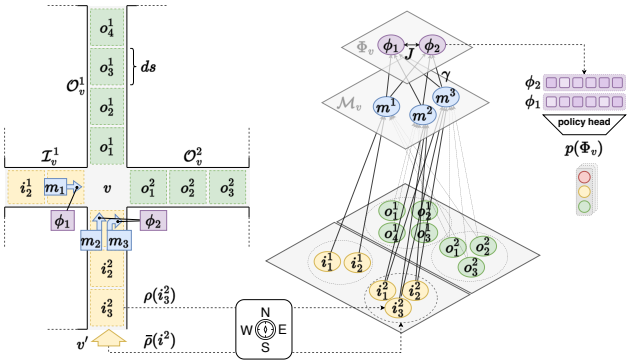

Figure 3: Our hierarchical state space encoding uses a position-encoded segment-density set on the lowest level. This information is embedded and aggregated to form movement representations, which then undergo another pass to the phase level. On the phase level, we have intra-level updates, otherwise information are passed down-to-top along the directed heterogenous graph structure.

Lane Partitioning As stated in Section 3, the density estimate over ℓ suffers from degeneracies. A state encoder using these estimates as inputs would inherit the degeneracies, which would smooth out update nuances. Instead, we bound the degeneracies to only act in limited sub-spaces. We define a metric ds to partition ℓ into $\frac{\ell}{ds}$ equally-sized segments, which defines a hyperparameter to control the resolution of the measure applied on top of lanes. In each segment, we estimate the density by $\rho(s) = \frac{1}{ds} C_s$. This factors out the number of vehicles for the input to the encoder (similar to the density estimate over ℓ). Such a representation ensures that the dynamical traffic system is fully described (for a proof refer to Wei et al. (2019a)).

While often overlooked in previous studies (Zheng et al. 2019; Oroojlooy et al. 2020; Zang et al. 2020; Yoon et al. 2021), the length of lanes or segments plays a crucial role in traffic dynamics. Our approach employs a uniform and constant segment length ds, thereby streamlining the input feature set compared to related works (Devailly, Larocque, and Charlin 2022, 2024). This design choice allows the policy to implicitly learn length-related characteristics, including segment capacity, enhancing the model's adaptability to diverse road networks compared to prior work (Wei et al. 2019a). However, ds has to be small enough to minimise the impact of local degeneracies, as discussed in Section 3.

Transition Prior Modelling only the dynamics within the boundaries of the intersection, would result in reactivity rather than proactivity, especially when ℓ is small. To fix this, we interpret the road-network as a coordination graph, which allows us to induce additional context to each agent $v \in \mathcal{V}$. We define the connectivity of the coordination graph by movements,

$$\begin{aligned} \mathcal{M}_{\leftarrow \ell} &:= \{(\ell, o') \mid i' = \ell; (i', o') \in \mathcal{M}\}, \\ \mathcal{M}_{\rightarrow \ell} &:= \{(i', \ell) \mid \ell = o'; (i', o') \in \mathcal{M}\}. \end{aligned} \tag{6}$$

This gives a single-hop receptive field for every $v \in \mathcal{V}$. This locally interdependent structure (Yi et al. 2024) can be inter-

preted as modelling communication channels between v and its adjacent neighbour intersections.[8] We use this to define a state transition prior

$$\bar{\rho}_\ell = \sum_{(i,o)\in\mathcal{M}_{\to\ell}} \rho(i_0) - \sum_{(i,o)\in\mathcal{M}_{\leftarrow\ell}} \rho(o_0), \tag{7}$$

where i_0 and o_0 are the closest segments to the intersection's centre. If $\bar{\rho}_\ell < 0$, more vehicles are going to leave ℓ.

Lane Coordinate Frames To break the permutation invariance of the segment set, we define the centre of the intersection as a reference point and induce a positional encoding on the segments relative to that point. We use sinusoidal positional encoding (Vaswani et al. 2017) along segments on each lane and over lanes. Instead of additive fusion (Vaswani et al. 2017), we concatenate the positional information with the density of the segment. This preserves both identities, which improves expressiveness without the need of separate processing (Yu et al. 2023). Thus, we can define the segment feature vector by $\mathbf{h}_s = [\rho(s) \| \operatorname{pe}(s) \| \bar{\rho}_\ell]^\top \in \mathbb{R}^d$ be the feature vector of a segment $s \in \ell$.

Segment-to-Movement Encoding We apply a graph attention network (Veličković et al. 2017) to learn the mapping $\mathbb{R}^{\frac{\ell}{ds}\times d} \mapsto \mathbb{R}^{d'}$. To improve expressiveness, we use dynamic scoring (Brody, Alon, and Yahav 2022) to compute attention weights

$$\alpha_s = \frac{\exp u(\mathbf{h}_s)}{\sum_{s'\in\mathcal{N}_\ell} \exp u(\mathbf{h}_{s'})} \quad \text{with} \quad u(\mathbf{h}_s) = \mathbf{a}_s^\top \sigma\left(\mathbf{W}_s \mathbf{h}_s\right), \tag{8}$$

where $\mathbf{a}_s \in \mathbb{R}^{d'}$ and $\mathbf{W}_s \in \mathbb{R}^{d\times d'}$ are learnable weights. $\mathcal{N}_\ell$ defines the segment set for lane ℓ and σ is a monotonic nonlinearity, like Leaky-ReLU. We then compute a representation for each movement $\mathbf{h}_m \in \mathbb{R}^{d'}$ by a weighted average of its segments, such that

$$\mathbf{h}_m = \sigma\left(\mathbf{b}_s + \underbrace{\hat{\mathbf{W}}_s \mathbf{h}_s}_{\text{residual}} + \sum_{s\in\mathcal{N}_i} \alpha_s \mathbf{W}_s \mathbf{h}_s + \sum_{s\in\mathcal{N}_o} \alpha_s \mathbf{W}_s \mathbf{h}_s\right), \tag{9}$$

where $\hat{\mathbf{W}}_s \in \mathbb{R}^{d\times d'}$ enables learnable residual connections and $\mathbf{b}_s \in \mathbb{R}^{d'}$ being the bias term. Movement nodes do not hold information initially, hence the update is independent of the original target node features $\mathbf{h}_m$.[9] To ensure that the neighbourhood aggregation runs in a numerically stable manner while allowing for a high degree of representational strength, the individual aggregation functions are implemented as weighted sums with multi-head attention. We compute attention over incoming and outgoing segments separately, but aggregate and update the movement node features in parallel.

[8] Agents are incentivized to cooperate rather than act solely in their self-interest. This can lead to more stable equilibria where multiple agents coordinate their strategies effectively.

[9] $\mathbf{h}_m$ is initialised with zeros, neutralising its impact in Eq. (9).

Movement-to-Phase Encoding The obtained movement node features $\{\mathbf{h}_m \mid m \in \mathcal{M}_v\}$ form another heterogeneous directed acyclic sub-graph with the phase nodes. Instead of a sparsified graph, we use a fully-connected bipartite structure with additional edge features. Each connection between a movement $m \in \mathcal{M}_v$ and a phase $\phi \in \Phi_v$ holds a scalar $\gamma_{m\phi} \in \{-1, 0, 1\}$ as an edge feature indicating whether a movement is prohibited, protected or permitted during a phase. In literature, often only permitted, or protected movements are considered (Zheng et al. 2019; Zang et al. 2020). We argue, that also the information about prohibited movements are essential to determine the energy of a phase. Furthermore, phase nodes are initialised by a binary flag $\mathbf{h}_\phi \in \{0, 1\}$ indicating whether the phase is currently active or not. This changes Eq. (8) and Eq. (9) to

$$u(\mathbf{h}_{m\phi}) = \mathbf{a}_m^\top \sigma\left(\mathbf{W}_m \mathbf{h}_m + \mathbf{W}_\phi \mathbf{h}_\phi + \mathbf{W}_\gamma \gamma_{m\phi}\right) \tag{10}$$

$$\text{and} \quad \mathbf{h}_\phi = \sigma\left(\mathbf{b}_m + \tilde{\mathbf{W}}_m \mathbf{h}_m + \sum_{m\in\mathcal{M}_v} \alpha_{m\phi} \mathbf{W}_m \mathbf{h}_m\right),$$

where $\mathbf{a}_{m\phi}, \mathbf{b}_m \in \mathbb{R}^{d'}$ and $\mathbf{W}_m, \tilde{\mathbf{W}}_m \in \mathbb{R}^{d'\times d'}$ are learnable weights. Attention weights are computed as in Eq. (8) but normalised over $\mathcal{M}_v$ instead. Contrary to Veličković et al. (2017), we embed node and edge features separately, which reduces the model complexity while still preserving expressiveness. Furthermore, we use the edge features and the initial phase flag only to compute the attention scores. Hence, the model can use γ to weight movement features during aggregation, but it does not make any further inferences from the movement information. As we use a directed acyclic graph, we do not face the identity issue discussed in general edge-based graph attention (Wang, Chen, and Chen 2021). This form of aggregation also preserves permutation invariance. In contrast to the level before, this is an important property for the encoding of phases, as they should be orientation independent (Zheng et al. 2019). The obtained phase node representations are further leveraged in an intra-level propagation phase, as discussed next.

Intra-Level Phase Propagation We model the connection between phases as a fully-connected homogeneous graph with Jaccard coefficients $J_{\phi\phi'} \in \mathbb{R}_+$ between each phase pair $\phi, \phi' \in \Phi_v$. The Jaccard coefficient encodes the intersection over the union of the green signals between the two phases. This structures the phase space by quantifying the relative differences between phases w.r.t. to their "green" portions. This results in the following intra-level update formulation

$$u(\mathbf{h}_{\phi\phi'}) = \mathbf{a}_\phi^\top \sigma\left(\mathbf{W}_\phi \mathbf{h}_\phi + \mathbf{W}_\phi \mathbf{h}_{\phi'} + \mathbf{W}_J J_{\phi\phi'}\right) \tag{11}$$

$$\text{and} \quad \mathbf{h}_\phi \leftarrow \sigma\left(\mathbf{b}_\phi + \tilde{\mathbf{W}}_\phi \mathbf{h}_\phi + \sum_{\phi\in\Phi_v} \alpha_{\phi\phi'} \mathbf{W}_\phi \mathbf{h}_\phi\right),$$

where $\mathbf{a}_\phi, \mathbf{b}_\phi \in \mathbb{R}^{d'}$ and $\mathbf{W}_\phi, \tilde{\mathbf{W}}_\phi \in \mathbb{R}^{d'\times d'}$ are learnable weights. Again, attention weights are computed as in Eq. (8) but normalised over Φ_v instead. After propagation, each node holds weighted information about all other phases, which renders a single layer sufficient.

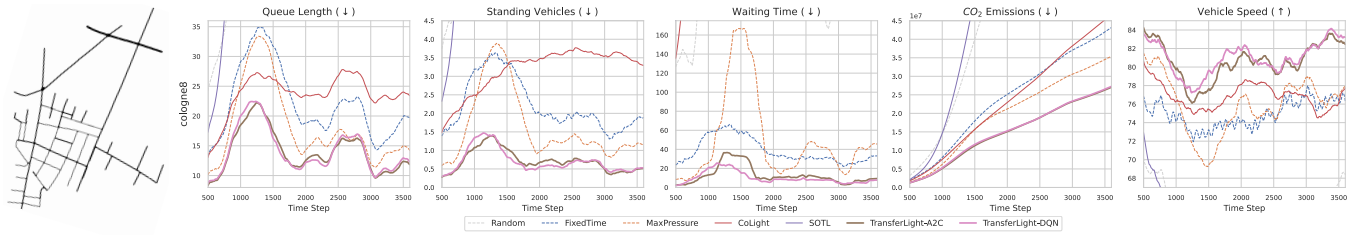

Figure 4: Test performances (moving averages) on *Cologne8* over 3600 simulated time steps.

Weight-Sharing Our universal state encoding function allows using the model for each intersection. In this way, *our model can be applied to any road-network size.* Moreover, by sharing parameters among agents, the algorithms are essentially encouraged to converge to a region in parameter space that works well for arbitrary intersections and traffic conditions, thereby promoting generalization.

5 Domain-Randomised Training

Domain Randomization (DR) is a powerful technique for bridging the sim-to-real gap (Tobin et al. 2017). By introducing sufficient variability in the simulated source domain during training, DR enables the agent to generalize its policy to the target domain, treating it as another variant within its learned distribution. The core principle of DR involves configuring the environment based on a randomly sampled configuration $\xi \sim \Xi$, where Ξ represents the space of possible domain parameters. Ξ contains all traffic-networks under some degree of freedom, as well as different forms of traffic dynamics. The agent's objective is to find an optimal policy π^* that maximizes the expected return across all possible environmental configurations, i.e., extending Eq. (1)

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_{\xi}\mathbb{E}_{\mathcal{C}^t}\mathbb{E}_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k R(t+k) \mid \mathcal{C}^t, \xi\right], \quad (12)$$

where $\mathcal{C}^t \sim \mu(\mathcal{C}^t \mid \pi; \xi)$ denotes the stationary distribution of the Markov chain under configuration ξ and policy π. We sample the static environmental characteristics (like the number of intersections and lane lengths) from a uniform distribution a priori. For the dynamics, we use traffic flow modelling to define each flow $f \in \mathcal{F}$ by its route, vehicle count, and departure times. To enhance realism and variability, we model departure times using a beta distribution with flow-specific parameters:

$$\mathcal{T}^f = \{t_{\max} b_k \mid b_k \sim \mathrm{Beta}(\alpha^f, \beta^f), 1 \leq k \leq C^f\}, \quad (13)$$

where α^f, β^f are sampled from a uniform distribution.[10] In literature, a Poisson process with a constant rate of $\frac{C^f}{t_{\max}}$ vehicles per second with $t \in [0, t_{\max}]$ is often used instead. However, the constant departure rates are often not realistic in practice (e.g. during rush hours). This approach allows for diverse departure patterns, including peaks and fluctuations, while still encompassing the possibility of constant departure rates.

[10]We sample a destination and target line segment and use the Dijkstra algorithm (Dijkstra 1959) to estimate the shortest path. We use $\alpha^f, \beta^f \sim \mathrm{Unif}(1, 10)$ in our experiments. The number of vehicles following the flow is sampled from a pool of C available vehicles in the simulation.

6 Experiments

The primary objective of our experiments is to show the ability of *TransferLight* to transfer its control policy to novel scenarios without requiring any kind of re-training or fine-tuning. In all experiments, *TransferLight* is trained on randomly generated road-networks with random traffic dynamics and tested on a yet unseen benchmark. This allows us to quantify the generalisability of our method explicitly. In Section 6.1, we analysed various performance measures on multiple benchmarks (test scenarios) with several well-known baselines. As arterial scenarios are of specific interest for the community (Wei et al. 2019a), we conduced a detailed investigation of our model's generalisability on such scenario types (see Section 6.2). The software specifications of our implementations can be found in our open-sourced code[11].

Exchangeable Policy Heads The learnable hierarchical state encoding Section 4 maps states to action (phase) energies. The policy control function maps from this action energy space to action probabilities. This results in maximum flexibility when it comes to the policy function. In this work, we chose a Double DQN (Hasselt, Guez, and Silver 2016) and a A2C (Peng et al. 2018) as policy heads, but any other can be used instead.

6.1 Generalising different Scales

A general traffic signal control policy should be able to generalise from single intersections to more complex road networks. We demonstrate this ability by conducting experiments on either end. Table 1 compares our models to different baselines on two single-intersection benchmarks. We analysed the number of vehicles, as for a single intersection this measure seems the most reasonable. We found that both *TransferLight* variants outperform all baselines on *Cologne1* and perform quasi on par with MaxPressure, causing the least congestion.

To analyse how are policy scales to more complex road networks, we conducted an experiment on *Cologne8* comprising 8 signalised intersections. We measured multiple

[11]`https://github.com/johSchm/TransferLight`

Table 1: Average number of standing vehicles (↓) over 3600 simulated time steps (TL = TransferLight).

	Cologne1	*Ingolstadt1*
Random	40.90 ±21.77	8.41 ±6.34
FixedTime	14.58 ±8.37	7.04 ±6.95
MaxPressure	8.00 ±5.22	1.88 ±1.38
TL-DQN	6.70 ±5.72	1.93 ±2.08
TL-A2C	7.21 ±7.04	2.30 ±3.30

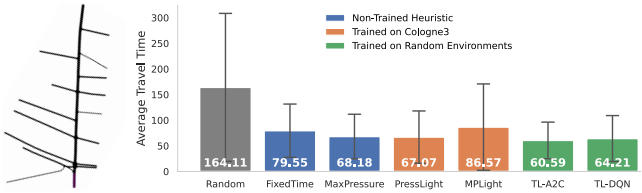

Figure 5: Average Travel Time on *Cologne3* over 3600 simulated time steps.

popular traffic performance indicators during testing. We found that both *TransferLight* variants outperformed all heuristic and trained baselines. Note that, CoLight (Wei et al. 2019b) and SOTL (Reztsov 2014) are explicitly trained on *Cologne8*, whereas *TransferLight* generalises from random road-networks. Both trained baselines failed to control a subset of intersections, leading to early congestions and hence the worse performance. The results in Fig. 4 undermine the ability of *TransferLight* to generalise also to more complex scenarios. In the appendix, we rise the problem complexity even more to identify *TransferLight's* generalisation limits.

6.2 Arterial Signal Progression

A special type of coordination is signal progression, which attempts to coordinate the onset of green times of successive intersections along an arterial street in order to move road users through the major roadway as efficiently as possible

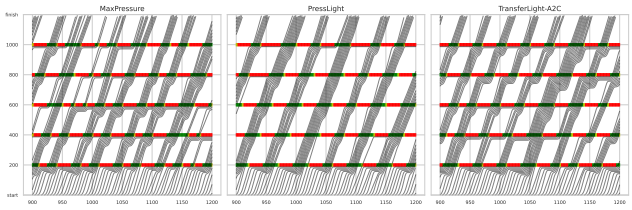

Figure 6: Signal progression comparison on a synthetic 5-intersection arterial scenario. PressLight (Wei et al. 2019a) was explicitly trained and designed to fit this specific scenario, whereas *TransferLight* generalises from random non-arterial road-networks.

(Wang, Abdulhai, and Sanner 2023). Intuitively, the hope here is to create a green wave in which green times are cascaded so that a large group of vehicles (also called a platoon) can pass through the arterial street without stopping.

PressLight (Wei et al. 2019a) and MaxPressure were shown to maximise throughput and minimise travel time in arterial environments. We compare *TransferLight* to these baselines while *not* being trained on arterial scenarios (other than PressLight). Here, the state transition prior is essential to provide geometric information to perform proactive decisions. Figure 6 shows the spatio-temporal signal progression plots, where each gray line represents the trajectory of a single vehicle. In the optimal case, vehicle trajectories form straight lines (i.e., they keep a constant velocity). We found that the zero-shot performance of our model can keep up with the performances of the baselines. In Fig. 5, we extended the experiment to a real-world scenario. We found that *TransferLight* was able to achieve the minimal travel time among the contesters, including MPLight (Chen et al. 2020) and PressLight. Our model learns a more robust and general policy from the DR-based training, enhancing its effectiveness in real-world environments characterized by greater variability.

7 Conclusion

We presented a novel framework designed for robust generalization across road-networks, diverse traffic conditions and intersection geometries. Our method can scale to any road-network through a decentralized multi-agent approach with global rewards and state transition priors to ensure proactive decisions. We used a heterogeneous and directed graph neural network to encode any intersection geometry, which we train using a novel log-distance reward function. Generalization is further fostered by domain randomization during training. This is particularly valuable for real-world applications where traffic conditions can vary significantly due to events, road closures, or long-term changes in urban mobility patterns.

Limitations and Future Work Our method shows already striking generalisation capabilities, which, however, need further improvement to cope with even larger road networks. In future work, we aim to extend the concept of symmetry breaking to the intersection's geometries. Mapping intersections to canonical forms, as in Jiang et al. (2024), collapses the state space to an exponentially smaller subspace. These canonical forms can be obtained from equivariant encodings (van der Pol et al. 2022) using canonicalisation priors (Kaba et al. 2023; Mondal et al. 2023) or by search (Schmidt and Stober 2024). This will drastically improve the sample efficiency of our model and render domain randomisation useless.

8 Acknowledgments

We would like to thank the Thorsis Innovation GmbH and Galileo Test-Track team for valuable support throughout this work. Furthermore, the authors acknowledge the financial support by the Federal Ministry of Education and Research

of Germany (BMBF) within the framework for the funding for the project PASCAL.

References

Brody, S.; Alon, U.; and Yahav, E. 2022. How Attentive are Graph Attention Networks? In *International Conference on Learning Representations (ICLR)*.

Chen, C.; Wei, H.; Xu, N.; Zheng, G.; Yang, M.; Xiong, Y.; Xu, K.; and Zhenhui. 2020. Toward A Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control. In *AAAI Conference on Artificial Intelligence*.

Choudhury, S.; Gupta, J. K.; Morales, P.; and Kochenderfer, M. J. 2021. Scalable Anytime Planning for Multi-Agent MDPs. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Chu, T.; Wang, J.; Codecà, L.; and Li, Z. 2019. Multi-Agent Deep Reinforcement Learning for Large-Scale Traffic Signal Control. *IEEE Transactions on Intelligent Transportation Systems*.

Devailly, F.-X.; Larocque, D.; and Charlin, L. 2022. IG-RL: Inductive Graph Reinforcement Learning for Massive-Scale Traffic Signal Control. *Trans. Intell. Transport. Sys.*

Devailly, F.-X.; Larocque, D.; and Charlin, L. 2024. Model-Based Graph Reinforcement Learning for Inductive Traffic Signal Control. *IEEE Open Journal of Intelligent Transportation Systems*.

Dijkstra, E. W. 1959. A note on two problems in connexion with graphs. *Numerische mathematik*.

Hasselt, H. v.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double Q-Learning. In *AAAI Conference on Artificial Intelligence*.

Jiang, H.; Li, Z.; Li, Z.; Bai, L.; Mao, H.; Ketter, W.; and Zhao, R. 2024. A General Scenario-Agnostic Reinforcement Learning for Traffic Signal Control. *Trans. Intell. Transport. Sys.*

Jiang, Y.; Kolter, J. Z.; and Raileanu, R. 2024. On the importance of exploration for generalization in reinforcement learning. In *International Conference on Neural Information Processing Systems (NeurIPS)*.

Kaba, S.-O.; Mondal, A. K.; Zhang, Y.; Bengio, Y.; and Ravanbakhsh, S. 2023. Equivariance with Learned Canonicalization Functions. In *International Conference on Machine Learning (ICML)*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

Korecki, M.; Dailisan, D.; and Helbing, D. 2023. How Well Do Reinforcement Learning Approaches Cope With Disruptions? The Case of Traffic Signal Control. *IEEE Access*.

Little, J. D. C.; Kelson, M. D.; and Gartner, N. H. 1981. MAXBAND : a versatile program for setting signals on arteries and triangular networks. Working papers.

Littman, M. L. 1994. Markov games as a framework for multi-agent reinforcement learning. In Cohen, W. W.; and Hirsh, H., eds., *Machine Learning Proceedings*.

Lopez, P. A.; Behrisch, M.; Bieker-Walz, L.; Erdmann, J.; Flötteröd, Y.-P.; Hilbrich, R.; Lücken, L.; Rummel, J.; Wagner, P.; and Wießner, E. 2018. Microscopic Traffic Simulation using SUMO. In *IEEE International Conference on Intelligent Transportation Systems*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

Mei, H.; Lei, X.; Da, L.; Shi, B.; and Wei, H. 2023. Libsignal: an open library for traffic signal control. *Machine Learning*.

Mondal, A. K.; Panigrahi, S. S.; Kaba, S.-O.; Rajeswar, S.; and Ravanbakhsh, S. 2023. Equivariant Adaptation of Large Pretrained Models. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Oroojlooy, A.; Nazari, M.; Hajinezhad, D.; and Silva, J. 2020. AttendLight: universal attention-based reinforcement learning model for traffic signal control. In *International Conference on Neural Information Processing Systems (NeurIPS)*.

Peng, B.; Li, X.; Gao, J.; Liu, J.; Chen, Y.-N.; and Wong, K.-F. 2018. Adversarial Advantage Actor-Critic Model for Task-Completion Dialogue Policy Learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Reztsov, A. 2014. Self-Organising Traffic Lights (SOTL) as an Upper Bound Estimate. *SSRN Electronic Journal*, 24.

Roess, R.; Prassas, E.; and McShane, W. 2004. *Traffic engineering*. Prentice Hall.

Schmidt, J.; Köhler, B.; and Borstell, H. 2024. Reviving Simulated Annealing: Lifting its Degeneracies for Real-Time Job Scheduling. In *Hawaii International Conference on System Sciences (HICSS)*.

Schmidt, J.; and Stober, S. 2024. Tilt your Head: Activating the Hidden Spatial-Invariance of Classifiers. In *International Conference on Machine Learning (ICML)*.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Urbanik, T.; Tanaka, A.; Lozner, B.; Lindstrom, E.; Lee, K.; Quayle, S.; Beaird, S.; Tsoi, S.; Ryus, P.; Gettman, D.; Sunkari, S.; Balke, K.; and Bullock, D. 2015. NCHRP Report 812: A Guide for Applying Context-Sensitive Solutions for Signalized Intersections.

van der Pol, E.; van Hoof, H.; Oliehoek, F. A.; and Welling, M. 2022. Multi-Agent MDP Homomorphic Networks. In *International Conference on Learning Representations (ICLR)*.

Varaiya, P. 2013. Max pressure control of a network of signalized intersections. *Transportation Research Part C: Emerging Technologies*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *International Conference on Neural Information Processing Systems (NeurIPS)*.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2017. Graph Attention Networks. *International Conference on Learning Representations (ICLR)*.

Wang, M.; Xiong, X.; Kan, Y.; Xu, C.; and Pun, M.-O. 2024. UniTSA: A Universal Reinforcement Learning Framework for V2X Traffic Signal Control. *IEEE Transactions on Vehicular Technology*.

Wang, X.; Abdulhai, B.; and Sanner, S. 2023. A Critical Review of Traffic Signal Control and a Novel Unified View of Reinforcement Learning and Model Predictive Control Approaches for Adaptive Traffic Signal Control. In *Handbook on Artificial Intelligence and Transport*.

Wang, Z.; Chen, J.; and Chen, H. 2021. EGAT: Edge-Featured Graph Attention Network. In *Artificial Neural Networks and Machine Learning (ICANN)*.

Webster, F. 1958. *Traffic Signal Settings*. Road research technical paper.

Wei, H.; Chen, C.; Zheng, G.; Wu, K.; Gayah, V.; Xu, K.; and Li, Z. 2019a. PressLight: Learning Max Pressure Control to Coordinate Traffic Signals in Arterial Network. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Wei, H.; Xu, N.; Zhang, H.; Zheng, G.; Zang, X.; Chen, C.; Zhang, W.; Zhu, Y.; Xu, K.; and Li, Z. 2019b. CoLight: Learning Network-level Cooperation for Traffic Signal Control. In *ACM International Conference on Information and Knowledge Management (CIKM)*.

Wei, H.; Zheng, G.; Gayah, V.; and Li, Z. 2021. Recent Advances in Reinforcement Learning for Traffic Signal Control: A Survey of Models and Evaluation. *SIGKDD Explor. Newsl.*

Wei, H.; Zheng, G.; Gayah, V. V.; and Li, Z. J. 2019c. A Survey on Traffic Signal Control Methods. *ArXiv*, abs/1904.08117.

Wei, H.; Zheng, G.; Yao, H.; and Li, Z. 2018. IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Wu, C.; Kim, I.; and Ma, Z. 2023. Deep Reinforcement Learning Based Traffic Signal Control: A Comparative Analysis. *Procedia Computer Science*. International Conference on Ambient Systems, Networks and Technologies Networks (ANT) and International Conference on Emerging Data and Industry 4.0 (EDI40).

Yi, Y.; Li, G.; Wang, Y.; and Lu, Z. 2024. Learning to share in multi-agent reinforcement learning. In *International Conference on Neural Information Processing Systems (NeurIPS)*.

Yoon, J.; Ahn, K.; Park, J.; and Yeo, H. 2021. Transferable traffic signal control: Reinforcement learning with graph centric state representation. *Transportation Research Part C: Emerging Technologies*.

Yu, R.; Wang, Z.; Wang, Y.; Li, K.; Liu, C.; Duan, H.; Ji, X.; and Chen, J. 2023. LaPE: Layer-adaptive Position Embedding for Vision Transformers with Independent Layer Normalization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zang, X.; Yao, H.; Zheng, G.; Xu, N.; Xu, K.; and Li, Z. 2020. MetaLight: Value-based Meta-reinforcement Learning for Traffic Signal Control. In *AAAI Conference on Artificial Intelligence*.

Zhang, Z.; Yang, J.; and Zha, H. 2020. Integrating Independent and Centralized Multi-agent Reinforcement Learning for Traffic Signal Network Optimization. In *International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Zheng, G.; Xiong, Y.; Zang, X.; Feng, J.; Wei, H.; Zhang, H.; Li, Y.; Xu, K.; and Li, Z. 2019. Learning Phase Competition for Traffic Signal Control. In *ACM International Conference on Information and Knowledge Management (CIKM)*.

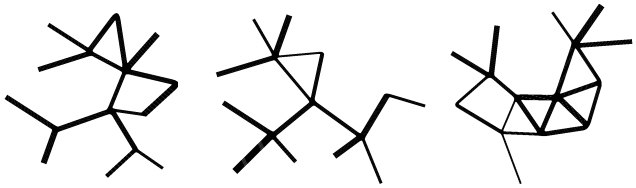

Figure 7: Three random road-network samples (static environments) used during training.

A Supplementary Material

A.1 Implementation Details

A replay buffer is introduced to decorrelate the experience tuples used for updating the parameters of the online DQN. For the A2C tuples are promptly utilized to perform immediate updates. This immediacy is crucial, as estimating the policy gradient necessitates the use of experience tuples generated from the current policy. All learnable functions are MLPs incorporating additional intermediate layers for layer normalization and dropout. This design aims to enhance training stability and convergence. We used a 64-dimensional latent space, which is significantly smaller than all our baselines, saving computational resources, allowing for better scalability and faster inference. We used 8 attention heads for all attention-based graph layer (see Section 4). All experiments are performed on an Nvidia A40 GPU (48GB) node with 1 TB RAM, 2x 24core AMD EPYC 74F3 CPU @ 3.20GHz, and a local SSD (NVMe). As the inference costs are generally extremely cheap, the available resources are only required to amplify training. More details can be found in our open-sourced code base.

Simulation Details We used the SUMO (Simulation of Urban MObility) (Lopez et al. 2018) during all our experiments. As in (Wei et al. 2019a), each action persists for a duration of 10 seconds before the next action can be chosen. To ensure safety, every transition from one phase to another involves a 3-second yellow-change interval followed by 2-second all-red interval to clear the intersection.

For our random training environments, we simulated passenger cars as vehicles only. The same is true for *cologne1*, *cologne3* and *cologne8*. In addition to vanilla passenger cars, *ingolstadt1*, *ingolstadt7* and *ingolstadt21* also include buses.

Segment Length We choose a segment length ds of 10 meters for our experiments. In theory, ds is only upper bounded to ℓ. However, we argue in favour of tighter bounds in practice. Using segment densities as inputs factors out the number of vehicles and enables the encoder to learn segment lengths implicitly (if needed). This requires a globally consistent segment length ds (as mentioned in Section 4). Hence, we drop possible remainders of $\frac{\ell}{ds}$. The impact is minor for any reasonable choices of ds, as the cut-off is done at the end of the lane (maximally distant to the intersection). This approach introduces a technical upper bound to the choice of ds, as short lanes with $\ell < ds$ would be dropped. Furthermore, choosing ds smaller than the average vehicle length h does not offer more valuable information for the encoding. Therefore, we have $h \leq ds \leq \min_{\ell \in \mathcal{I} \cup \mathcal{O}} \ell$.

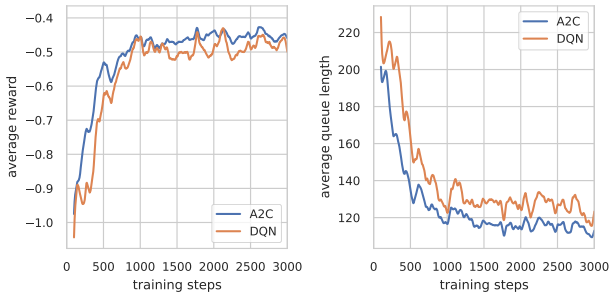

Figure 8: Average log-pressure reward (↑) and the average queue length (↓) over 3000 training steps.

Training Details For optimisation, AdamW (Loshchilov and Hutter 2017) with a learning rate of $1e-3$ and otherwise default settings is utilised. Furthermore, we used mini-batches of 64 SAR (state, action, reward) samples. We operate within a finite horizon of $0 \geq t \geq T$. We also include a convergence illustration in Fig. 8. We found that both model versions converge within 3000 steps (as the performance stays within reasonable error-bounds constant afterwards). We skipped the first 100 steps to let the traffic spawn in the simulation and develop a natural flow.

Baseline Details All heuristics (incl. Random, FixedTime, and MaxPressure) are custom implementations. All trainable baselines and related performance results are obtained using LibSignal (Mei et al. 2023). Nonetheless, we used the same routines to compute the high-level performance indicators presented in our performance plots.

A.2 Theory

Runtime Complexity Signals on the segment level are embedded in parallel and aggregated by Eq. (9). Let $S_v := \sum_{\ell \in \mathcal{I}_v \cup \mathcal{O}_v} \frac{\ell}{ds}$ be the number of segments at v. Due to the quadratic complexity of the attention mechanism in each of the three encoding layers, we get

$$\mathcal{O}\left(S_v|\mathcal{M}_v| + |\mathcal{M}_v||\Phi_v| + \Phi_v^2 + P_v\right), \tag{14}$$

where $P_v := \sum_{\ell \in \mathcal{I}_v \cup \mathcal{O}_v} |\mathcal{M}_{\rightarrow \ell}| + |\mathcal{M}_{\leftarrow \ell}|$ is the number of movements required to estimate the transition priors in Eq. (7). Note that $P_v \ll |\mathcal{M}|$ as only the prior only considers movements connecting two intersections. For a reasonable segment length ds, $S_v|\mathcal{M}_v|$ is the dominant term in Eq. (14). As each intersection is controlled by weight-tied agents, policies can be computed in parallel for all v, disentangling the number of intersections from the runtime complexity.

A.3 Ablation Studies

Reward Comparison Fig. 11 compares the performance gains through our symmetry-breaking log-distance reward. We found that our log-distance reward improves all three target performance indicators over the simulated test span. These empirical results underpin our theoretical claims in Section 3.

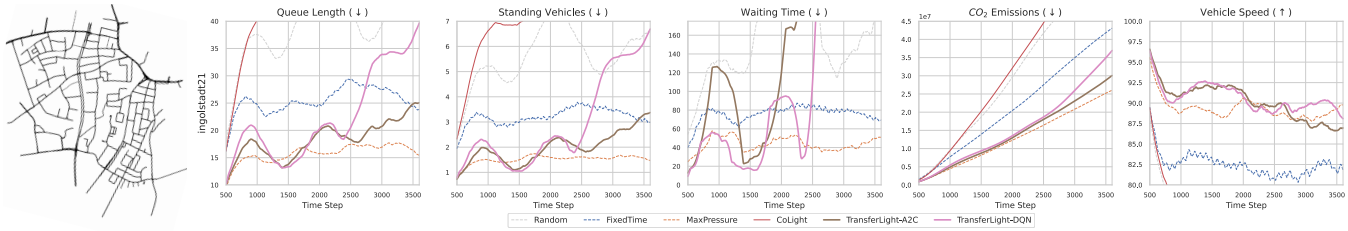

Figure 9: Test performances (moving averages) on *Ingolstadt21* over 3600 simulated time steps. This is an example of the limits of generalisation capabilities of *TransferLight*. Around step 2400 a congestion is builds up around a few intersections which miscalculated some phase energies. Afterwards, it was not able to resolve the knot and the congestion spread across the network.

pe	$\gamma_{m\phi}$	$J_{\phi\phi'}$	$\bar{\rho}_\ell$	ATT (↓)
	✓	✓	✓	92.13 ±58.18
✓		✓	✓	144.52 ±107.83
✓	✓		✓	92.15 ±58.22
✓	✓	✓		92.80 ±58.23
✓	✓	✓	✓	**91.71 ±56.51**

Figure 10: Ablation study of positional encoding pe, edge features $\gamma_{m\phi}$ and $J_{\phi\phi'}$, and our transition prior $\bar{\rho}_\ell$ comparing Average Travel Time (ATT) on *cologne8*. Left: visualization of the cologne8 network. Right: ablation results.

Impact of Positional Encoding When lane segments are relatively short on average, the inclusion of positional encoding at the segment level has limited influence on model performance. This condition is particularly prevalent in urban environments, where signalized intersections are typically located in close proximity to one another. Under such circumstances, positional encoding can be effectively disregarded. Evidence supporting this claim is presented in Fig. 10, where the omission of positional encoding resulted in only a negligible increase in average travel time.

Impact of Transition Prior The removal of the state transition prior during both training and testing resulted in the second-largest performance degradation observed in our ablation study (see Fig. 10). The transition prior plays a pivotal role in enabling the model to make proactive decisions rather than merely reactive ones, a capability that is especially critical in dense, urban environments. However, the observed impact of this prior was less substantial than initially anticipated. This discrepancy warrants further investigation, and we plan to explore the underlying reasons in greater detail in future work.

Impact of Edge Features To introduce a measure of similarity between traffic signal phases (in terms of their green signal overlap), we employed the Jaccard index at the phase-to-phase level. As demonstrated in Fig. 10, this feature provided only a marginal improvement in performance. At the movement-to-phase level, we incorporated edge feature encodings to represent whether a movement was prohibited, protected, or permitted. By a wide margin, this feature had the most significant impact on overall performance. We hypothesize that this added inductive bias enables the model to cluster movements more effectively before decoding them into phase energy distributions. Consequently, the model focuses primarily on traffic densities that can be directly alleviated by selecting specific phases. Investigating this claim more thoroughly is another key point for potential follow-up work.

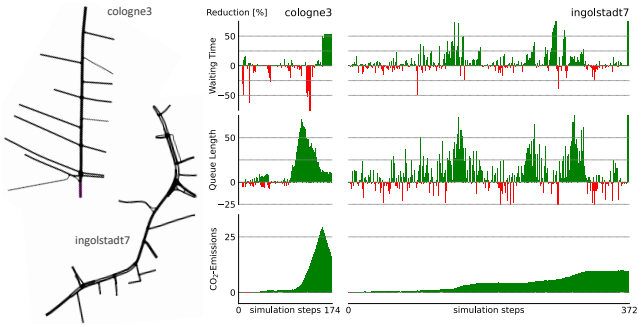

Figure 11: Waiting time, queue length, and emission reduction using our log-distance pressure reward (Eq. (5)) compared to the commonly used pressure reward (Eq. (2)).

A.4 Limits of Generalisability

The ability to generalise is of course facing limits at some range of problem complexity. We performed an additional experiment on *ingolstadt21* comprising 21 intersections in a narrow urban environment. Figure 9 compares our method to various baselines under different performance measures on this benchmark scenario. After around 1200 time steps, *TransferLight* with either head starts diverging into a sub-optimal sequence of phases. On the long run, this leads to congestions, which in turn lead to performance decreases among all measures. We dedicate our future work to indicate the causing factors and prevent such situations to occur (under reasonable traffic demands).