HeteroLight: A General and Efficient Learning Approach for Heterogeneous Traffic Signal Control

Yifeng Zhang[1], Peizhuo Li[1], Mingfeng Fan[2], Guillaume Sartoretti[1]

***Abstract*— Efficient and scalable adaptive traffic signal control is crucial in reducing congestion, maximizing throughput, and improving mobility experience in ever-expanding cities. Recent advances in multi-agent reinforcement learning (MARL) with parameter sharing have significantly improved the adaptive optimization of large-scale, complex, and dynamic traffic flows. However, the limited model representation capability due to shared parameters impedes the learning of diverse control strategies for intersections with different flows/topologies, posing significant challenges to achieving effective signal control in complex and varied real-world traffic scenarios. To address these challenges, we present a novel MARL-based general traffic signal control framework, called HeteroLight. Specifically, we first introduce a General Feature Extraction (GFE) module, crafted in a decoder-only fashion, where we employ an attention mechanism to facilitate efficient and flexible extraction of traffic dynamics at intersections with varied topologies. Additionally, we incorporate an Intersection Specifics Extraction (ISE) module, designed to identify key latent vectors that represent the unique intersection's topology and traffic dynamics through variational inference techniques. By integrating the learned intersection-specific information into policy learning, we enhance the parameter-sharing mechanism, improving the model's representation diversity among different agents and enabling the learning of a more efficient shared control strategy. Through comprehensive evaluations against other state-of-the-art traffic signal control methods on the real-world Monaco traffic network, our empirical findings reveal that HeteroLight consistently outperforms other methods across various evaluation metrics, highlighting its superiority in optimizing traffic flows in heterogeneous traffic networks.**

I. Introduction

With the rapid growth of population and travel demands in urban areas, increasing traffic congestion has brought significant challenges to the existing traffic management systems. This congestion results in longer travel times, higher fuel consumption, and increased pollution, all of which significantly reduce the efficiency and comfort of urban mobility. Traditional traffic signal control methods, such as fixed-time control [1] and actuated control [2], [3], have been widely deployed in urban areas. However, these methods struggle to manage the dynamic nature of urban traffic, highlighting the need for more adaptable control solutions.

Recently, Multi-agent Reinforcement Learning (MARL) approaches have shown great potential in solving various complicated control tasks [4], [5], [6]. Given the complex nature of traffic management systems, centralized control methods, which let one central entity control all intersections, becomes impractical. Thus, the community has turned the attention to decentralized MARL methods for adaptive traffic signal control (ATSC), where the traffic network is modeled as a multi-agent system and each intersection is treated as an autonomous learning agent [7], [8]. However, those independent learning approaches often suffer from environmental instability, as surrounding agents simultaneously update their policies. The discrepancies in policy learning among agents can lead to unstable environments, thus impeding agents from learning efficient control strategies. Parameter sharing (PS), where all agents learn a shared policy, has become a popular mechanism applied in MARL algorithms, as it shows significant performance in improving data efficiency and reducing environmental instability in homogeneous multi-agent traffic signal control (MATSC) tasks [9], [10], [11], [12], [13], [14], [15], [16], [17]. However, the heterogeneity of the real-world traffic networks poses a significant challenge for conventional PS-based methods, as agents may have different state and action spaces due to the varying topologies and traffic flows of intersections. To tackle the heterogeneous traffic signal control problem, approaches such as FRAP [18], AttendLight [13], and OAM [17] have been developed, providing different universal learning frameworks for intersections with any configuration. However, they do not fully consider the static and dynamic specifics of intersections, which may cause the learned policy lack of diversity, thus leading to sub-optimal solutions.

To address these challenges, we introduce a novel MARL framework named HeteroLight, an efficient and scalable parameter-sharing-based learning method tailored for heterogeneous traffic signal control. Specifically, we first propose a General Feature Extraction (GFE) module, designed in a decoder-only fashion, which employs a cross attention mechanism to aggregate the crucial traffic dynamics for each available phase. This enhances the agent's ability to proficiently manage traffic signals by correlating the signal phase with traffic dynamics via the detailed movement states, thus enabling our method to adapt to intersections with diverse topologies. To further enhance the parameter-sharing mechanism, we introduce an Intersection Specifics Extraction (ISE) module, which integrates a Variational Autoencoder (VAE) that is designed to generate intersection-specific latent vectors through variational inference techniques. In particular, our ISE module takes the combination of the current traffic state vector, phase vector and the intersection topology vector as inputs, to reconstruct predictions for the next traffic state. This enables agents to integrate unique intersection

[1]Y. Zhang, P. Li, and G. Sartoretti are with the Department of Mechanical Engineering, National University of Singapore (E-mail: yifeng@u.nus.edu, e0376963@u.nus.edu, guillaume.sartoretti@nus.edu.sg).

[2]M. Fan is with School of Traffic and Transportation Engineering, Central South University, China (E-mail: mingfan2001@gmail.com).

topology details, along with the traffic flows and transition dynamics into a compact latent space. By incorporating time-variant latent vectors into the decision-making process, our approach greatly improves the model's capability to represent diverse traffic situations, thereby facilitating a more efficient shared policy learning among heterogeneous agents.

We conduct comprehensive evaluation experiments using the open-source SUMO simulator over the real-world heterogeneous Monaco traffic network (featuring 28 signalized intersections) under complex synthetic traffic demands. The results indicate that our method HeteroLight outperforms all benchmark methods in almost all evaluation metrics. Notably, we demonstrate that incorporating such learned latent vectors, which represent the static and dynamic characteristics of intersections, into the decision-making and policy learning process greatly improves the agents' representation capability for diverse traffic scenarios, thus leading to a substantial increase in performance. In doing so, this work not only underscores the superiority of HeteroLight in tackling heterogeneous traffic signal control challenges, but also offers important insights on enhancing the parameter-sharing framework for heterogeneous multi-agent systems.

II. Related Work

Traditional traffic signal control methods can be broadly classified into fixed-time control and adaptive control. Fixed-time control, as discussed by Roess et al. [1], operates based on a pre-determined phase cycle and timing of phases, yet it struggles to adapt to complex, changing traffic patterns. On the other hand, adaptive control systems, such as SCOOT [2] and SCATS [3] adjust signal plans in response to real-time traffic conditions, leveraging data collected from loop detectors for more responsive management. Furthermore, the advanced max-pressure control [19] optimizes traffic flow at intersections by minimizing the difference in the stopped vehicle counts between upstream and downstream roads.

Recently, MARL has emerged as a promising approach for ATSC tasks, demonstrating great potential in improving traffic flow efficiency. The majority of existing studies [8], [9], [10], [15], [20], [21], [12], [11], [16], [22], [23] have concentrated on the development of decentralized, parameter sharing-based MARL frameworks for network-wide ATSC systems. Specifically, PressLight [9] introduced the concept of pressure into the state and reward definitions of RL agents, designed to jointly optimize arterial traffic flows. To further enhance the agent cooperation and collaboration, approaches like CoLight [10], STMARL [24], and GPLight [11] have utilized graph neural networks to facilitate efficient communication and spatio-temporal feature extraction among connected agents. While NC-HDQN [21] concentrated on analyzing the correlations between adjacent agents, it then utilized these relationships to adjust agents' observations and rewards to foster neighborhood cooperation. SocialLight [12] presented a decentralized MARL algorithm with a refined counterfactual-based advantage calculation to achieve scalable cooperation. Furthermore, MetaVIM [16] enhanced policy generalizability of agents across different neighborhood sizes through the incorporation of latent variables, and designed an intrinsic reward to ensure a stable training process in dynamic traffic environments.

In heterogeneous traffic signal control, where agents encounter varied topologies and traffic flows, traditional parameter-sharing methods struggle to devise effective solutions due to the distinct agent state and action spaces. Independent learning approaches, like IA2C and MA2C [7], were developed to allow each agent to learn its own policy, yet they often lead to sub-optimal outcomes due to environmental instability. To overcome these issues, Zheng et al. introduced FRAP [18], a parameter-sharing method that leverages the principles of phase competition, tailored for intersections with diverse topologies and traffic patterns. MPLight [14] extended this by integrating pressure concepts with FRAP for effective learning in mega-scale traffic systems. AttendLight [13] introduced an attention-based learning framework for universal policy learning applicable to any intersection configuration. Similarly, Liang et al.'s Option-Action RL Framework (OAM) [17] simplified phase selection into lane options, offering a versatile solution for various intersection topologies. While these innovations offer greater flexibility, they still face challenges from the shared model's limited capability to represent diverse traffic scenarios, often resulting in sub-optimal solutions.

III. Background

A. Traffic Terminology

Definition 1 (Incoming and outgoing lanes): An incoming lane directs vehicles towards an intersection and an outgoing lane guides them away. Each road at the intersection comprises several lanes. The sets of these lanes are denoted as $\mathcal{L}_{in}$ for incoming and $\mathcal{L}_{out}$ for outgoing lanes, respectively.

Definition 2 (Traffic movements and movement status): A traffic movement is a specific route that vehicles take to navigate through an intersection, linking an incoming lane with an outgoing lane. In practice, an incoming lane can participate in multiple traffic movements as it may connect to several outgoing lanes. We denote a traffic movement from incoming lane l_{in} to outgoing lane l_{out} as $m_{(l_{in} \to l_{out})}$. The activation status of the movement is indicated by $m_{(l_{in} \to l_{out})} = 1$, allowing vehicles on the incoming lane to proceed. Conversely, $m_{(l_{in} \to l_{out})} = 0$ means vehicles on the incoming lane are prohibited from passing.

Definition 3 (Traffic signal phases): Traffic signal phases are implemented at intersections to ensure safe and efficient traffic management. Each traffic phase, denoted as p, consists of a set of non-conflicting traffic movements that are activated simultaneously, defined by $p = \{m = 1 \,|\, m \in \mathcal{M}^p\}$. Here, $\mathcal{M}^p$ represents the set of traffic movements for phase p. Given $\mathcal{P}$ as the set of all possible phases and $\mathcal{M}$ as the set of all traffic movements at an intersection, we establish $\mathcal{M} = \bigcup_{p \in \mathcal{P}} \mathcal{M}^p$, indicating that the complete set of movements is the union of all movements across the phases.

Definition 4 (Traffic agents and traffic networks): Traffic agents control intersection flows by managing phases and

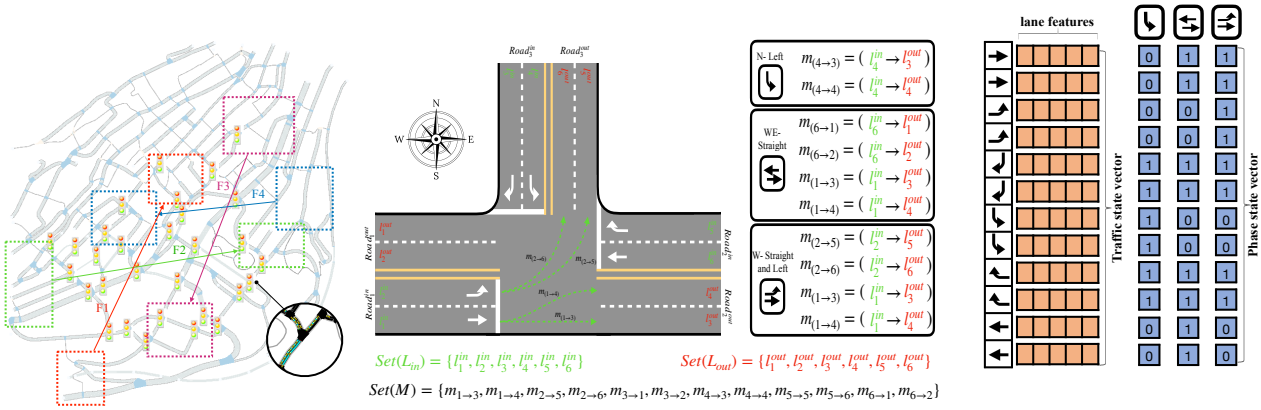

Fig. 1. (a) The heterogeneous real-world Monaco traffic network, which includes 28 signalized intersections and a set of synthetic traffic flows. (b) A three-armed intersection featuring three available phases, six incoming lanes, six outgoing lanes and twelve traffic movements. (c) Traffic state and phase state definitions in HeteroLight, with elements of both state ordered according to the same sequence of the traffic movements at the intersection.

phase timings. The traffic network is a multi-agent system composed of multiple traffic agents. These networks can be divided into heterogeneous networks, where agents have diverse topology structures (heterogeneous agents), and homogeneous networks, with identical topology structures throughout (homogeneous agents).

In real-world traffic networks, intersections are typically heterogeneous, i.e., they have different attributes such as lane counts, lane lengths, speed limits, signal phase settings, and traffic flows. Fig. 1(b) depicts a three-armed intersection with three incoming roads and three outgoing roads, where each road features two lanes (totalling six incoming lanes and six outgoing lanes). This results in twelve traffic movements by assuming that each incoming lane can be connected to multiple outgoing lanes. We also implement three traffic signal phases for this three-armed intersection, with the specifics of the corresponding activated traffic movements detailed on the right side of Figure 1(b).

B. Multi-agent Reinforcement Learning

Given a fully decentralized setting with each intersection controlled by an independent RL agent, we formulate the MATSC problem as a MARL problem, which can be represented as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) [25]. This Dec-POMDP can be formally defined as a tuple $\langle \mathcal{I}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathfrak{P}, \mathcal{R}, \gamma, \rho_0 \rangle$, with $\mathcal{I} = \{1, 2, \ldots, N\}$ representing the set of agents within a traffic network and $s \in \mathcal{S}$ indicating the global traffic states that cannot be observed by agents. For each time/decision step t, each agent $i \in \mathcal{I}$ draws a local observation o_i^t via the observation function $\mathcal{O}(s_t, i)$ and selects an action $a_i^t \in \mathcal{A}_i$ according to its policy $\pi(\cdot \mid o_i^t)$, which forms a joint action $a_t \in \mathcal{A}$. After executing a_t within the environment, each agent i receives an individual reward r_i^t from the reward function $\mathcal{R}_i(o_i^t, a_i^t)$. Subsequently, this joint action leads to the next state s^{t+1}, as governed by the transition function $\mathfrak{P}(s_{t+1} \mid s_t, a_t)$. Finally, γ represents the discount factor, and ρ_0 indicates the distribution of initial states. Thus, the ultimate goal of the multi-agent traffic system is to find an optimal joint policy π^* that can maximize the expected discounted return over all agents: $J(\pi) = \mathbb{E}_\tau \left[\sum_{i=1}^{|\mathcal{I}|} \sum_{t=1}^{t_e} \gamma^t r_i^t \right]$, where $\tau = \{(o^t, a^t, r^t)\}_{t=0}^{t_e}$ denotes the global trajectory with sequence length t_e.

IV. Methodology

A. RL Agent Design

In this section, we introduce the essential state, action, and reward definitions for our HeteroLight agents as follows:

1) State/Observation: In ATSC problems, lane-feature vectors, which aggregate various lane-specific data like queue lengths, vehicle counts, vehicle velocities, densities, and pressure, have been widely utilized to represent local traffic conditions in previous studies [12], [10], [14], [20], [13], [7]. Here, for each time step t, we define the state vector for a single traffic movement $m_{(l_{in} \to l_{out})}$ at an intersection i as a combination of five lane features, expressed as:

$$S_i^m(t) \in \mathbb{R}^5 = [P^{in}(t), Q^{in}(t), Q^{out}(t), N^{in}(t), N^{out}(t)], \tag{1}$$

where $P^{in}(t)$ indicates the current movement activation status, $Q^{in}(t)$ and $Q^{out}(t)$ the number of stopped vehicles (queue length) at the incoming and outgoing lanes for movement m, respectively, while $N^{in}(t)$ and $N^{out}(t)$ denote the number of moving vehicles on the respective incoming and outgoing lanes. All these lane features can be collected via digital cameras at intersections. Thus, the local traffic state vector for a single intersection i can be represented as:

$$S_i^t = S_i(t) \in \mathbb{R}^{|\mathcal{M}_i| \times 5} = [S_i^m(t) \mid m \in \mathcal{M}_i], \tag{2}$$

which includes the states of all available traffic movements. Additionally, we define the time-invariant phase state vector used to indicate the activation status of traffic movements for a given phase $p \in \mathcal{P}_i$ at the intersection as:

$$G_i^p \in \mathbb{R}^{|\mathcal{M}_i|} = [1 \text{ if } m \in \mathcal{M}_i^p \text{ else } 0 \mid m \in \mathcal{M}_i], \tag{3}$$

where $\mathcal{M}_i$ represents the set of all traffic movements of intersection i, and $\mathcal{M}_i^p$ denotes the subset of traffic movements activated by phase p. Thus, given the total phase set $\mathcal{P}_i$, we construct the comprehensive phase state vector G_i as:

$$G_i \in \mathbb{R}^{|\mathcal{P}_i| \times |\mathcal{M}_i|} = [G_i^p \mid p \in \mathcal{P}_i]. \tag{4}$$

A detailed illustration of a three-armed intersection's local traffic state and phase state definitions is presented in Fig. 1(c). Moreover, we formulate the time-invariant intersection topology vector of intersection i as follows:

$$I_i = [T_{\text{tl}}, L^{\text{in}}, V_{\text{max}}^{in}, N_{\text{l}}^{in}, N_{\text{m}}^{in}, L^{\text{out}}, V_{\text{max}}^{out}, N_{\text{l}}^{out}], \quad (5)$$

where T_{tl} (a one-hot vector) specifies the type of intersection/phase settings. L^{in} represents the average lane length of all incoming roads. V_{max}^{in} indicates the average maximum speed allowed on these incoming roads, N_{l}^{in} denotes the number of lanes on the incoming roads, and N_{m}^{in} the total count of traffic movements across all incoming roads. Similar metrics are included for the outgoing roads: L_{out}, V_{max}^{out}, and N_{l}^{out} denote the average lane length, maximum speed, and the number of lanes, respectively, for the outgoing roads. These components offer a comprehensive overview of the intersection's layout and traffic regulations, presenting a multidimensional view of its unique attributes.

2) Action: In this study, we define the action space for each agent as its finite set of collision-free traffic phases, where agents simultaneously select and implement a phase from these sets for a pre-determined duration, without being bound to a fixed cycle. This action setting is widely adopted in previous ATSC methods [9], [10], [14], [7], [12], since it can effectively skip the selection of unnecessary phases, thus maximizing control flexibility and overall efficiency.

3) Reward: We define the reward structure for each agent as the negative sum of queue lengths measured by lane-area detectors (with an effective detection range of 50 meters) installed at the incoming lanes near the intersection, which can be formulated as: $R(t) = -\left(\sum_{l_{in} \in \mathcal{L}_{in}} q_{l_{in}}\right)$, where $q_{l_{in}}$ denotes the queue length detected on the incoming lane l_{in}.

B. HeteroLight

The architecture of HeteroLight is illustrated in Fig. 2, where, for each agent, the GFE module takes its current traffic state vector and all phase vectors as input, utilizing a multi-head cross-attention mechanism to calculate the state-aggregated feature vector for each phase. Meanwhile, the ISE module combines each phase vector with the traffic state vector and the intersection topology vector to devise the latent vectors for that phase via variational inference techniques (specifically, a VAE). The GFE module's aggregated feature vector and the ISE module's latent vector, corresponding to the same phase vector, are then concatenated to calculate the policy function and value function. Detailed descriptions of the structure of our HeteroLight are provided below:

1) General Feature Extraction: To effectively facilitate parameter sharing mechanism for heterogeneous traffic signal control, we propose a General Feature Extraction (GFE) module, designed in a decoder-only manner, which aims to relax the assumption of fixed input and output dimension in conventional parameter-sharing approaches and enables efficient feature extraction of crucial intersection traffic dynamics for all different kinds of intersections/agents. Within our GFE module, we first transform the traffic state vector of agent i at decision time step t, denoted as S_i^t, into a higher-dimensional feature vector through a Multi-layer Perceptron (MLP), which comprises two linear layers. The resulting state feature vector $\mathbf{h}_s$ is represented as:

$$\mathbf{h}_s \in \mathbb{R}^d = \mathbf{MLP}_s^{(2)}\left(S_i^t\right), \quad (6)$$

where d denotes the feature dimensions. We then input this state feature vector into a recurrent neural network, specifically a Gated Recurrent Unit (GRU) [26], used to selectively integrate crucial historical local traffic state information while discarding the irrelevant ones. The resultant hidden feature vector $\mathbf{h}_s^{'}$ is derived as follows:

$$\mathbf{h}_s^{'} \in \mathbb{R}^d = \mathbf{GRU}\left(\mathbf{h}_s, \mathbf{h}_{(s,t-1)}^{GRU}\right), \quad (7)$$

where $\mathbf{h}_{(s,t-1)}^{GRU}$ denotes the hidden vector produced by the GRU at the preceding decision time step $t-1$. Similarly, we transform the complete phase state vector for the intersection G_i into a higher-dimensional feature vector using an additional MLP composed of two linear layers. The resulting phase feature vector $\mathbf{h}_p$ is then calculated as:

$$\mathbf{h}_p \in \mathbb{R}^{|\mathcal{P}_i| \times d} = \mathbf{MLP}_p^{(2)}\left(G_i\right). \quad (8)$$

We further employ a multi-head cross-attention mechanism, as introduced by [27], to generate phase-conditioned state feature vectors. Specifically, for each head h (totalling 4 heads), the query vector $\mathbf{Q}_h$, derived from the phase feature vector $\mathbf{h}_p$, is calculated as $\mathbf{Q}_h \in \mathbb{R}^{|\mathcal{P}_i| \times d} = \mathbf{h}_p\, W_h^Q$. The key vector $\mathbf{K}_h$ and the value vector $\mathbf{V}_h$ are generated from the aggregated state feature vector $\mathbf{h}_s^{'}$ produced by the GRU, calculated as: $\mathbf{K}_h \in \mathbb{R}^{1 \times d} = \mathbf{h}_s^{'} W_h^K$ and $\mathbf{V}_h \in \mathbb{R}^{1 \times d} = \mathbf{h}_s^{'} W_h^V$, respectively. Here, W_h^Q, W_h^K, and W_h^V are the learnable parameters of the cross-attention mechanism. The attention vector for each head is then determined through the scaled-dot product attention mechanism, represented as:

$$\text{Attention}_h(\mathbf{Q_h}, \mathbf{K_h}, \mathbf{V_h}) = \text{softmax}\left(\frac{\mathbf{Q}_h \left(\mathbf{K}_h\right)^T}{\sqrt{d}}\right)\mathbf{V}_h. \quad (9)$$

Lastly, we get the output through a linear layer after concatenating the attention vectors from all heads $\mathbf{h}_{sp} \in \mathbb{R}^{|\mathcal{P}_i| \times d} = \text{Concat}\left(\text{Attention}_1, \ldots, \text{Attention}_4\right) W^O$, where W^O represents the learnable parameters for the output layer. Hence, the GFE module's output $\mathbf{h}_{sp}$ aggregates crucial traffic states for each available phase at the intersection, guided by the specified phase state vectors, which will be further utilized to compute the output policy/action probabilities.

2) Intersection Specifics Extraction: Despite the effectiveness of parameter sharing in learning a scalable decentralized control strategy in complex traffic scenarios, it often leads to sub-optimal solutions, which are likely due to limited model representation power. To further enhance the parameter-sharing mechanism, we introduce the integration of latent vectors, derived through variational inference techniques, which aims to create a more accurate low-dimensional representation that can effectively capture the static topology and dynamic traffic patterns of various intersections. Specifically,

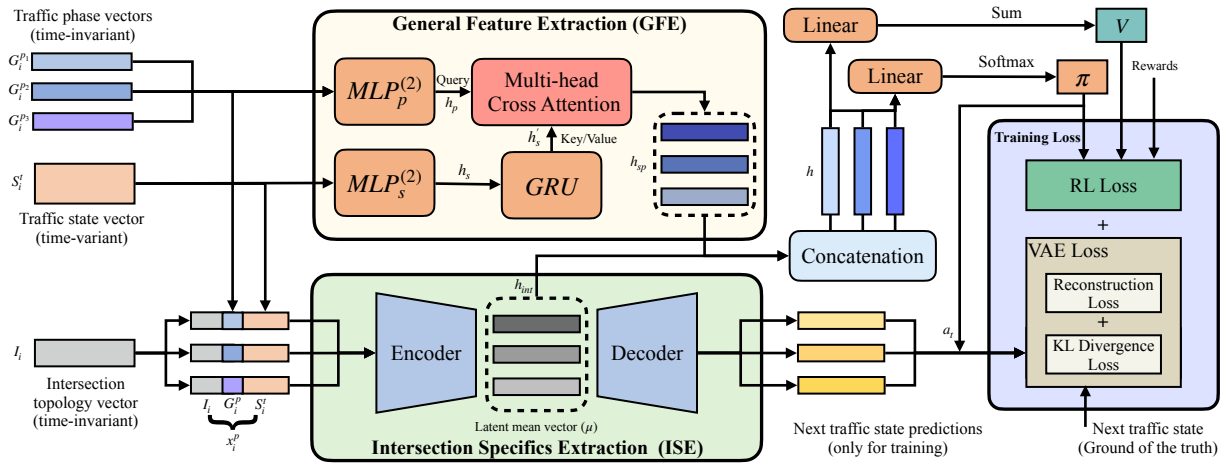

Fig. 2. Overall learning framework of HeteroLight, which incorporates a General Feature Extraction (GFE) module to facilitate efficient and flexible traffic dynamics extraction for varied intersections, as well as an Intersection Specifics Extraction (ISE) module to generate intersection-specific latent vectors.

this is achieved with a VAE [28], where the input of the agent i with a given phase p at time step t, denoted as $x_i^p(t) = [S_i^t, G_i^p, I_i]$, includes the current traffic state vector S_i^t, the phase state vector G_i^p, and the intersection topology vector I_i. Within the VAE structure, the encoder, parameterized by ξ, is designed to approximate the true posterior distribution $p(z_i^p \mid x_i^p)$ with a variational distribution $q_\xi(z_i^p \mid x_i^p)$. Particularly, it first encodes the input x_i^p to the parameters of a Gaussian distribution (the mean vector μ_i^p and the log-variance vector σ_i^p) as described by:

$$(\mu_i^p, \sigma_i^p) = \textbf{Encoder}_\xi(x_i^p). \tag{10}$$

Subsequently, a latent variable z_i^p is sampled from this distribution: $z_i^p = \mu_i^p + \sigma_i^p \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. The decoder, parameterized by ϕ, utilizes the latent variable z_i^p to reconstruct the prediction of next state $s_{i,p}^{t+1}$, expressed as:

$$s_{i,p}^{t+1} \sim p_\phi\left(s_{i,p}^{t+1} \mid z_i^p\right) = \textbf{Decoder}_\phi(z_i^p). \tag{11}$$

Importantly, we generate multiple predictions, each linked to a specific input phase vector, but only the prediction for the selected phase (denoted as $\bar{p}$) is actively used during each training step. The training objective aligns with maximizing the Evidence Lower Bound (ELBO) [28], denoted as L_i^{vae}, which serves as a proxy for maximizing the log-likelihood $\log p\left(\hat{s}_i^{t+1} \mid x_i^{\bar{p}}\right)$ of observing the next state $\hat{s}_i^{t+1}$ (the true next state vector) given $x_i^{\bar{p}}$. The ELBO is calculated as:

$$\mathbb{E}_{q_\xi(z_i^{\bar{p}}|x_i^{\bar{p}})}\left[\log p_\phi\left(\hat{s}_i^{t+1}|z_i^{\bar{p}}\right)\right] - \mathrm{KL}\left[q_\xi(z_i^{\bar{p}}|x_i^{\bar{p}}) \| p(z_i^{\bar{p}}|x_i^{\bar{p}})\right], \tag{12}$$

where the first term (reconstruction loss) denotes the expected log likelihood of the true next state under the decoder, $\mathbb{E}[\log p_\phi\left(\hat{s}_i^{t+1}|z_i^{\bar{p}}\right)]$. This emphasizes the reconstruction accuracy from latent space to output predictions, and maximizing this term encourages the VAE to generate decoder outputs $s_{i,\bar{p}}^{t+1}$ that are close to the true next states $\hat{s}_i^{t+1}$. Meanwhile, the Kullback-Leibler (KL) divergence (the second term) ensures that the posterior distribution $q_\xi(z_i^{\bar{p}} \mid x_i^{\bar{p}})$ stays close to the prior distribution $p(z_i^{\bar{p}}|x_i^{\bar{p}})$. For simplicity, the prior $p(z_i^{\bar{p}} \mid x_i^{\bar{p}})$ is often assumed to be a standard normal distribution $p(z_i^{\bar{p}}) = \mathcal{N}(0, I)$ that is not conditioned on $x_i^{\bar{p}}$.

In summary, for each available phase state vector at the intersection, we construct a VAE input and derive the corresponding intermediate latent vectors μ_i^p and σ_i^p. We then form the intersection-specific feature vector by computing the latent mean vectors for all phases, denoted as $\mathbf{h}_{int} \in \mathbb{R}^{|\mathcal{P}_i| \times d_{vae}} = [\mu_i^p \mid p \in \mathcal{P}_i]$. These latent vectors are essentially used for prediction reconstruction (i.e., predicting the actual next traffic state via supervised learning), and thus incorporate crucial information of the traffic flow dynamics. Our key insight here, is that augmenting the VAE input with additional phase state vectors and intersection topology vectors for different agents results in diverse and unique latent expressions, thereby significantly boosting the model's capability to represent various traffic scenarios.

3) Policy and Value Output: After obtaining the aggregated feature vector $\mathbf{h}_{sp}$ and the intersection-specific latent feature vector $\mathbf{h}_{int}$ from our GFE module and ISF module for all available phase vectors respectively, we concatenate them together to yield the final feature vector $\mathbf{h} \in \mathbb{R}^{|\mathcal{P}_i| \times (d + d_{vae})} = \textbf{Concat}(\mathbf{h}_{sp}, \mathbf{h}_{int})$. This feature vector is then used to calculate the policy function and value function of agent i through two different linear layers (denoted as $\mathbf{f}_\pi$ and $\mathbf{f}_v$, respectively), which can be expressed as:

$$\pi_i^t \in \mathbb{R}^{|\mathcal{P}_i| \times 1} = \textbf{Softmax}(\mathbf{f}_\pi(\mathbf{h})), V_i^t \in \mathbb{R}^1 = \textbf{Sum}(\mathbf{f}_v(\mathbf{h})). \tag{13}$$

C. Policy Optimization

We adopt the popular RL algorithm, Proximal Policy Optimization (PPO) [29], to update the policy function π_θ, parameterized by θ, and the value function V_Φ, parameterized by Φ that is shared among all agents within the network. Specifically, the policy loss for agent i is defined as:

$$L_i^a(\theta) = -\mathbb{E}_t\left[\min\left(\kappa_i^t(\theta)\hat{A}_i^t, \mathrm{clip}\left(\kappa_i^t(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i^t\right)\right], \tag{14}$$

where $\kappa(\theta)$ is the ratio of the probability between the current policy and the old policy, $\hat{A}$ is the advantage function calculated through the General Advantage Estimate (GAE) methods [30], and ϵ is used to determine the clip range. Additionally, the value loss in PPO is defined as follows:

$$L_i^c(\Phi) = \mathbb{E}_t \left[\left(r_i^t + \gamma V_{\Phi,i}^{t+1} - V_{\Phi,i}^t \right)^2 \right], \quad (15)$$

which aims to minimize the mean square error between the temporal difference errors (where r_i^t is the individual reward) and the predicted values. To encourage exploration and prevent premature convergence to sub-optimal policies, we also add an entropy loss $L_i^e(\theta)$, which is calculated as the expectation of the policy entropy. Combining these RL losses with the VAE loss $L_i^{vae} = -ELBO(\xi, \phi)$, derived from the ELBO in Eq. (12) for the ISE module, the final optimization loss for all N agents is formally written as:

$$L(\theta, \Phi) = \frac{1}{N} \sum_{i=1}^{N} (L_i^a + c_1 L_i^c - c_2 L_i^e + c_3 L_i^{vae}), \quad (16)$$

where c_1, c_2, and c_3 are constant coefficients that balance the values of the value loss, entropy loss, and VAE losses. To enhance training efficiency through batch processing, we employ a padding mechanism to standardize the lengths of traffic state vectors and phase vectors across all intersections, ensuring alignment with the maximum movement and phase counts for the given network. Padding elements are masked out during the decision-making and training processes.

V. Experiments

A. Experiment Settings

We evaluate our proposed method, HeteroLight, by comparing its efficacy against both conventional control strategies and learning-based methods on the open-source microscopic traffic simulator SUMO [31]. For all experiments, we align with the MA2C experimental settings with a 5-second green duration and a 2-second yellow light. Selecting the same phase results in a 5-second green light, while changing phases triggers a 2-second yellow light followed by a 3-second green light. Each episode runs for 3600 seconds, corresponding to 720 decision steps. For training, our hyperparameter configuration is as follows: a discount factor and GAE factor of 0.95, an actor learning rate of 0.0003, and a critic learning rate of 0.0005. The MLPs' feature dimensions are set to 128, and the VAE's latent dimension is set to 20. We set the scaling coefficients for value loss, entropy loss, and VAE prediction loss at 0.5, 0.0003, and 0.0001, respectively. Additionally, the clip ratio and update epochs for PPO learning are fixed at 0.2 and 6. In parallel, we train all learning-based baselines using their default parameter configurations within the same experimental setup.

Our evaluation spans 10 episodes, each initialized with distinct random seeds, to ensure consistency by using identical seeds for corresponding episodes. We adopt the traffic evaluation metrics as defined in MA2C [7], which includes average queue length (veh), average vehicle speed (m/s), trip completion rate (veh/s), average intersection delay (s/veh), average trip time (s), and average trip delay (s).

B. Traffic Datasets

In this study, we conduct experiments on a real-world heterogeneous Monaco traffic network, incorporating synthetic traffic flows that vary with time [7]. The Monaco map utilized in our experiments, depicted in Fig. 1(a), consists of 28 signalized intersections with varied topology structure and phase settings. To test the efficacy of various control methods in ATSC tasks, we adopt a series of time-variant traffic flows that are defined in the MA2C paper [7]. Specifically, four groups of traffic flows are generated within the network, each scaled as multiples of a "unit" flow of 325 vehicles per hour, with origins and destinations (O-D) randomly selected within the mapped area, as shown in Fig. 1(a). The first two groups F_1 and F_2 are simulated over the initial 40 minutes, following a pattern of $[1, 2, 4, 4, 4, 2, 1]$ unit flows at 5-minute intervals. The remaining groups F_3 and F_4 are introduced during a shifted time window from 15 minutes to 55 minutes.

C. Compared Methods

1) **Greedy** (conventional): Chooses the phase that releases the maximum queue length at intersections.
2) **IQL-LR** [7]: A linear regression-based Independent Q-Learning (IQL) algorithm where each local agent independently learns and adapts its unique policy.
3) **IQL-DNN** [7]: An enhanced IQL variant using deep neural networks (DNNs) instead of linear regression method for more accurate Q-function approximation.
4) **IA2C** [7]: Builds on IQL-LR, adopting the advantage actor-critic algorithm (A2C) for policy learning.
5) **MA2C** [7]: An advanced MARL approach that incorporates observations and fingerprints of nearby agents into ego agent's state for stable training, and also integrates neighbors' rewards to promote cooperation.
6) **AttendLight [13]**: An attention-based learning framework with parameter-sharing, designed for managing heterogeneous traffic signals, adaptable to various intersection layouts and signal phase configurations.
7) **IPPO-S**: A simplified variant of our algorithm that focuses on parameter sharing, retaining GFE module while omitting the ISE module for ablation study.

All independent learning methods, namely IQL-LR, IQL-DNN, IA2C, MA2C, allocate distinct parameters to each agent for independent updates. In contrast, methods such as AttendLight, IPPO-S, and HeteroLight employ parameter-sharing, allowing all agents to learn common parameters.

D. Results and Analysis

1) Overall Performance: Our comparative results are presented in Fig. 3, where we measure and record the values of all traffic metrics at each simulation step (totalling 3600 steps) and then compute the averages and standard deviations of these metrics across a consistent series of 10 test episodes. Our results show that the Greedy approach performs well for low-demand traffic but struggles with high-demand scenarios, leading to longer queue lengths as traffic increases. Independent learning methods like IQL-LR, IQL-DNN, and IA2C outperform the Greedy approach in managing queues under heavy traffic, benefiting from each agent's ability to independently adapt its signal control. However, this independence may cause policy variance and environmental instability, making it harder to develop efficient control strategies. MA2C improves upon independent learning by

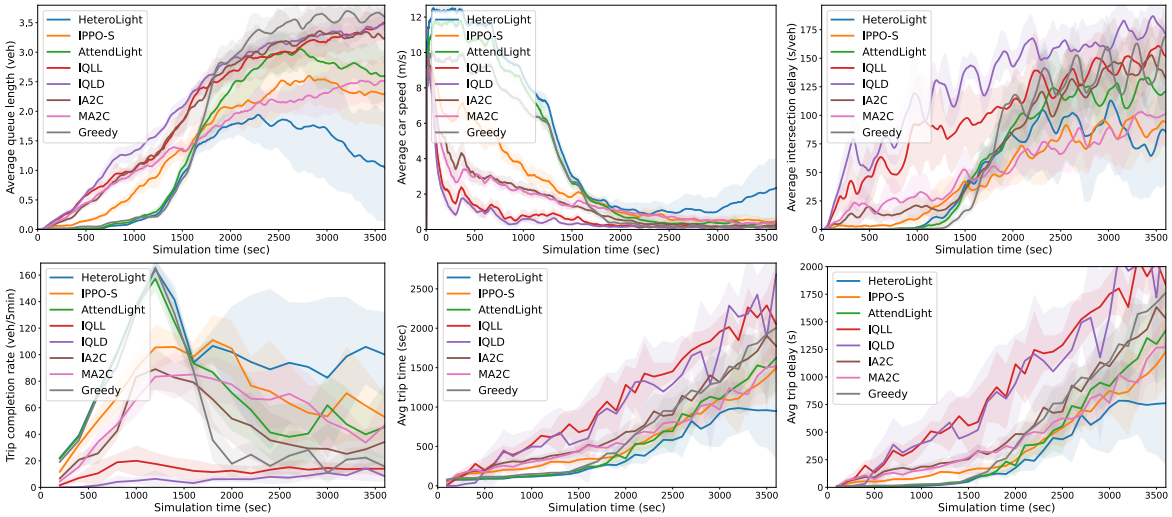

Fig. 3. Curves of six traffic metrics, namely queue length, speed, intersection delay, completion rate, trip time, and trip delay, over 3600 simulation steps for all methods during testing. Solid lines represent average values across 10 episodes, with shaded areas indicating standard deviations.

TABLE I

TEMPORAL AVERAGES (STANDARD DEVIATIONS) AND PEAKS (TROUGHS) ACROSS FIVE TRAFFIC METRICS FOR ALL METHODS OVER 10 EPISODES ON THE MONACO TRAFFIC NETWORK. THE TRAFFIC METRICS INCLUDE AVERAGE QUEUE LENGTH (↓), AVERAGE VEHICLE SPEED (↑), AVERAGE INTERSECTION DELAY (↓), TRIP COMPLETION RATE (↑), AND AVERAGE TRIP TIME (↓). HERE, ↓ DENOTES METRICS WHERE LOWER IS BETTER, AND ↑ WHERE HIGHER IS BETTER. THE BEST VALUES IN EACH COLUMN ARE BOLDED, AND THE SECOND-BEST VALUES ARE UNDERLINED.

Metrics (Average)	Temporal Averages (Standard Deviations)					Temporal Peaks (Troughs)				
	Queue Length ↓ (veh)	Speed ↑ (m/s)	Intersection Delay ↓ (s/veh)	Trip Completion Rate ↑ (veh/s)	Trip Time ↓ (s)	Queue Length ↓ (veh)	Speed ↑ (m/s)	Intersection Delay ↓ (s/veh)	Trip Completion Rate ↑ (veh/s)	Trip Time ↓ (s)
HeteroLight(ours)	**0.99 (0.74)**	**4.75 (4.39)**	50.25 (40.29)	**0.48 (0.26)**	404.28 (465.37)	**1.97 (0)**	**14.31 (0)**	116.95 (0)	1.60 (0)	**2365.00 (47.00)**
IPPO-S(ours)	1.50 (0.93)	2.43 (2.34)	**47.60 (34.65)**	0.36 (0.23)	544.66 (493.71)	2.65 (0)	12.60 (0)	**101.13 (0)**	**1.70 (0)**	2753.00 (64.00)
AttendLight	1.58 (1.24)	4.01 (4.39)	62.64 (52.90)	0.35 (0.26)	410.94 (473.54)	3.09 (0)	13.42 (0)	134.65 (0)	1.70 (0)	2500.00 (51.00)
MA2C	1.54 (0.78)	1.60 (1.45)	53.34 (28.88)	0.28 (0.20)	632.63 (505.03)	2.56 (0)	12.01 (0)	105.58 (0)	1.20 (0)	3079.00 (65.00)
IA2C	2.07 (1.18)	1.53 (1.74)	71.31 (53.00)	0.22 (0.19)	653.11 (538.80)	3.41 (0)	12.98 (0)	154.61 (0)	1.20 (0)	3092.00 (64.00)
IQL-LR	2.08 (1.13)	0.75 (1.23)	100.87 (41.03)	0.07 (0.08)	1172.54 (751.42)	3.56 (0)	13.10 (0)	163.83 (0)	0.50 (0)	3347.00 (62.00)
IQL-DNN	2.22 (1.12)	0.57 (1.15)	127.65 (44.74)	0.03 (0.06)	1569.26 (820.82)	3.52 (0)	12.96 (0)	189.10) (0)	0.50 (0)	3543.00 (64.00)
Geedy	1.91 (1.56)	3.37 (3.83)	70.08 (63.80)	0.26 (0.29)	**357.35 (459.75)**	3.73 (0)	13.40 (0)	171.43 (0)	1.70 (0)	2443.00 (60.00)

fostering communication and shared rewards among agents, resulting in significantly shorter queue lengths. AttendLight, utilizing parameter sharing, performs well in low-demand scenarios but faces challenges as traffic volume increases. Finally, HeteroLight excels over other methods by consistently keeping queue lengths low, demonstrating superior adaptability and recovery in high-demand conditions, particularly noticeable after 2500 seconds of simulation. We believe that the superior performance of HeteroLight may be attributed to its efficient parameter-sharing GFE module, which enhances data efficiency and reduces environmental instability, as well as to its ISE module, which integrates intersection specifics to further improve the shared policy learning in diverse traffic scenarios. Furthermore, HeteroLight shows comparable performance with other methods in minimizing intersection delay, while it surpasses other baseline approaches in maximizing average vehicle speed and trip completion rate, exhibiting its exceptional ability in easing congestion and optimizing overall network throughput. It also achieves the shortest average trip times and delays throughout the whole simulation, further highlighting its effectiveness in improving travel efficiency for heterogeneous traffic networks.

Additionally, we calculate the averages, standard deviations, maximum and minimum values for each traffic metric over 10 testing episodes and summarize the results in Table V-B. Here, HeteroLight excels in almost all assessed traffic metrics, particularly in maintaining the lowest queue length (0.99 veh) and highest vehicle speed (4.75 m/s), indicating superior traffic flow and congestion management. It also achieves the shortest trip time (404.28 s) and a high trip completion rate (0.48 veh/s), suggesting efficient vehicle throughput. In comparison, IPPO-S and AttendLight only achieve more moderate performance, while IA2C, IQL-LR, and IQL-DNN exhibit suboptimal performance in the face of dynamic, heterogeneous traffic scenarios. Conversely, the Greedy control method outperforms other methods in terms of average trip time, potentially because the calculation of trip time includes only vehicles that have completed their journeys. This results in an inaccurate estimation of the actual average trip time, given the substantial number of vehicles still present within the traffic network.

2) Ablation Study: Evaluation results from Fig. 3 and Table V-B show that, compared with the universal learning method AttendLight, IPPO-S (our ablation variant which removes the ISE module) only achieves modest improvements in average queue length and intersection delay, while AttendLight performs slightly better in average speed and trip time. This suggests that our GFE module can match the performance of AttendLight's universal model with greater parameter efficiency, likely due to its decoder-only design. Furthermore, comparisons between HeteroLight and IPPO-S underscore the significant role of the ISE module in enhanc-

ing shared policy learning, with HeteroLight outperforming IPPO-S in almost all metrics. We believe that this emphasizes how introducing intersection-specific latent vectors through variational inference can effectively enhance the agents' representation ability, resulting in marked performance gains.

VI. Conclusion

In this work, we introduce HeteroLight, a novel MARL framework for heterogeneous traffic signal control. Specifically, we introduce a General Feature Extraction (GFE) module that employs a cross-attention mechanism to enable efficient and adaptable feature extraction at intersections of varied topologies. We also develop an Intersection Specifics Extraction (ISE) module to dynamically generate latent vectors capturing intersection topology and traffic distribution information through variational inference techniques. By integrating these latent vectors into the decision-making process, we enhance agents' ability to represent diverse traffic scenarios, thus facilitating improved shared policy learning. We evaluate HeteroLight against various control methods on the real-world Monaco network, demonstrating its superiority in managing heterogeneous traffic flows.

In future work, we aim to further test the robustness and generalizability of HeteroLight across a broader range of heterogeneous traffic networks. Furthermore, we plan to investigate our method's application in other heterogeneous multi-agent/-robot systems, enhancing it as a general approach to facilitate efficient shared strategy learning for agents/robots with diverse roles or state/action spaces.

Acknowledgement

This research is supported by A*STAR, CISCO Systems (USA) Pte. Ltd and National University of Singapore under its Cisco-NUS Accelerated Digital Economy Corporate Laboratory (Award I21001E0002).

References

[1] R. P. Roess, *Traffic engineering*. United states of Anerica, 2004.

[2] P. Hunt, D. Robertson, R. Bretherton, and M. C. Royle, "The scoot on-line traffic signal optimisation technique," *Traffic Engineering & Control*, vol. 23, no. 4, 1982.

[3] L. PR, "Scats: A traffic responsive method of controlling urban traffic control/pr lowrie," *Roads and Traffic Authority*, 1992.

[4] O. Vinyals, T. Ewalds, S. Bartunov, P. Georgiev, A. S. Vezhnevets, M. Yeo, A. Makhzani, H. Küttler, J. Agapiou, J. Schrittwieser, *et al.*, "Starcraft ii: A new challenge for reinforcement learning," *arXiv preprint arXiv:1708.04782*, 2017.

[5] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Neural Information Processing Systems (NIPS)*, 2017.

[6] M. Damani, Z. Luo, E. Wenzel, and G. Sartoretti, "Primal _2: Pathfinding via reinforcement and imitation multi-agent learning-lifelong," *IEEE RA-L*, vol. 6, no. 2, pp. 2666–2673, 2021.

[7] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 3, pp. 1086–1095, 2019.

[8] H. Wei, G. Zheng, H. Yao, and Z. Li, "Intellilight: A reinforcement learning approach for intelligent traffic light control," in *KDD '18*, pp. 2496–2505, 2018.

[9] H. Wei, C. Chen, G. Zheng, K. Wu, V. Gayah, K. Xu, and Z. Li, "Presslight: Learning max pressure control to coordinate traffic signals in arterial network," in *KDD '19*, pp. 1290–1298, 2019.

[10] H. Wei, N. Xu, H. Zhang, G. Zheng, X. Zang, C. Chen, W. Zhang, Y. Zhu, K. Xu, and Z. Li, "Colight: Learning network-level cooperation for traffic signal control," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1913–1922, 2019.

[11] Y. Liu, G. Luo, Q. Yuan, J. Li, L. Jin, B. Chen, and R. Pan, "Gplight: grouped multi-agent reinforcement learning for large-scale traffic signal control," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 199–207, 2023.

[12] H. Goel, Y. Zhang, M. Damani, and G. Sartoretti, "Sociallight: Distributed cooperation learning towards network-wide traffic signal control," *arXiv preprint arXiv:2305.16145*, 2023.

[13] A. Oroojlooy, M. Nazari, D. Hajinezhad, and J. Silva, "Attendlight: Universal attention-based reinforcement learning model for traffic signal control," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4079–4090, 2020.

[14] C. Chen, H. Wei, N. Xu, G. Zheng, M. Yang, Y. Xiong, K. Xu, and Z. Li, "Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, pp. 3414–3421, 2020.

[15] X. Zang, H. Yao, G. Zheng, N. Xu, K. Xu, and Z. Li, "Metalight: Value-based meta-reinforcement learning for traffic signal control," in *AAAI Conf. Artif. Intell.*, vol. 34, pp. 1153–1160, 2020.

[16] L. Zhu, P. Peng, Z. Lu, and Y. Tian, "Metavim: Meta variationally intrinsic motivated reinforcement learning for decentralized traffic signal control," *IEEE Trans. Knowl. Data Eng.*, 2023.

[17] E. Liang, Z. Su, C. Fang, and R. Zhong, "Oam: An option-action reinforcement learning framework for universal multi-intersection control," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, pp. 4550–4558, 2022.

[18] G. Zheng, Y. Xiong, X. Zang, J. Feng, H. Wei, H. Zhang, Y. Li, K. Xu, and Z. Li, "Learning phase competition for traffic signal control," in *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 1963–1972, 2019.

[19] P. Varaiya, "Max pressure control of a network of signalized intersections," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 177–195, 2013.

[20] L. Zhang, Q. Wu, J. Shen, L. Lü, B. Du, and J. Wu, "Expression might be enough: Representing pressure and demand for reinforcement learning based traffic signal control," in *International Conference on Machine Learning*, pp. 26645–26654, PMLR, 2022.

[21] C. Zhang, Y. Tian, Z. Zhang, W. Xue, X. Xie, T. Yang, X. Ge, and R. Chen, "Neighborhood cooperative multiagent reinforcement learning for adaptive traffic signal control in epidemic regions," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 25157–25168, 2022.

[22] M. Wang, X. Xiong, Y. Kan, C. Xu, and M.-O. Pun, "Unitsa: A universal reinforcement learning framework for v2x traffic signal control," *IEEE Transactions on Vehicular Technology*, 2024.

[23] H. Jiang, Z. Li, Z. Li, L. Bai, H. Mao, W. Ketter, and R. Zhao, "A general scenario-agnostic reinforcement learning for traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[24] Y. Wang, T. Xu, X. Niu, C. Tan, E. Chen, and H. Xiong, "Stmarl: A spatio-temporal multi-agent reinforcement learning approach for cooperative traffic light control," *IEEE Transactions on Mobile Computing*, vol. 21, no. 6, pp. 2228–2242, 2020.

[25] F. A. Oliehoek, C. Amato, *et al.*, *A concise introduction to decentralized POMDPs*, vol. 1. Springer, 2016.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[28] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[29] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[30] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," *arXiv preprint arXiv:1506.02438*, 2015.

[31] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *21st IEEE Int. Conf. Intell. Transp. Syst.*, IEEE, 2018.