GPLight: Grouped Multi-agent Reinforcement Learning for Large-scale Traffic Signal Control

Yilin Liu[1,2], **Guiyang Luo**[1*], **Quan Yuan**[1,2], **Jinglin Li**[1], **Lei Jin**[3], **Bo Chen**[1] and **Rui Pan**[1]
[1]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]State Key Laboratory of Integrated Services Networks, Xidian University, Xian 710126, China
[3]School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China
{liuyilin10, luoguiyang, yuanquan, jlli, jinlei, Czb199871, panrui805}@bupt.edu.cn

Abstract

The use of Multi-agent reinforcement learning (MARL) methods in coordinating traffic lights (CTL) has become increasingly popular, treating each intersection as an agent. However, existing MARL approaches either treat each agent absolutely homogeneous, i.e., same network and parameter for each agent, or treat each agent completely heterogeneous, i.e., different networks and parameters for each agent. This creates a difficult balance between accuracy and complexity, especially in large-scale CTL. To address this challenge, we propose a grouped MARL method named GPLight. We first mine the similarity between agent environment considering both real-time traffic flow and static fine-grained road topology. Then we propose two loss functions to maintain a learnable and dynamic clustering, one that uses mutual information estimation for better stability, and the other that maximizes separability between groups. Finally, GPLight enforces the agents in a group share the same network and parameter. This approach reduces complexity by promoting cooperation within the same group of agents while reflecting differences between groups to ensure accuracy. To verify the effectiveness of our method, we conducted experiments on both synthetic and real-world datasets, with up to 1,089 intersections. Compared with state-of-the-art methods, our experiment results demonstrate the superiority of our proposed method, especially in large-scale CTL.

1 Introduction

In recent years, there has been an unprecedented trend in coordinating and controlling traffic lights. This trend has been shown to be effective in improving the efficiency and robustness of road networks [Jiang *et al.*, 2021]. With the development of AI technology and the availability of large-volume traffic data, learning-based control approaches have shown great potential in solving traffic signal control (TSC) problems. In particular, multi-agent reinforcement learning (MARL) has shown great potential as a promising solution [Luo *et al.*, 2020], as it enables coordinated control and global optimization of large-scale traffic lights without the need for manual intervention.

In the past few years, there are many representative research achievements in TSC. Wei's team has come up with a lot of work worthy of reference. IntelliLight [Wei *et al.*, 2018] highlights the importance of features and emphasizes that different agents in the same environment should give greater weight to important features when making decisions. To avoid the need for heuristic design of reinforcement learning (RL) parameters, PressLight [Wei *et al.*, 2019a] maps the rewards in the multi-agent system directly to the pressure values defined in traffic. FRAP [Zheng *et al.*, 2019] recognizes the spatial symmetry of the same agent model at different times and improves the generalization ability of the model. CoLight [Wei *et al.*, 2019b] introduces the graph attention mechanism (GAT [Velickovic *et al.*, 2017]) in the multi-agent system, considering the interaction of the surrounding agents for the first time. In addition, in the latest study, EMVLight [Su *et al.*, 2022] provides a good solution for emergency vehicles through TSC. However, none of the above methods have been applied in large-scale scenarios. MPLight [Chen *et al.*, 2020] and OAM [Liang *et al.*, 2022] are applied to large-scale traffic scenarios, but the heterogeneity of each intersection as well as the influence relationship between intersections are not considered.

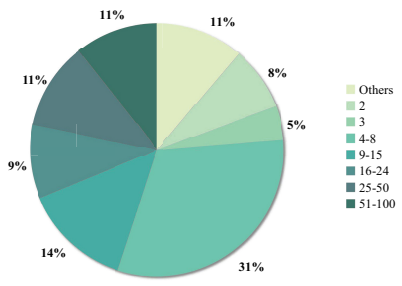

Figure 1: Number of intersections in previous studies.

*Corresponding author.

We investigated the number of intersections discussed in the TSC field, which is shown in Figure 1. The results showed in that 96% of the studies were conducted in the scenario with less than 100 intersections. Therefore, large-scale TSC is still an immature field requiring special attention. Coordinating large-scale traffic lights using MARL is a practical requirement, but it is a challenging problem to solve. On the one hand, it is difficult for all agents to use only a single network representation and strategy [Smith, 1937] since 1) large-scale road intersections have complex and abundant patterns, which requires an extremely large and deep neural network. On the other hand, if each agent applies a different neural network, there would be an extremely large number of parameters, which has exceedingly low efficiency and high complexity. Furthermore, intersections that are far apart could have strong spatial-temporal correlation and dynamics, which increase the connection complexity between agents.

To this end, we introduce GPLight, which extracts spatial-temporal features from intersections in real-time and clusters them into different groups. Agents in each group share the same neural network. We first consider real-time traffic flow and static fine-grained road topologies to dynamically divide intersections into different groups by mining the similarity between environments. Grouping agents can reduce the scale of the multi-agent system and improve training efficiency. We use GCN [Kipf and Welling, 2016] network to extract features, and then introduce two loss functions to carry out fine-grained partitioning of multi-agent systems. MI Loss aims to keep agents change smoothly, while Gather Loss aim to let agents find their partners in the same group. The grouping results will be transmitted into the improved QMIX [Rashid *et al.*, 2018] network for training. By mining the similarities among intersections in large-scale intersections, intersections with high similarity can be grouped together for cooperation, even if they are far away, thereby reducing the complexity of the system. At the same time, differences between different groups are preserved, allowing GPLight to strike a balance between accuracy and complexity. This process breaks through the traditional TSC method, which only considers the mutual influence of adjacent intersections.

We evaluated GPLight using both synthetic and real-world datasets with up to 1,000 intersections. The experimental results demonstrate that our proposed multi-agent grouping approach, which incorporates dynamic features, enables agents to share policies more effectively and dynamically. The traffic system scheduled by GPLight achieves better efficiency in coordinating large-scale traffic lights compared to existing methods.

The contributions of this work is threefold:

- We comprehensively extract both dynamic and static features of each agent to create its embedding. By considering the real-time dynamic traffic flow and the real road topology, the similarity between intersections can be better mined.
- We propose a MARL algorithm for large-scale traffic light intersections, which divides the multi-agent system into different groups to reduce complexity. Different from the traditional method which only considers the interaction between adjacent intersections, it allows intersections that show similarities in the whole region to cooperate even though they are far apart.
- We conduct extensive experiments on large-scale multiple scenarios with up to 1,089 intersections, including both synthetic and real-world datasets. Experimental results show that GPLight can achieve better performance in terms of total travel time than other advanced TSC methods in large-scale scenarios.

2 Related Work

Intelligent transportation is an important direction of future urban construction [Luo *et al.*, 2022a; Luo *et al.*, 2022b; Luo *et al.*, 2023]. TSC has been studied in the field of transportation for many years. In recent years, there has been a growing interest in combining TSC with MARL. In this approach, the road network is treated as the observation and the signal phase combination as the action set. The signal phase is defined as a set of permissible traffic movements [Zheng *et al.*, 2019]. For example, at an intersection shown in Figure 2, there are eight signal phases combinations to choose from.

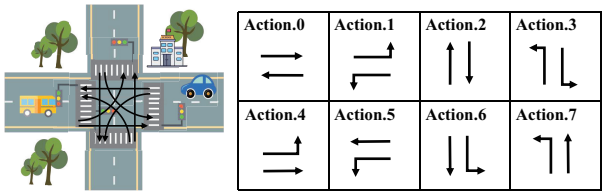

Figure 2: Signal phase and corresponding action set of crossroads.

In the past, there are many well-known studies of TSC combined with MARL. Intellilight [Wei *et al.*, 2018] uses deep Q-Network (DQN), which states the queue length of each lane, the total number of vehicles at the intersection, the updated wait time, the image of each vehicle's position at the intersection, the current agent's action and the next action. Presslight [Wei *et al.*, 2019a] proposes the use of max pressure as an input feature and reward to maximize throughput in the traffic network. A study in 2020 [Jamil *et al.*, 2020] proposed to integrate rewards obtained by different methods into the training process. Each reward has an independent network to learn Q value, and then votes to get the final action and interact with the environment. GeneraLight [Zhang *et al.*, 2020] designed a traffic flow generator based on Wasserstein generation adversal network, which improves the adaptability of MARL model to dynamic traffic flow and enhances the generalization ability of MARL model. FedLight [Ye *et al.*, 2021] considers collaborative optimization between intersections and proposes a combination of federated learning and RL. Jiang et al. [Jiang *et al.*, 2021] used multi-time scale model training to learn appropriate strategies for optimal control of traffic signals and dynamic lanes. MACAR [Yu *et al.*, 2021] realizes active communication between agents by considering the effect of the synchronization of agents. It consists of an active communication agent network (CAN) involving a message propagation graph neural network (MPGNN) and

a traffic prediction network (TFN). By using predictive information, action value bias during the training process is mitigated to help correct the agent's future actions.

To sum up, as a multi-agent problem, each intersection can be considered as an agent in TSC. Since there are thousands of intersections in a city, solving this problem is extremely challenging. While creating a model for each agent would ensure accuracy, it would require significant resources and difficult training. Therefore, we propose grouping the multi-agent system to identify similar intersections in the entire area for collaboration, thereby reducing the complexity of the training process.

3 Preliminary

GPLight treats TSC as a multi-agent systems, by grouping agents into clusters for achieving better accuracy-complexity tradeoff in coordinating large-scale traffic lights. This section introduces the MARL in TSC.

GPLight considers TSC as a multi-agent task which can be modelled by a distributed multi-agent partially observable Markov decision process (Dec-POMDP) [Oliehoek and Amato, 2016] $L = \langle N, S, A, P, \mathbb{O}, O, R, n, \gamma \rangle$, where $N = \{ 1,2,\cdots, n \}$ is the set of n agents. S is a finite set of states. For each agent $i \in N$, they can observe only partial environment o_t^i at each time step t, where o_t^i is part of the state $s_t \in S$. In our scenario, o_t^i includes both the real time traffic flow as well as road network topology. A is the set of joint actions, where the action a_t^i of each agent i at time step t involves the signal phase combinations that can be selected at the current intersection. Take Figure 2 as an example, for a four-way intersection, there is a total of eight actions that can be selected [Chen *et al.*, 2020]. For an agent i, it will choose an appropriate action $a_t^i \in A$ based on the observation o_t^i at time t. The agent will keep this action until the next decision is made. P is the transition probability function. γ is the discount factor whose value space is [0,1). Each intersection is controlled by an RL agent. We consider the system partially observable, which means agent i can only derive an observation $o_i \in \mathbb{O}$ from the observation $O(s,i)$. Given the traffic situation and current traffic signal phase, the goal of the agent is to take an optimal action $a \in A$ to maximize the cumulative reward R at each time step t. Our goal is to expect the overall traffic situation to become more unimpeded. Therefore, we consider combining each agent's reward with the length of the queue at the intersection. The reward R of each agent at time t is obtained by the reward function $S \times A_1 \times ... \times A_n \to \mathbb{R}$. Here, for a certain intersection i, we assume that $z_t^{i,l}$ is the queue length of vehicles in the approaching lane l at time t. We define its reward as R_t^i=$-\sum_l z_t^{i,l}$. Each agent has history phases τ_i. The joint strategy π generates a joint action-value function:

$$Q_{tot}^{\pi}(s,a) = \mathbb{E}_{s_{0:\infty},a_{0:\infty}}[\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a, \pi]. \tag{1}$$

4 The proposed model: GPLight

In this section, we propose a grouped multi-agent reinforcement learning method GPLight that can extract static and dynamic features of intersections, then group them for efficient training. It uses mutual information to divide the whole multi-agent system into different agent groups so as to enhance the efficiency of shared learning among agents. As shown in Figure 3, our model mainly includes three parts: Feature Extraction, Group Cohesion and Q-Learning.

GPLight first mines the similarity of agents considering both real-time traffic flow and static road topology, and then maintains a learnable and dynamic clustering to group agents. Since the road topology information is non-euclidean data, we apply the GCN network to extract features for each intersection. As previous research [Kipf and Welling, 2016] suggests that GCN embedding (even with random weights) can automatically clustering when extracting features, the feature extraction process enables a coarse-grained clustering. However, such a clustering is not meticulously designed to guarantee best performance for subsequent MARL tasks. To address this issue, we propose two loss functions, namely Mutual Information (MI) Loss and Gather Loss, to supervise the GCN network for a fine-grained clustering. While MI Loss is used to ensure steady changes of the agents, Gather Loss is applied to maximize the separability between clusters. Furthermore, the clustering is also supervised by the task (MARL Loss), i.e., traffic lights coordination performance. These three losses act as supervisory signals to guide the GCN network to extract ample features, which ultimately lead to the best grouping result for the MARL task. Finally, agents in the same group will share the same network parameters to make decisions.

4.1 Feature Extraction

We model the traffic network in the multi-agent scenario as a graph $G = (V,E)$, where V is the set intersections and E means the road connections in between. Each intersection is treated as an agent. Each agent $i \in V$ has a partial observation o_t^i at time step t. o_t^i includes 1) static features, such as the number of lanes, length, speed limits, type of roads, as well as the local road topology; and 2) dynamic features, such as the real time traffic flow as well as the current signal phase. We concatenates the state and dynamic features as vector $x_i \in \mathbb{R}^M$, which represents the partial observation o_t^i, where M is the feature dimension. All nodes' features can be represented by a matrix $X_{n \times M}$, where n represents the number of nodes. The input of GCN at each layer is the adjacency matrix Z and node feature H, where $H_0 = X$. The final layer feature propagation formula improved by GCN is as follows,

$$f(H^{(l+1)}, Z) = \sigma\Big(\tilde{C}^{-\frac{1}{2}} \tilde{Z} \tilde{C}^{-\frac{1}{2}} H^{(l)} W^{(l)}\Big), \tag{2}$$

where $\tilde{C}$ is a matrix introduced to normalize Z.

For the feature extraction model, we construct a GCN network with several layers, and the activation function adopts ReLU and Softmax respectively, so the overall forward propagation formula is as follows:

$$f(X, Z) = softmax\Big(\hat{Z}\,\mathrm{ReLU}\big(\hat{Z} X W^{(0)}\big) W^{(1)}\Big). \tag{3}$$

Through the above, static and dynamic features are extracted and processed by the GCN embedding layer. The re-

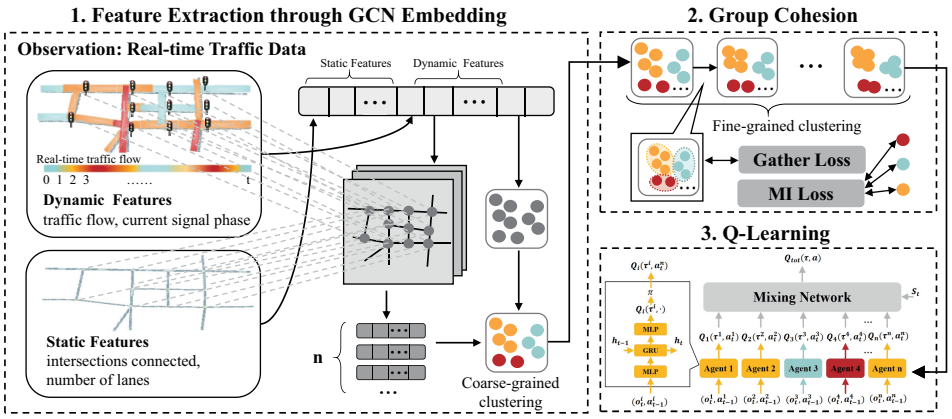

Figure 3: GPLight general framework includes Feature Extraction, Group Cohesion and Q-Learning.

sult obtained by the embedding will be input to the next part for grouping cohesion.

4.2 Group Cohesion

In this section, we introduce two loss functions for better group cohesion.

MI Loss : stability maintenance

Given the features extracted by GCN, each agent i has an embedding ρ_i. In order to adapt to the dynamic environment and avoid rapid change that leads to learning instability, we propose to ensure slow changes of agents by maximizing $I(\tau_i;\rho_i \mid o_i)$, the conditional mutual information between the individual actions and the group given the current observation. However, estimating and maximizing mutual information can often be challenging. To address this, we refer to the work of ROMA[Lhaksmana *et al.*, 2018], which introduced a variational posterior estimate to derive a tractable lower bound for mutual information targets [Wainwright and Jordan, 2008; Alemi *et al.*, 2016]:

$$I(\rho_i^t;\tau_i^{t-1}|o_i^t) \geq \mathbb{E}_{\rho_i^t,\tau_i^{t-1},o_i^t}\left[log\frac{q_\xi(\rho_i^t|\tau_i^{t-1},o_i^t)}{p(\rho_i^t|o_i^t)}\right], \quad (4)$$

where $\tau_i^{t-1} = (o_i^0, a_i^0, \cdots, o_i^{t-1}, a_i^{t-1})$. q_ξ is the variational estimator parameterised with ξ , and we call it an encoder, which uses a GRU [Cho *et al.*, 2014] to encode the history of the agent's observation and behavior. The first loss function we use is rewritten from the lower bound of Eq. 4 as follows:

$$\mathcal{L}_I(\theta_\rho,\xi) = \mathbb{E}_{(\tau_i^{t-1},o_i^t)\sim B}\Big[B_{KL}\left[p(\rho_i^t|o_i^t) \parallel q_\xi(\rho_i^t|\tau_i^{t-1},o_i^t)\right]\Big], \quad (5)$$

where parameters θ_ρ are conditioned on ρ_i, B is a replay buffer and $B_{KL}[\cdot\|\cdot]$ is the KL divergence operator.

Gather Loss : separate different groups

Furthermore, we have to separate different groups in order to 1) make agents in the same multi- agent group have more similar features to ensure the accuracy of the shared decision, and 2) maximize differentiation of different multi-agent groups to ensure that the grouping is reasonable. To achieve distinguishable groups, we have the following formula to minimize the similarity between agent i and agent j [Lhaksmana *et al.*, 2018]:

$$\begin{aligned} &\underset{\theta_\rho,\xi,\phi}{minimize}\, U^t_{\phi_{2,0}} \\ &subject\ to\ I\left(\rho_i^t,\tau_j^{t-1}|o_j^t\right) + u_\phi\left(\tau_i^{t-1},\tau_j^{t-1}\right) > 1, \forall i \neq j, \end{aligned} \quad (6)$$

where matrix $U_\phi = (u_{ij})$, $u_{ij} = u_\phi(\tau_i,\tau_j)$ is used to measure the difference in the distribution of agent i and agent j by historical local states comparison. The meaning of subscript (i.e., 2,0) is the Frobenius norm. $I(\rho_i;\tau_j)$ represents the mutual information between agent i and agent j. The values of u and I are both in $[0,1]$. We want to minimize the non-zero elements in matrix U while maximizing the sum of I and u. The purpose of this is that we expect to maximize I preferentially, that is, to enhance compactness within multi-agent groups. In this way, the value of u will be high only when the mutual information I of the two agents is low, which means their difference is large. Thus, the multi-agent group becomes compact and the distinction between groups becomes more obvious.

Similarly, we construct an upper bound as the second loss function we will use:

$$\begin{aligned} \mathcal{L}_U(\theta_\rho,\phi,\xi) = \mathbb{E}_{(\tau^{t-1},o^t)\sim B,\rho^t\sim p(\rho^t|o^t)} \quad & U^t_{\phi_F} \\ - \sum_{i\neq j} \min\{q_\xi(\rho_i^t|\tau_j^{t-1},o_j^t) \quad & \\ + u_\phi(\tau_i^{t-1},\tau_j^{t-1}),1\}, & \end{aligned} \quad (7)$$

where F represents Frobenius norm, τ^{t-1} joint distribution and o^t is the joint observation.

4.3 Q-Learning

QMIX [Rashid *et al.*, 2018] is a multi-agent reinforcement learning algorithm, which is suitable for Dec-POMDP

[Oliehoek and Amato, 2016]. QMIX is characterized by centralized training and distributed execution application framework.

Through group cohesion in the previous step, it is possible for us to combine centralized training with multi-agent clustering. As shown in Figure 4, at time t, we input the clustering result of the multi-agent system into QMIX network. In this way, different types of multi-agents can be obtained, and the same multi-agents will share the same DRQN network. This step is obtained by averaging the parameters of each layer of DRQN. Therefore, the same group of multi-agents will have a high degree of similarity in decision-making. Instead, different types of multi-agents will be distinguished by differences in the network. At the same time, the group belonging to the same agent will constantly change at different times.

In order to make more use of the state information S_t of the system, S_t is mixed into Q_{tot} (the sum of q values of each agent by linear transformation) through the hyper network, rather than just as the input of the mixing network. In the process of updating, the idea of traditional Deep Q-Network is used, samples will be sampled from the replay buffer. In this way, we get the final loss function in the process of Q-learning as follows [Rashid *et al.*, 2018]:

$$\mathcal{L}_{TD}(\theta) = \Big[r + \gamma max_{a'} Q_{tot}(s', a'; \theta^-) - Q_{tot}(s, a; \theta)\Big]^2, \tag{8}$$

where θ^- are periodically updated parameters for the target network.

From what has been discussed above, the final learning goal of our GPLight framework consists of three Loss functions:

$$\mathcal{L} = \mathcal{L}_{TD} + \lambda_I \mathcal{L}_I + \lambda_U \mathcal{L}_U, \tag{9}$$

where λ_I and λ_U are scaling factors.

5 Experiments

5.1 Settings

We run our experiments on CityFlow [Zhang *et al.*, 2019], a traffic simulator. Compared to SUMO [Lopez *et al.*, 2018], CityFlow is a highly concurrent multi-threaded system with significantly faster simulation. In our experiments, each car has its own set of parameters, e.g., acceleration, maximum speed, which greatly improves the realism of the traffic simulation environment.

As each car makes its way from start position to destination. GPLight schedules the traffic light of all intersections, which would influence the moving speed of all vehicles since Vehicles follow the traffic rules. According to the traditional setting, each green signal is followed by three seconds of yellow light and two-second all red time.

5.2 Dataset

Synthetic Data. In the synthetic dataset, we will use two kinds of maps. They are made up of different number of intersections. Synthetic maps are generated via Cityflow and include road attributes such as the number of lanes and road speed limits. Each road at the intersection has three lanes with 3 meters in width, lanes between two intersections is different in length.

- $Grid_{10\times10}$-*Uni.* In this type of map, traffic flows in one direction. We uniformly set it to run west to east and north to south. The west→east traffic flow is 300 vehicles/lane/hour, and the north→south traffic flow is 90 vehicles/lane/hour.
- $Grid_{33\times33}$-*Bi.* In this type of map, there are 1,089 intersections, traffic flows in both directions. Vehicles moving in from east, west, north, south. The east↔west traffic flow is 300 vehicles/lane/hour, and the north↔south traffic flow is 90 vehicles/lane/hour. The structure of the road network is heterogeneous, thus the setting is more realistic.

Real-world Data. We also experiment with real traffic data. For the convenience of subsequent comparative experiments, we continue to use the real maps of Hangzhou, Jinan in China and NewYork in USA. Their road network structure can be imported from OpenStreetMap, as shown in Figure 4. A detailed comparison to the three real-world datasets are shown in Table 1.

Table 1: Comparison between Real-world Datasets

	$D_{Hangzhou}$	D_{Jinan}	$D_{NewYork}$
Intersections	16	12	196
Average arrival (vehicles/300s)	526.63	250.70	240.79
Primary roads	86	183	195
Secondary roads	117	164	306
Trunk links	26	33	27

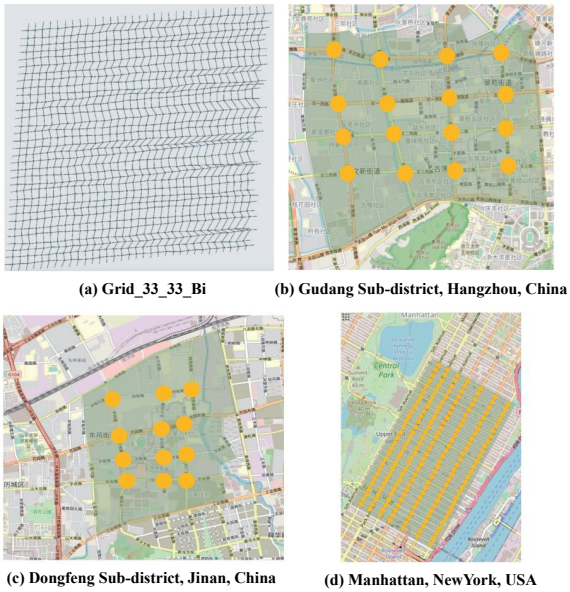

Figure 4: (a) is Synthetic Map with 1,089 intersections. (b)-(d) are Real-world Maps with 16, 12, 196 intersections. The green areas on the maps are the ones we use. The intersections within the yellow circles will be used in the experiment.

Table 2: Performance on synthetic data and real-world data.

Method	$Grid_{10\times10}$-Uni	$D_{Hangzhou}$	D_{Jinan}	$D_{NewYork}$
Fixedtime [Koonce and Rodegerdts, 2008]	345.81	718.29	814.09	2125.97
MaxPressure [Varaiya, 2013]	319.28	416.82	487.52	1826.78
IntelliLight [Wei *et al.*, 2018]	308.97	402.68	461.47	1952.11
CoLight [Wei *et al.*, 2019b]	286.04	356.88	355.41	1534.36
MPLight [Chen *et al.*, 2020]	305.65	348.12	417.51	1673.68
GPLight	**260.37**	**301.45**	**307.52**	**1284.98**

5.3 Baseline

Our experiment mainly compare with two types of methods, traditional traffic signal control methods as well as deep reinforcement learning based signal control methods. Here are the details:

- *Fixedtime* [Koonce and Rodegerdts, 2008]. In the method of Fixedtime, intersection traffic signals are in accordance with the pre-set timing scheme. Traffic signal light changes periodically.
- *MaxPressure* [Varaiya, 2013]. In MaxPressure, the purpose of traffic signal control is to minimize the pressure at the intersection and balance the length of vehicle queue on the lanes connected with the intersection.
- *IntelliLight* [Wei *et al.*, 2018]. It essentially uses a deep Q-learning network (DQN). The agent at each intersection is completely independent, regardless of adjacency and parameter sharing. The reward value is set as the weighted result of the six evaluation indexes.
- *CoLight* [Wei *et al.*, 2019b]. Graph neural network is introduced in CoLight. It takes into account the influence of surrounding intersections on the current intersection by introducing graph attention network and some multi-head calculations.
- *MPLight* [Chen *et al.*, 2020]. MPLight combined with MARL to conduct experiments at large-scale intersections. It sets up a DQN at each intersection.

5.4 Evaluation Metric

The main purpose of controlling the traffic signal lights at the intersection is to make the vehicles pass through the intersection more efficiently. In order to achieve this goal, we usually set some indicators to evaluate the efficiency [Wei *et al.*, 2019c]. In our experiment, we choose **Travel Time** to evaluate the performance of signal control algorithm. It is defined as the average time taken by all vehicles during their journey.

5.5 Effect verification of Group Cohesion

In this section, we visualize the results of Group Cohesion to demonstrate the feasibility of the method.

We process the data of Group Cohesion in GPLight, hoping to see its effect intuitively. It is worth noting that since we want to visualize the results, we compress the embedding of GCN to two dimensions.

We run the experiment on the synthetic map. Figure 5 shows the clustering results of Group Cohesion. It can be seen that GPLight can divide groups effectively and the results change in real time with different traffic conditions at different times. Figure 6 shows the results of Group Cohesion combined with the road network. Different from traditional TSC in which only the influence of adjacent intersections is considered, we can see from the figure that in GPLight, it is possible to cooperate even when intersections are far apart. This is because the topological structure and traffic features are likely to be highly similar even if intersections are far apart. GPLight processes the comprehensive features of the intersections to mine these similarities.

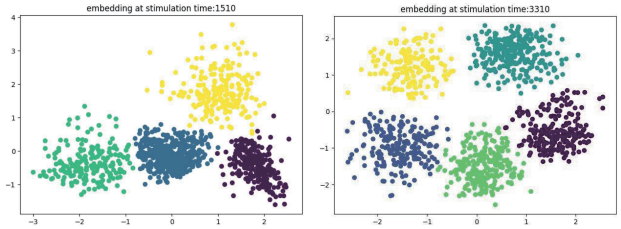

Figure 5: Visualization of GPLight clustering effect. It shows the distribution of intersection after GPLight feature extraction. At different times, the results of clustering change with the change of features.

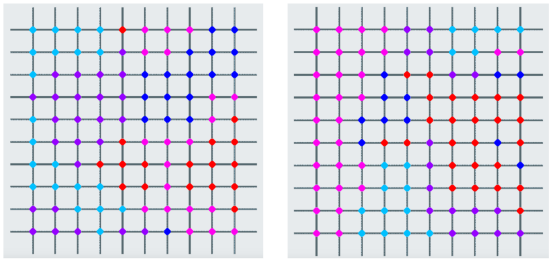

Figure 6: The intersections of the same color are in a group, which means that their extracted features have a high degree of similarity. Note that the grouping at each intersection changes in real time. Intersections far apart but similar can affect each other.

5.6 Performance Comparison

In this section, we show the performance of GPLight and compare it with conventional transportation methods and RL methods.

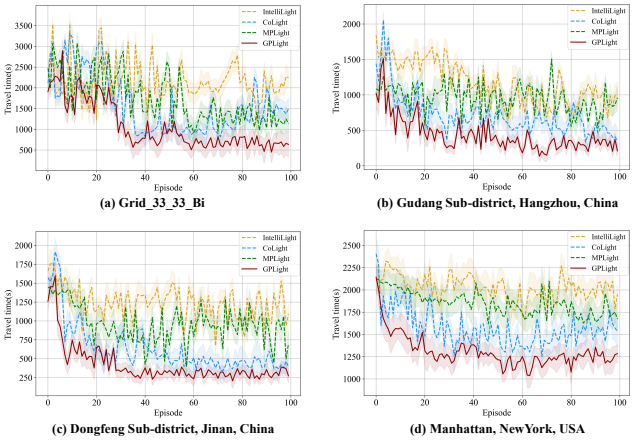

Figure 7: Convergence speed of IntelliLight, CoLight, MPLight and GPLight. Performance of GPLight is best.

Table 3: Performance on large scale.

Method	$Grid_{33\times33}$-Bi	Improvement
Fixedtime	2032.56	63.8%
MaxPressure	2171.13	66.1%
IntelliLight	2246.23	67.3%
CoLight	1493.72	50.8%
MPLight	1134.01	35.1%
GPLight	**735.51**	-

Overall Analysis. Table 2 shows GPLight's comparison to the other five approaches, including two traditional TSC approaches and three advanced TSC approaches in MARL. According to experimental result, GRLight has an average improvement of 21.6% compared with the two traditional methods (Fixedtime and MaxPressure) on the synthetic datasets. On real-world datasets, GPLight has improved 53.3% over Fixedtime and 31.4% improvement over MaxPressure on average. This is because conventional traffic light controls do not apply to traffic conditions that change over time. The traffic signal control methods which are adjusted according to the real-time traffic flow are more suitable for our life.

We also compare GPLight with three advanced MARL-based TSC methods (IntelliLight, CoLight and MPLight). As we can see from the table, the performance of GPLight is significantly better. GRLight achieves an average 13.2% improvement over the other three methods on synthetic datasets. In addition, GPLight averages 30.4% improvement over IntelliLight, 15.1% improvement over CoLight and 21% improvement over MPLight on real-world datasets, which proves its superior performance. We can see that the RL-based approaches are significantly superior to the traditional TSC approaches. This is because the TSC methods based on RL can flexibly make judgements derived from the current states of the intersections, which makes a great contribution to the changing traffic situation at every moment.

Furthermore, we conducted an experiment on a large and irregular road network. As shown in Table 3, GPLight demonstrates its superiority more prominently on $Grid_{33\times33}$-Bi maps than on the others, highlighting the effectiveness of our approach in large-scale TSC. The reason for this is that we have mined intersections with high similarity among large scale intersections and grouped them for cooperation. The experimental results demonstrate that this cooperation model significantly improves efficiency.

In conclusion, GPLight groups multiple agents, which not only ensures the diversity between different groups, but also reduces the difficulty of training within the same group. The experimental results prove that GPLight can effectively group multi-agent systems and achieve superior performance, which is particularly evident in large-scale TSC.

Convergence Analysis. In Figure 7, we compare GPLight with IntelliLight, CoLight and MPLight's convergence rate during training. The metric used is the average travel time of vehicles evaluated at each episode. CoLight's convergence trend is similar to GPLight's, but GPLight performs best compared to the other three advanced RL-based TSC approaches. This is reflected in three aspects, respectively, initial performance after the first episode, learning time to achieve a pre-expected goal, and the final learning result. From this, we can conclude that our model GPLight learned the best way to make decisions and achieved good results in overall average travel time while maintaining excellent convergence rates.

5.7 Ablation Experiments

In the above experiments, we can see that GPLight has shown excellent results. As GPLight is mainly composed of three modules: GCN Embedding, Group Cohesion and QMIX. To prove the effectiveness of each module, we conduct ablation experiments. It can be seen from Table 4 that multi-agent reinforcement learning effects without GCN embedding or Group cohesion will become worse, which proves that both GCN Embedding and Group Cohesion have played indispensable roles in GPLight.

Table 4: Ablation experiments of GPLight

Method	$Grid_{33\times33}$-Bi
QMIX	1225.63
QMIX+Group cohesion	862.24
QMIX+GCN embedding	1013.78
QMIX+Group cohesion+GCN embedding	**735.51**

6 Conclusion

This paper proposes a MARL traffic signal control method named GPLight, which balances accuracy and complexity in large-scale TSC by grouping agents with a high degree of similarity. In the future, we will focus on following aspects: 1) heterogeneous intersections; 2) MARL algorithms for grouped agents. As agents are divided into groups, agents within a group collaborate for higher training efficiency and lower complexity, and agents belong to different groups cooperate for better traffic efficiency. Taking these interactions into consideration, we will propose novel MARL methods to improve the training efficiency and stability.

Acknowledgments

This paper is supported in part by the the National Key Research and Development Program of China under Grant 2022YFB4300403, the Natural Science Foundation of China under Grant 62102041, Grant 62272053, and in part by the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant 2022QNRC001.

References

[Alemi *et al.*, 2016] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

[Chen *et al.*, 2020] Chacha Chen, Hua Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yuanhao Xiong, Kai Xu, and Zhenhui Li. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3414–3421, 2020.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[Jamil *et al.*, 2020] Abu Rafe Md Jamil, Kishan Kumar Ganguly, and Naushin Nower. Adaptive traffic signal control system using composite reward architecture based deep reinforcement learning. *IET Intelligent Transport Systems*, 14(14):2030–2041, 2020.

[Jiang *et al.*, 2021] Qize Jiang, Jingze Li, Weiwei Sun SUN, and Baihua Zheng. Dynamic lane traffic signal control with group attention and multi-timescale reinforcement learning. IJCAI, 2021.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Koonce and Rodegerdts, 2008] Peter Koonce and Lee Rodegerdts. Traffic signal timing manual. Technical report, United States. Federal Highway Administration, 2008.

[Lhaksmana *et al.*, 2018] Kemas M Lhaksmana, Yohei Murakami, and Toru Ishida. Role-based modeling for designing agent behavior in self-organizing multi-agent systems. *International Journal of Software Engineering and Knowledge Engineering*, 28(01):79–96, 2018.

[Liang *et al.*, 2022] Enming Liang, Zicheng Su, Chilin Fang, and Renxin Zhong. Oam: An option-action reinforcement learning framework for universal multi-intersection control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4550–4558, 2022.

[Lopez *et al.*, 2018] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2575–2582. IEEE, 2018.

[Luo *et al.*, 2020] Guiyang Luo, Hui Zhang, Haibo He, Jinglin Li, and Fei-Yue Wang. Multiagent adversarial collaborative learning via mean-field theory. *IEEE Transactions on Cybernetics*, 51(10):4994–5007, 2020.

[Luo *et al.*, 2022a] Guiyang Luo, Hui Zhang, Xiao Wang, Quan Yuan, Jinglin Li, and Fei-Yue Wang. Acp based large-scale coordinated route planning: From perspective of cyber-physical-social systems. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1842–1849. IEEE, 2022.

[Luo *et al.*, 2022b] Guiyang Luo, Hui Zhang, Quan Yuan, Jinglin Li, and Fei-Yue Wang. Estnet: embedded spatial-temporal network for modeling traffic flow dynamics. *IEEE transactions on intelligent transportation systems*, 23(10):19201–19212, 2022.

[Luo *et al.*, 2023] Guiyang Luo, Hui Zhang, Quan Yuan, Jinglin Li, Wendong Wang, and Fei-Yue Wang. One size fits all: A unified traffic predictor for capturing the essential spatial–temporal dependency. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2023.

[Oliehoek and Amato, 2016] Frans A Oliehoek and Christopher Amato. *A concise introduction to decentralized POMDPs*. Springer, 2016.

[Rashid *et al.*, 2018] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 4295–4304. PMLR, 2018.

[Smith, 1937] Adam Smith. *The wealth of nations [1776]*, volume 11937. na, 1937.

[Su *et al.*, 2022] Haoran Su, Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. Emvlight: A decentralized reinforcement learning framework for efficient passage of emergency vehicles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4593–4601, 2022.

[Varaiya, 2013] Pravin Varaiya. The max-pressure controller for arbitrary networks of signalized intersections. In *Advances in dynamic network modeling in complex transportation systems*, pages 27–66. Springer, 2013.

[Velickovic *et al.*, 2017] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *stat*, 1050:20, 2017.

[Wainwright and Jordan, 2008] Martin J Wainwright and Michael Irwin Jordan. *Graphical models, exponential families, and variational inference*. Now Publishers Inc, 2008.

[Wei *et al.*, 2018] Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2496–2505, 2018.

[Wei *et al.*, 2019a] Hua Wei, Chacha Chen, Guanjie Zheng, Kan Wu, Vikash Gayah, Kai Xu, and Zhenhui Li. Presslight: Learning max pressure control to coordinate traffic signals in arterial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1290–1298, 2019.

[Wei *et al.*, 2019b] Hua Wei, Nan Xu, Huichu Zhang, Guanjie Zheng, Xinshi Zang, Chacha Chen, Weinan Zhang, Yanmin Zhu, Kai Xu, and Zhenhui Li. Colight: Learning network-level cooperation for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1913–1922, 2019.

[Wei *et al.*, 2019c] Hua Wei, Guanjie Zheng, Vikash Gayah, and Zhenhui Li. A survey on traffic signal control methods. *arXiv preprint arXiv:1904.08117*, 2019.

[Ye *et al.*, 2021] Yutong Ye, Wupan Zhao, Tongquan Wei, Shiyan Hu, and Mingsong Chen. Fedlight: Federated reinforcement learning for autonomous multi-intersection traffic signal control. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 847–852. IEEE, 2021.

[Yu *et al.*, 2021] Zhengxu Yu, Shuxian Liang, Long Wei, Zhongming Jin, Jianqiang Huang, Deng Cai, Xiaofei He, and Xian-Sheng Hua. Macar: Urban traffic light control via active multi-agent communication and action rectification. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2491–2497, 2021.

[Zhang *et al.*, 2019] Huichu Zhang, Siyuan Feng, Chang Liu, Yaoyao Ding, Yichen Zhu, Zihan Zhou, Weinan Zhang, Yong Yu, Haiming Jin, and Zhenhui Li. Cityflow: A multi-agent reinforcement learning environment for large scale city traffic scenario. In *The world wide web conference*, pages 3620–3624, 2019.

[Zhang *et al.*, 2020] Huichu Zhang, Chang Liu, Weinan Zhang, Guanjie Zheng, and Yong Yu. Generalight: Improving environment generalization of traffic signal control via meta reinforcement learning. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1783–1792, 2020.

[Zheng *et al.*, 2019] Guanjie Zheng, Yuanhao Xiong, Xinshi Zang, Jie Feng, Hua Wei, Huichu Zhang, Yong Li, Kai Xu, and Zhenhui Li. Learning phase competition for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1963–1972, 2019.