Spatio-Temporal Sequence Modeling for Traffic Signal Control

Qian Sun
The Division of Emerging Interdisciplinary Areas, The Hong Kong University of Science and Technology
Hong Kong SAR, China
qsunal@connect.ust.hk

Le Zhang*
Jingbo Zhou
Baidu Research, Baidu Inc.
Beijing, China
zhangle0202@gmail.com
zhoujingbo@baidu.com

Rui Zha
School of Computer Science, University of Science and Technology of China
Hefei, China
crui0210@gmail.com

Yu Mei
Chujie Tian
Department of Intelligent Transportation System, Baidu Inc.
Beijing, China
whqyqy@hotmail.com
tianchujie@baidu.com

Hui Xiong*
Thrust of Artificial Intelligence, The Hong Kong University of Science and Technology (Guangzhou), China
Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
Hong Kong SAR, China
xionghui@ust.hk

Abstract

Traffic Signal Control(TSC), a pivotal and challenging research area in the transportation domain, aims to alleviate congestion at urban intersections by optimizing vehicular flows from different inflow directions. While large efforts have been focused on using Reinforcement Learning(RL) based methods to tackle the TSC problem, it possesses constraints such as unpredictable training duration and risks of online exploration, limiting its real-world deployment. Recently, offline RL has emerged as a new solution by transitioning from learning through online interactions to deriving policies from pre-collected datasets, which guarantees a safer and more efficient learning process. However, existing offline methods overlook the crucial temporal and spatial intricacy among data from different traffic signals at different timesteps, which leads to suboptimal performance. To this end, in this paper, we present an innovative formulation of the offline TSC problem by introducing a spatio-temporal graph to model the historical Markov Decision Process sequences across all traffic signals within the road network. Along this line, we propose STLight, a novel spatio-temporal sequence modeling approach to predict optimal actions for the signals from historical data, accounting for the inherent inter-dependencies among them. Specifically, we incorporate a spatio-temporal encoder to represent states, actions, and returns by capturing dynamic and spatially dependent information. The ordered space-time-aware representations are further fed to the Action Decoder to predict signal phase actions in an auto-regressive manner, accounting for the hidden dependencies between the actions and the reward and state tokens. Furthermore, to adaptively handle tasks with different levels of congestion scenarios, we incorporate space-aware return-based contrastive learning to automatically differentiate data samples with disparate traffic flow patterns. Finally, extensive experiments conducted on two public real-world traffic datasets clearly demonstrate the superior performance of the proposed model over both the state-of-the-art online and offline traffic signal control baselines.

CCS Concepts

• **Applied computing** → **Transportation**.

Keywords

Reinforcement Learning; Traffic Signal Control

ACM Reference Format:
Qian Sun, Le Zhang, Jingbo Zhou, Rui Zha, Yu Mei, Chujie Tian, and Hui Xiong. 2024. Spatio-Temporal Sequence Modeling for Traffic Signal Control. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679998

1 Introduction

The rapid urban development, along with the surge in the number of vehicles, has caused severe traffic congestion in numerous cities worldwide. This problem results in increased travel times for commuters and, more importantly, negatively impacts the environment. [1, 10, 15]. Controlling the traffic signals at intersections, known as Traffic Signal Control (TSC), is a vital approach to addressing this issue [8, 13]. With the advancements in machine learning, numerous works have studied Reinforcement Learning (RL) methods to model the TSC task, and have outperformed traditional TSC methods by modeling the complicated spatio-temporal relationships among the traffic signals [13, 15, 17]. Nevertheless,

*Corresponding authors.

CIKM '24, October 21–25, 2024, Boise, ID, USA

ACM ISBN 979-8-4007-0436-9/24/10
https://doi.org/10.1145/3627673.3679998

these RL methods necessitate learning through extensive online explorations. In the early training stage, the model often exhibits poor performance due to severe congestion caused by the trail-and-errors, constraining the practical deployment of these models in the real world. To address these challenges, recent works have started to model the TSC task using offline RL [18, 21]. They propose to directly learn the policy from fixed datasets leveraging approaches like Conservative Q-learning or sequence modeling, bypassing any online interactions. Although these offline models also achieve competitive performance, they still encounter several limitations. Primarily, they ignore the dynamic spatial dependencies among the data samples from different intersections, a crucial feature that online models have adeptly recognized and extensively validated [13, 15]. Additionally, they fail to adaptively discern different levels of TSC tasks according to their inherent congestion patterns represented by the traffic flows.

Due to the deficiencies of existing methods as mentioned above, we propose a spatio-temporal sequence modeling approach for TSC. In particular, we formulate the task as a sequence of Markov Decision Processes(MDP) characterized by states, actions, and returns (accumulated rewards) from individual traffic signals. By incorporating the road network topological structure, we propose our model, STLight, to predict optimal actions for signals from both the spatial and temporal perspectives. Specifically, we first design a spatio-temporal encoder to comprehensively represent states, actions, and returns by considering dynamic and spatially related information. The space-time aware representations are further transformed into ordered sequences to predict signal phase actions through the Action Decoder auto-regressively. Furthermore, to adaptively handle tasks with different types of congestion scenarios, we employ return-driven contrastive learning to automatically differentiate data samples with disparate traffic flow patterns. Extensive experiments conducted on two public real-world datasets demonstrate the superior performance of our model over the state-of-the-art traffic signal control baselines.

2 Preliminary

Sequence modeling for decision making aims to predict the actions based on historical MDP trajectories in an auto-regressive manner [3]. To prepare the offline data for sequence modeling, the MDP tokens including states s, actions a, and rewards r are structured into trajectories formulated as $\tau = (R_1, s_1, a_1, \ldots, R_K, s_K, a_K)$, where $R_t = \sum_{t'=t}^{K} r_{t'}$ represents the cumulative reward from current timestep t to the trajectory end K, aiming to guide the auto-regressive action prediction by target returns. Driven by the efficacy of Transformer in sequence modeling for decision making [7, 24, 25] as well as inter-signal collaborative decision making for RL-based TSC [13, 15], we structure the TSC trajectories into sequences τ consisting of global constructs. Specifically, the return tokens $R \in \mathbb{R}^{K\times N\times L}$ are the negative of total queue length on each incoming lane $l \in \{1, 2, \ldots L\}$ [2, 14], accumulated to the end of each sequence of length K, where N is the total number of traffic signals within the road network. The state tokens $S \in \mathbb{R}^{K\times N\times D}$, where D is the feature space dimension, are observations consisting of current phase encoding, total number of running and waiting vehicles in the incoming and outgoing lanes within the effective detection range [22], and the pressure between upstream lanes and downstream lanes [2]. The action tokens $A \in \mathbb{R}^{K\times N\times 1}$ are the indices of the possible green phases, where each phase controls a particular set of traffic movements dictating directional flows, such as going straight, turning left, or turning right. Hence, the offline TSC problem can be formulated as follows: given the offline trajectories τ as well as the road network graph $G \in \mathbb{R}^{N\times N}$, we train the model to fit the phase actions auto-regressively. During model evaluation, we enforce the model to predict the best phase actions driven by the maximum possible target return.

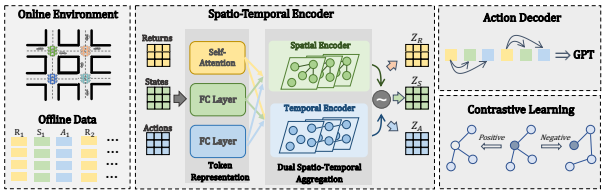


Figure 1: An overview of the proposed STLight framework.

3 Methodology

The proposed STLight framework is illustrated in Figure 1, comprised of three main components: 1) the Spatio-Temporal Encoder derives the spatially and temporally enhanced representations for states, actions, and returns; 2) the Autoregressive Action Decoder predicts actions in a causal manner; 3) Return-driven Contrastive Learning enhances the model's capability in discriminating trajectories samples into specific auxiliary tasks, each reflecting a unique congestion pattern. We provide in-depth explanations below.

3.1 Spatio-Temporal Encoder

The input sequences are first processed through a spatio-temporal encoder to obtain both temporally and spatially aggregated representations. Specifically, the encoder is comprised of a token representation module to encode the input tokens, followed by a dual spatio-temporal aggregation module to capture the dynamic and inter-signal dependencies within the embeddings.

3.1.1 Token Representation. Given the diverse dimensionalities of the heterogeneous input tokens, we first attempt to map these tokens into a unified representation space. Specifically, we employ fully connected layers $f_s(\cdot)$ and $f_a(\cdot)$ on the states and actions, respectively, resulting in representations $H_S \in \mathbb{R}^{K\times N\times d}$ and $H_A \in \mathbb{R}^{K\times N\times d}$, respectively, where d is the embedding dimension. For the returns, we claim that existing methods [18, 21] that directly apply vanilla neural networks on returns for each intersection overlook the intrinsic correlation among different lanes. To provide a more effective and comprehensive return representation, we introduce a lane-level self-attention mechanism. To be concrete, multiple lanes within a specific interaction are fed as basic tokens into the self-attention module, which is explicitly defined as follows:

$$\hat{R}^{i,j,k} = \text{MultiHeadAtt}\left(R^{i,j,k}W_Q, R^{i,j,k}W_K, R^{i,j,k}W_V\right), \quad (1)$$

where the input $R^{i,j,k}$ denotes the traffic feature for the k-th lane of the j-th intersection at the i-th timestep. The matrices $W_{Q,K,V} \in \mathbb{R}^{1\times d}$ are learnable parameters. On such basis, we further aggregate

the lane-level representations to obtain the return token embeddings by applying a pooling operation over all lanes within a single intersection, following: $H_R^{i,j} \leftarrow \text{Pooling}\left(\hat{R}^{i,j,1}, \ldots, \hat{R}^{i,j,L}\right)$.

3.1.2 Dual Spatio-Temporal Aggregation. With the token embeddings mapped into the unified embedding space, we further model them in both spatial and temporal dimensions, which serve as two pivotal aspects of our representation strategy. For the equations below, we use $\mathbf{M} \in \{\mathbf{R},\mathbf{S},\mathbf{A}\}$ to denote the modality.

Spatial Encoder. We first develop a space-centric encoder to capture the correlations among tokens across different traffic signals. Typical spatial message passing strategies such as Graph Convolution Networks [23] and Graph Attention Networks [12] on the established road network falls short in accurately representing the inter-signal correlations since they merely leverage the given node adjacency graph. Therefore, as inspired by [4, 19], besides applying graph message passing on the input embeddings H_M with the graph adjacency G, we parallely apply a transformer-like architecture, which employs a learnable position encoding P_M^S initialized with the adjacency G. Further, a linear mapping function is applied on the token embedding combined with the position encoding, i.e., $H_M^S = f_S\left(H_M \parallel P_M^S\right)$, where f_S is the linear mapping function, and $\parallel$ denotes the concatenation operation. Then, the space-aware representations can be obtained by applying multi-head attention on H_M^S with residual connections, where the multi-head attention employs the similar operation as Equation 1. The representation from the attention mechanism $\hat{H}_M^S$ is further combined with the representation from the direct graph message passing $\tilde{H}_M^S$ to obtain the spatially enhanced representation Z_M^S through an importance weighting mechanism, following $Z_M^S = \tilde{w} \odot \tilde{H}_M^S + \hat{w} \odot \hat{H}_M^S$, where $\tilde{w}$ and $\hat{w}$ represent two learnable parameters, and $\odot$ denotes the element-wise multiplication. On such basis, we can adaptively uncover the intricate interconnections among the traffic signals.

Temporal Encoder. In addition to the spatial correlation among traffic signals, it is also crucial to capture the temporal dynamics of the traffic patterns across different timesteps. Analogous to the spatial encoder above, we adopt a temporal position encoding denoted as matrix $P_M^T \in \mathbb{R}^{K\times K}$, which is initialized by discrete one-hot embeddings representing K-timesteps. Subsequently, the encoding is consistently concatenated with the hidden token representations at each node, obtaining $H_M^T = f_T\left(H_M \parallel P_M^T\right)$, where f_T denotes the linear mapping function. In this phase, we also leverage a temporal-oriented multi-head attention mechanism with residual connections to capture the latent temporal dependencies across different timesteps for each traffic signal to obtain the temporally enhanced representation Z_M^T.

So far, both Z_M^S and Z_M^T have been learnt separately. To facilitate the multi-source information integration, we then apply a gating mechanism to fuse the hidden embeddings in both spatial and temporal dimensions to obtain the spatio-temporal aware representation Z_M [5].

3.2 Return-Driven Action Decoder

Given the spatio-temporally enhanced representations of the returns, states, and actions, we then attempt to predict the next action with the causal decoder. In this phase, inspired by [6], to efficiently "index" the representations of states and actions based on the returns, we incorporate a return-based embedding sub-space transformation scheme to transform the input data into distinct subspaces within the input-dimension. Specifically, for each timestep, we encode the return representation Z_R into the sate and action representations. The process can be formulated with $\tilde{Z}_S = Z_S \odot Z_R$ and $\tilde{Z}_A = Z_A \odot Z_R$, where $\odot$ denotes element-wise product between two vectors. In this way, the return-encoded representations $\tilde{Z}_S$ and $\tilde{Z}_A$ can serve as inputs for the casual decoder to predict actions tokens. Accordingly, the input trajectory is transformed into following structure: $\tau^t = \left(\tilde{z}_S^1, \tilde{z}_A^1, \tilde{z}_S^2, \tilde{z}_A^2, ..., \tilde{z}_S^t\right)$. Subsequently, the re-organized sequence is fed into a transformer decoder [9] to predict subsequent actions autoregressively using a casual self-attention mask, following: $p_A^t = \text{TransformerDecoder}\left(\tau^t\right)$. Since the TSC actions are discrete phases, we formulate our prediction task as a classification problem. Accordingly, we employ the cross-entropy loss as the optimization objective, as shown below:

$$\mathcal{L}_p = -\sum_{i=1}^{C} p_A^t \log\left(a^t\right), \tag{2}$$

where C is the total number of phases.

3.3 Return-based Contrastive Learning

Since we formulate the task as a *return*-guided action prediction task, we further design an auxiliary task to contrastively enhance the discriminability of the return representations. Specifically, given a specific anchor return token, we employ two data augmentation techniques to obtain the corresponding positive and negative samples. For the positive sample, denoted as R^+, we mask the input features of the anchor return with a constant, such as 0. To generate the negative sample R^-, we process each timestep individually and perform a row-wise shuffle on the feature matrices within the timestep. Consequently, for each anchor return token, there is precisely one positive and one negative. Then, we employ a binary discriminator $\mathbf{D}: \mathbb{R}^d \times \mathbb{R}^d \rightarrow [0, 1]$ to classify the anchor-positive pairs and anchor-negative pairs. We further leverage the binary cross-entropy loss to optimize the contrastive learning process:

$$\mathcal{L}_c = -y \log\left(\sigma\left(f\left(R, R^+\right)\right) + (1-y)\log\left(1 - \sigma\left(f\left(R, R^-\right)\right)\right)\right), \tag{3}$$

where σ denotes the sigmoid activation function, and y indicates the label of pairwise inputs. The goal for such design is to encourage the modeling of topological gaps, and challenge the model to discern graph structures from random ones, strengthening its capacity to recognize spatial patterns. Hence, the final loss function is: $\mathcal{L} = \mathcal{L}_p + \alpha * \mathcal{L}_c$, where the hyperparameter α adjusts the weight of $\mathcal{L}_c$.

4 Experiments

4.1 Experimental Setup

We evaluate the performance of our model on two public real-world benchmark datasets [8, 15, 16], namely *Hangzhou-4x4* and *Jinan-3x4*, with total number of traffic signals 16 and 12 respectively. To obtain the offline data, we collect the trajectories by training a state-of-the-art RL-based TSC model, i.e., AdvancedColight [22], and saving

Table 1: Model Performance Comparison.

Algorithm	*Hangzhou-4x4*			*Jinan-3x4*		
	AQL	AP	ATT	AQL	AP	ATT
MaxPressure	40.3	13.5	291.6	223.4	74.6	276.2
CoLight	38.6	12.3	290.0	214.0	71.5	271.9
AdvancedCoLight	24.5	9.4	272.5	152.9	48.8	247.4
BehaviorCloning	26.4	9.7	279.2	159.3	52.1	249.5
DecisionTransformer	25.3	9.7	275.4	159.1	50.5	252.6
DataLight	23.5	9.2	272.3	154.5	49.9	249.0
TransformerLight	24.5	9.5	273.3	155.2	50.4	249.2
STLight	**21.8**	**8.1**	**270.5**	**150.4**	**48.1**	**245.8**

trajectories of states, actions, and rewards at each timestep. Furthermore, we preprocess the long trajectories with episode lengths by creating slices with sequence length K=4 iteratively. During model evaluation, we leverage the traffic simulator *CityFlow* [20] as the environment for real-time traffic simulations. Each training/evaluation epoch lasts 3600s, while the green time duration for all possible phases is set to 15s. We train all models with 100 epochs and online evaluation is performed every 10 epochs. Evaluation results demonstrate the average of the last 5 evaluation epochs.

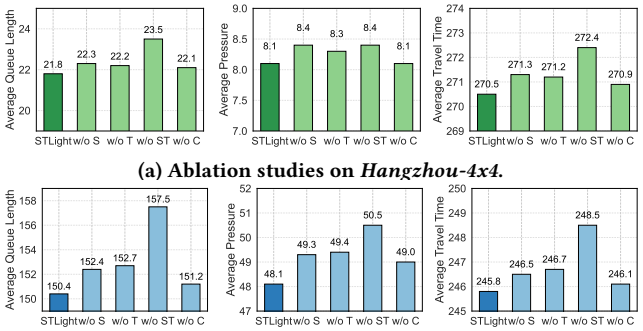
4.2 Compared Methods

We compare *STLight* with three categories of benchmark methods:
Heuristic Approach. MaxPressure [11] is a classical rule-based method that selects the phase based on pressure of queued vehicles between different incoming and outgoing directions.
Online RL. Colight [15] and AdvancedColight [22] are multi-agent DQN models with GAT for neighbor information aggregation.
Offline RL. Behavior Cloning [3] is an imitation learning baseline that reproduces actions given states as inputs. Decision Transformer [3] is a sequence-modeling based method that predicts actions given historical trajectories of rewards and states autoregressively. DataLight [21] is an offline RL model with Conservative Q-Learning. TransformerLight[18] adopts a gated Transformer for causal signal phase action prediction.

In terms of evaluation metrics, we select three evaluation metrics commonly adapted in the TSC task, including Average Queue Length(*AQL*), Average Pressure(*AP*), and Average Travel Time(*ATT*).

4.3 Overall Performance

We report the comparative analysis between our model and the baselines in Table 1. The results demonstrate that STLight outperforms competing methods on both the *Hangzhou-4x4* and *Jinan-3x4* datasets. Specifically, offline models like the Decision Transformer and TransformerLight eliminate the need for online explorations while maintaining competitive performance, which validate the efficacy of the shift from online-RL based TSC methods to offline modeling. Among these offline approaches, our model surpasses the best-performing DataLight model by 7.2% in average queue length evaluated on the *Hangzhou* dataset. This underscores the significance of sequential modeling and capturing dependencies within the MDP sequences since DataLight learns from individual MDP tokens rather than sequences. Moreover, relative to the offline sequence-modeling method TransformerLight which is also adapted from Decision Transformer, our approach registers improvements



(a) Ablation studies on *Hangzhou-4x4*.

(b) Ablation studies on *Jinan-3x4*.

Figure 2: Ablation studies on two cities.

of 4.6% and 1.4% in average pressure and average travel time respectively on the *Jinan* dataset, demonstrating the effectiveness of spatio-temporal sequence modeling in the traffic signal control task. Overall, experimental results demonstrate the superiority of our model in the TSC task compared to the SOTA baselines.

4.4 Ablation Studies

To evaluate the effectiveness of different modules of our model, we conduct the ablation studies with the following variants including *STLight-S* which removes the spatial encoder, *STLight-T* which removes the temporal encoder, *STLight-ST* which is the model without the spatio-temporal encoder, and *STLight-C*, the variant that excludes the return-based contrastive learning module. As shown by the experimental results on both datasets in Figure 2a and Figure 2b, we can conclude that both the spatial and temporal encoders are necessary in contributing to the overall performance. Without the spatio-temporal encoder, average queue length drops by 7.8% and 4.7% respectively on *Hangzhou* and *Jinan*. Furthermore, the importance of the contrastive learning module is underscored by the performance drop of 0.1% in average travel time on the two datasets after removing the contrastive learning module.

5 Conclusion

In this study, we introduced *STLight*, a novel spatio-temporal sequence modeling method for offline traffic signal control. We first formulated the task as a sequence of Markov Decision Processes spanned over the spatial dimension. Then, we proposed our model consisting of a spatio-temporal encoder that discerns spatial dependencies among traffic signals and the temporal dependencies across different timesteps. Besides, we enhanced the return-driven sequence-modeling method by representing target returns as auxiliary tasks to facilitate adaptive decision-making. Our model achieves superior performance among all compared baselines, which draws the conclusion that STLight provides a feasible solution to facilitate the deployment of learning-based TSC methods in the real world.

6 acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.92370204), National Key R&D Program of China (Grant No.2023YFF0725001), Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2023A03J0008), Education Bureau of Guangzhou Municipality.

References

[1] James Ault and Guni Sharon. 2021. Reinforcement learning benchmarks for traffic signal control. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

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

[3] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems* 34 (2021), 15084–15097.

[4] Zulong Diao, Xin Wang, Dafang Zhang, Yingru Liu, Kun Xie, and Shaoyao He. 2019. Dynamic spatial-temporal graph convolutional neural networks for traffic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 890–897.

[5] Albert Gu, Caglar Gulcehre, Thomas Paine, Matt Hoffman, and Razvan Pascanu. 2020. Improving the gating mechanism of recurrent neural networks. In *International conference on machine learning*. PMLR, 3800–3809.

[6] Sachin G Konan, Esmaeil Seraj, and Matthew Gombolay. 2023. Contrastive decision transformers. In *Conference on Robot Learning*. PMLR, 2159–2169.

[7] Jiaqi Liu, Peng Hang, Jianqiang Wang, Jian Sun, et al. 2023. MTD-GPT: A Multi-Task Decision-Making GPT Model for Autonomous Driving at Unsignalized Intersections. *arXiv preprint arXiv:2307.16118* (2023).

[8] Hao Mei, Xiaoliang Lei, Longchao Da, Bin Shi, and Hua Wei. 2022. LibSignal: An Open Library for Traffic Signal Control. *arXiv preprint arXiv:2211.10649* (2022).

[9] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[10] Qian Sun, Le Zhang, Huan Yu, Weijia Zhang, Yu Mei, and Hui Xiong. 2023. Hierarchical reinforcement learning for dynamic autonomous vehicle navigation at intelligent intersections. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4852–4861.

[11] Pravin Varaiya. 2013. Max pressure control of a network of signalized intersections. *Transportation Research Part C: Emerging Technologies* 36 (2013), 177–195.

[12] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat* 1050, 20 (2017), 10–48550.

[13] Yanan Wang, Tong Xu, Xin Niu, Chang Tan, Enhong Chen, and Hui Xiong. 2020. STMARL: A spatio-temporal multi-agent reinforcement learning approach for cooperative traffic light control. *IEEE Transactions on Mobile Computing* 21, 6 (2020), 2228–2242.

[14] Hua Wei, Chacha Chen, Guanjie Zheng, Kan Wu, Vikash Gayah, Kai Xu, and Zhenhui Li. 2019. 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*. 1290–1298.

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

[16] Hua Wei, Guanjie Zheng, Vikash Gayah, and Zhenhui Li. 2021. Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation. *ACM SIGKDD Explorations Newsletter* 22, 2 (2021), 12–18.

[17] Libing Wu, Min Wang, Dan Wu, and Jia Wu. 2021. DynSTGAT: Dynamic spatial-temporal graph attention network for traffic signal control. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2150–2159.

[18] Qiang Wu, Mingyuan Li, Jun Shen, Linyuan Lü, Bo Du, and Ke Zhang. 2023. TransformerLight: A Novel Sequence Modeling Based Traffic Signaling Mechanism via Gated Transformer. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA).

[19] Mingxing Xu, Wenrui Dai, Chunmiao Liu, Xing Gao, Weiyao Lin, Guo-Jun Qi, and Hongkai Xiong. 2020. Spatial-temporal transformer networks for traffic flow forecasting. *arXiv preprint arXiv:2001.02908* (2020).

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

[21] Liang Zhang and Jianming Deng. 2023. Data Might be Enough: Bridge Real-World Traffic Signal Control Using Offline Reinforcement Learning. *arXiv preprint arXiv:2303.10828* (2023).

[22] Liang Zhang, Qiang Wu, Jun Shen, Linyuan Lü, Bo Du, and Jianqing Wu. 2022. Expression might be enough: Representing pressure and demand for reinforcement learning based traffic signal control. In *International Conference on Machine Learning*. PMLR, 26645–26654.

[23] Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. 2019. Graph convolutional networks: a comprehensive review. *Computational Social Networks* 6, 1 (2019), 1–23.

[24] Zhiyue Zhang, Hongyuan Mei, and Yanxun Xu. 2023. Continuous-Time Decision Transformer for Healthcare Applications. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 6245–6262.

[25] Kesen Zhao, Lixin Zou, Xiangyu Zhao, Maolin Wang, and Dawei Yin. 2023. User Retention-oriented Recommendation with Decision Transformer. In *Proceedings of the ACM Web Conference 2023*. 1141–1149.