Accident Analysis and Prevention 144 (2020) 105655

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Multi-Objective reinforcement learning approach for improving safety at intersections with adaptive traffic signal control*

Yaobang Gong*, Mohamed Abdel-Aty, Jinghui Yuan, Qing Cai

Department of Civil, Environmental and Construction Engineering, University of Central Florida, Orlando, FL 32816, USA

ARTICLE INFO

Keywords:
Traffic safety
Adaptive Signal control
Multi-objective reinforcement learning
Deep learning

ABSTRACT

Adaptive traffic signal control (ATSC) systems improve traffic efficiency, but their impacts on traffic safety vary among different implementations. To improve the traffic safety pro-actively, this study proposes a safety-oriented ATSC algorithm to optimize traffic efficiency and safety simultaneously. A multi-objective deep reinforcement learning framework is utilized as the backend algorithm. The proposed algorithm was trained and evaluated on a simulated isolated intersection built based on real-world traffic data. A real-time crash prediction model was calibrated to provide the safety measure. The performance of the algorithm was evaluated by the real-world signal timing provided by the local jurisdiction. The results showed that the algorithm improves both traffic efficiency and safety compared with the benchmark. A control policy analysis of the proposed ATSC revealed that the abstracted control rules could help the traditional signal controllers to improve traffic safety, which might be beneficial if the infrastructure is not ready to adopt ATSCs. A hybrid controller is also proposed to provide further traffic safety improvement if necessary. To the best of the authors' knowledge, the proposed algorithm is the first successful attempt in developing adaptive traffic signal system optimizing traffic safety.

1. Introduction

As one of the most important Active Traffic Management (ATM) strategies, Adaptive Traffic Signal Control (ATSC) helps improve traffic efficiency of signalized arterials and urban roads by adjusting the signal timing in response to the dynamic traffic demand. However, the safety effects of state-of-practice ATSCs are not consistent. Some studies show that the installation of ATSCs reduces the number of crashes (Ma et al., 2016; Khattak et al., 2018). While another study concludes that the crash frequency before and after the implementation of ATSCs is not significantly different (Fink et al., 2016). A study on traffic conflicts, which is one of the surrogate safety measures, even found that there is a considerable increase in both frequency and severity of conflicts following the installation of the ATSC (Tageldin et al., 2014). The mixed evidence raises the concern of ATSC's safety impact.

This study advocates designing an ATSC that is able to ensure or improve traffic safety, i.e. a safety-oriented ATSC. Several recent studies have found that signal timing is related to the crash occurrence at signalized arterials and intersections (Yuan et al., 2019, 2018; Yuan and Abdel-Aty, 2018). By applying the models developed by the aforementioned studies, the safety-oriented ATSC is able to reduce the crash occurrence by dynamically optimizing its signal timing in response to different traffic conditions. Moreover, the proposed safety-oriented ATSC could also serve as a strategy of pro-active traffic safety management to improve traffic safety of arterials and urban roads like other ATM strategies (e.g. Ramp Metering and Variable Speed Limits) do for freeways (Abdel-Aty et al., 2006; Wang et al., 2017; Yu and Abdel-Aty, 2014).

Although the safety-oriented ATSC aims at optimizing traffic safety, it should not be detriment to efficiency. Therefore, we proposed an ATSC system that utilizes a multi-objective framework to simultaneously optimize traffic efficiency and safety. A real-time crash risk model (Abdel-Aty et al., 2004) is applied to generate the indicator of near-future crash likelihood. The multi-objective reinforcement learning algorithm is used for optimization. The proposed algorithm was tested in a simulated real-world isolated intersection. Its performance in terms of delay and crash risk reduction was compared with a replicated field controller and an ordinary ATSC optimizing only traffic efficiency. A discussion about the control policy of proposed safety-oriented ATSC and the potential impact of different signal configuration is also provided.

* This paper has been handled by associate editor Tony Sze

* Corresponding author.
E-mail address: gongyaobang@knights.ucf.edu (Y. Gong).

https://doi.org/10.1016/j.aap.2020.105655
Received 27 February 2020; Received in revised form 28 May 2020; Accepted 15 June 2020
Available online 14 July 2020
0001-4575/ © 2020 Elsevier Ltd. All rights reserved.

2. Related work

2.1. ATCS considering traffic safety

There have been several studies on optimizing signal timing considering traffic safety (Stevanovic et al., 2015, 2013; Zhu et al., 2019). Almost all of them employed surrogate safety measures as the safety indicator and all these studies focused on fixed-timing controllers. The burden of extending such fixed-timing controllers to safety-oriented ATSC is that the most modern ATSCs are designed to be pro-active/predictive. In other words, the ATSC needs to know how the signal timings affect the future safety condition. Therefore, to use the surrogate safety measures (e.g. traffic conflicts) as the safety indicator, a prediction model of the future surrogate safety measures needs to be developed.

To the best of the authors' knowledge, only one published study (Sabra et al., 2013) developed a non-parametrical traffic conflict prediction model and proposed a safety-oriented ATCS accordingly. The study developed a four-stage algorithm tuning the cycle length, splits, offsets, and left-turn phase sequence sequentially. At each stage, the "predicted" number of traffic conflicts is used to evaluate the signal timing tuned by the control algorithm. If the "predicted" traffic conflict of new signal timing is greater than that of the current one, the controller keeps using the current signal timing. Otherwise, the new signal timing is applied. The proposed algorithm is tested in a simulated real-life arterial corridor and a simulated real-life grid network. For the arterial case, although the proposed algorithm reduces the number of traffic conflicts compared with a coordinated actuated signal optimized by the authors, it does increase the number of traffic conflicts compared with the existing field controllers. For the grid network, the algorithm increases the number of traffic conflicts compared with a coordinated actuated signal optimized by the authors, and no testing results of existing field operation are provided. Therefore, the ability of the proposed algorithms in improving traffic safety is not conclusive.

2.2. Multi-objective ATSCs using reinforcement learning

In the past decades, reinforcement learning (RL) algorithms have been widely applied to develop ATSCs (Yau et al., 2017). Signal control agents using RL algorithms to learn a *policy,* which maps the perceived environment (e.g. traffic condition), i.e. *state*, to *actions* taken by the controller. The *rewards* received from the environment, which represent the objective, direct the agents to learn an optimized *policy*. The discussion of reinforcement learning will be elaborated on in the next section.

Some studies have proposed RL-based ATSC with multiple optimization objectives. Based on the way to achieve the multi-objective optimization, they could be classified into three different types.

The first type is an ATSC which is able to switch its objective dynamically. Houli et al. (2010) developed an ATSC with three different backend single-objective RL algorithms with different goals. But only one algorithm is activated according to the traffic condition. When the current traffic condition is free flow, the goal of the ATSC is minimizing the number of stops. When it is under medium traffic condition, the goal turns to minimize the overall waiting time. When there exists congestion, the goal is switched to minimize queue length to avoid queue spillover. This type of ATSC is not suitable for this study as the safety and efficiency have to be simultaneously optimized.

The second type of algorithms creates a synthetic reward to account for multiple objectives simultaneously. The most straightforward approach is using the simple/weighted average of multiple rewards. Each reward is associated with a policy goal. Khamis and Gomaa (2014) proposed an ATCS with 7 different objectives. Five of them indicate different aspects of traffic efficiency, one represents the fuel consumption and the last one is claimed to be "safety reward". The so-called safety reward is not any safety measure but essentially the average speed of the vehicles. The potential issue of this type of algorithm is that its convergence is not thematically proved since the synthesizing reward is no longer the decomposition of any policy goals. It should be noted that the reward of several single-objective ATSCs (Muresan et al., 2018; Van Der Pol and Oliehoek, 2016; Vidhate and Kulkarni, 2017) could also be the average of several components.

The third type of algorithm is multi-objective RL (MORL). Although similar to the second type, MORL manipulates the *value function* rather than the rewards. *Value function* represents the expected long-term reward, which implies the goal. In such algorithms, different *value functions* are learned independently using different rewards, while the second type of algorithms learns a signal *value function* by the single synthetic reward. This is beneficial to the convergence of the algorithm especially when the objectives are irrelevant. For a multi-objective ATSC using the aforementioned algorithm, the control agent could either choose the action based on the weighted average of multiple *value functions* or use one or more of them as the thresholds. The study conducted by Jin and Ma (2015) utilized the later method to assign priority to arterials.

According to the review of existing studies, there is still a dearth of research in developing an effective ATSC optimizing traffic efficiency and safety. To optimize the two objectives simultaneously, multi-objective reinforcement learning (MORL) is selected as the backend algorithm.

3. Background

3.1. Real-time crash risk models

The first step of optimizing the signal timing for traffic safety is understanding how they are correlated. In the pro-active perspective, the impact of a specific signal timing on the future crash potential needs to be quantified. Recently, researchers (Yuan et al., 2019, 2018; Yuan and Abdel-Aty, 2018a) have proposed real-time crash risk models to examine the relationships between future crash potential and traffic conditions including signal timing. The basic assumption underlying real-time crash risk models is that there exist certain conditions that are relatively more "crash-prone" than the others, which could be called "crash precursors". For example, conditions that are just before the crash occurrence would be regarded as "crash condition". By comparing the characteristics of "crash conditions" with "non-crash conditions", crash precursors could be identified. Like other binary classification models, the output of real-time risk models indicates the forecasted crash potential. It could be the probability of the crash or the odds of crash versus non-crash. Similar to the efficiency measures like delay, the forecasted crash potential could be directly employed by "predictive/pro-active" controllers to assess the future safety effect of a candidate signal timing in real-time.

3.2. Multi-objective reinforcement learning

RL (Sutton and Barto, 2018) is a goal-oriented machine learning algorithm. It learns to achieve the *goal(s)* over discrete time intervals by interacting with the environment. In each time interval, an RL *agent* observes the *state s* of the environment, takes an *action a* accordingly based on its knowledge *policy* π, receives the feedback *reward r* (could also be a penalty) from the environment, which accumulates to the long-term *goal*, and transits to the *next state* s' with the state *transition probability P*. During the learning process, it keeps updating its *policy* by maximizing the expectation of the long-term *reward*, which is *value function* of value-based RL, until it converges to the *optimal policy* π^*. For the single objective RL problem, the *Q value*, or action value, refers to the expected long-term discounted reward for selecting *action a* at *states* following the *policy* π, is defined as:

$$Q_\pi(s, a) = E[R_t | s_t = s, a_t = a] \quad (1)$$

And it decomposes into the Bellman equation:

$$Q_\pi(s, a) = \sum_{s',r} P(s', rs, a)[r + \gamma \sum_{a'} \pi(a's')Q_\pi(s', a')] \quad (2)$$

The discount factor γ indicates the importance of future rewards. A higher γ means the future reward is more important.

For Q-value-based RL, the *optimal policy* π^* guides the agent to choose actions that maximize the Q value. Thus, the optimal Q value function is defined as:

$$Q^*(s, a) = \max_p Q_\pi(s,\mathrm{a}) \quad (3)$$

And the optimal policy is obtained by:

$$\pi^*(s) = \operatorname{argmax}_a Q^*(s, a) \quad (4)$$

In the context of Multi-Objective Reinforcement Learning (MORL), multiple *goals* are optimized simultaneously. In MORL, each objective has its associated reward and value function. Thus, Q-values are expressed as Q vector $MQ_\pi(s, a)$:

$$MQ_\pi(s, a) = [Q_\pi^1(s,\mathrm{a}),\ Q_\pi^2(s,\mathrm{a}), ..., Q_\pi^n(s,\mathrm{a})]^T \quad (5)$$

Intuitively, the optimal Q vector is defined as

$$MQ^*(s, a) = \max_p MQ_\pi(s,\mathrm{a}) \quad (6)$$

The "maximum operation" of a vector could have different definitions. Generally, there are two ways for MORLs handling the "maximum operation": single-policy MORL approach and multi-policy MORL approach (Liu et al., 2015). Single policy approaches aim to find the best single policy representing the preferences or the trade-off among the objectives. Several different algorithms are developed to determine and express the preferences or trade-off, such as linear/non-linear weighted sum approach, W-learning, AHP approach, ranking approach, and geometric approach, etc. Multi-policy MORL aims at approximating the Pareto front by a set of policies. The Pareto front is a set of Pareto non-dominated solutions. If any objective of solution could not be improved without sacrificing at least one other objective, the solution is a Pareto non-dominated solution.

4. Algorithm

In this study, the control problem is formulated into the MORL setting: the signal controller acts as an RL *agent*; it observes the traffic condition of the intersection and the current signal status as the *state*; it directly selects the appropriate phase as its *action*; the waiting time of vehicles acts as the efficiency *reward* while the risk score derived from the real-time crash risk model acts as the safety *reward*; and the *goals* of the agents are reducing the delay (efficiency) and future crash potential (safety). Weighted sum approach is selected to develop a single policy MORL and one of the famous deep reinforcement learning algorithms Double Dueling Deep Q Network (3DQN) is utilized as the backend learning algorithm. The details of the algorithm are elaborated on in the subsections below.

4.1. State

The *state* used in this study includes two components: a binary matrix indicating the traffic condition of the intersection and the current activated signal phase. For the traffic condition, the proposed algorithm employs a "camera-like" virtual traffic detector to capture the locations of individual vehicles occupying the intersection approaches. It provides more heterogeneous travel information than aggregated traffic parameters. For example, blockage of the left-turning lane could be captured. In order to simulate the limited detection range of traffic cameras, the virtual traffic detectors only detect the vehicles within a certain distance from the stop line. Fig. 1 shows how the traffic condition matrix is generated. "Virtual-loop" concept that is widely used in video detection is applied, which is basically a short segment of an intersection approach. If a "virtual-loop" is occupied by a vehicle, the corresponding element in the traffic state matrix turns from 0 to 1. The length of each virtual loop detectors in this study is 15 feet and the maximum number of loop detectors for each lane is 20.

The current activated signal phase is also recorded as a part of the *state*. The signal phase is defined as the combination of two or more non-conflicting vehicular movements. Figs. 2,3 shows a typical eight-phase schema of a four-way intersection. An "interphase" refers to the clearance time including yellow time and all-red clearance. The current signal phase is coded as a vector with a length of n + 1, where n is the number of phases of the signal. The last digit indicates whether the phase is interphase or not.

4.2. Action

The *action* of the agent is selecting the appropriate signal phase at each time interval based on the current *policy*. If a phase changing occurs, the controller will activate the interphase to clear the intersection.

Several other rules are applied to restrict the arbitrary selection of actions to ensure traffic safety and overcome some fundamental limitations of RL-based ATSC:

1) *Ensure minimum green time (g_{min})*: the minimum green time concept is used in actuated signal control to satisfy the driver's expectation (Arroyo et al., 2015). If a minimum green time is set too low (or even omitted) and violates the driver's expectation, there exists a risk of increased rear-end crashes. Therefore, the controller is configured not to allow the change phase if the minimum green time is not satisfied. The values of minimum green times are set to be the same as the ones used in the field to avoid double investigation. However, if there is no existing signal control, the values should be set based on the local traffic signal timing manual.

2) *Default phase (p_0)*: If there is no vehicle at the intersection, theoretically the RL-based ATSC randomly selects a phase to activate during the learning stage. This is detrimental to both traffic efficiency and safety. Therefore, a default phase that represents the major approach through movements is set to avoid the random phase changes.

3) *Maximum allowed waiting time ($t_{maxwaiting}$):* The benefit of setting a maximum allowed waiting time is ensuring fair travel rights. Consider an extreme case. There is only one vehicle waiting on the minor approach to turn left, while there are one hundred vehicles that are waiting on the major approach going through. As one of the objectives of ATSC is reducing the TOTAL delay, the controller would favor clearing the major approach, which results in excessively long waiting time for the vehicle on the minor approach. Therefore, a maximum allowed waiting time is configured to prevent the occurrence of such a situation.

4.3. Rewards

Two rewards are designed for traffic efficiency and safety. As for the efficiency, the goal is minimizing the travel time of a vehicle, or the delay, which theoretically is the difference between the actual and expected travel time. However, it is not feasible to obtain the travel time or delay as they are only available when the vehicle reaches its destination. Thus, the cumulative waiting time of the queued vehicles is used as the goal indicator. The reward representing the traffic efficiency is defined as the difference of the current and previous cumulative waiting time of all vehicles:

$$\boldsymbol{r_{te}} = -(\boldsymbol{W_{t+1}} - \boldsymbol{W_t}) \quad (11)$$

where W_{t+1}, W_t are the waiting time of step $t + 1$ and t. The reward could be interpreted as such: when the vehicles are queued, the agent will be penalized; and when the queued vehicles are discharged, the agent will be rewarded.

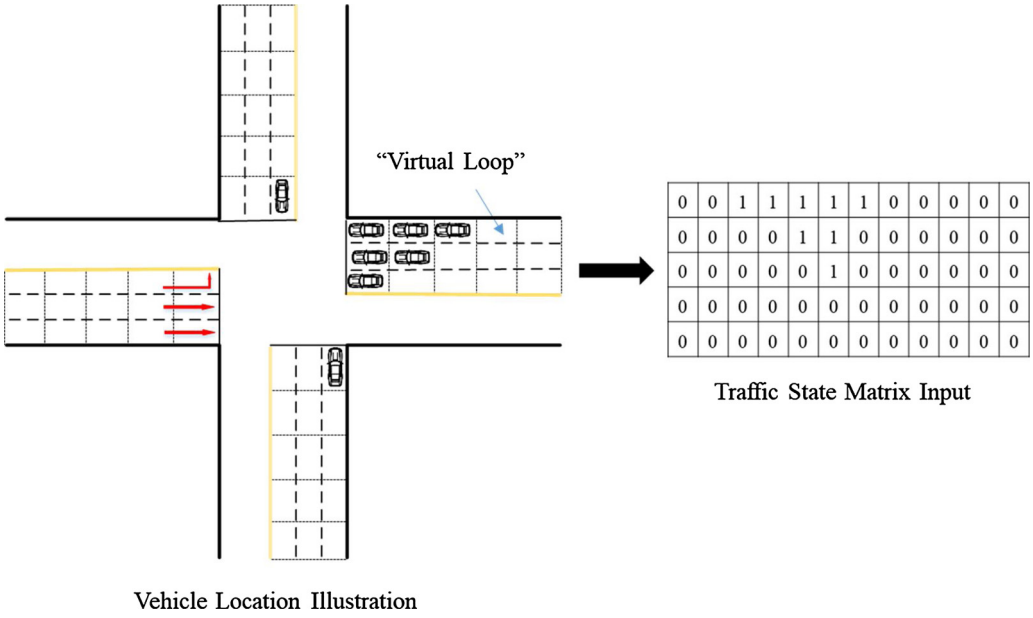

Fig. 1. Traffic State Representation (Gong et al., 2019, with permission).

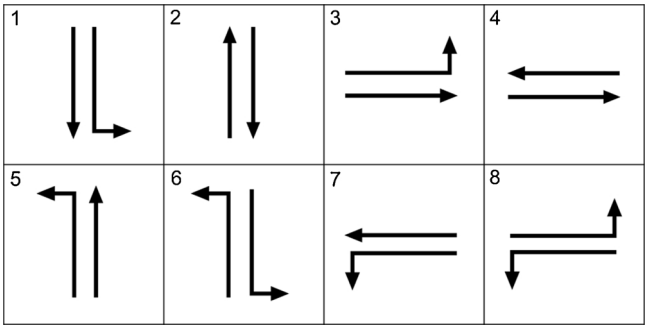

Fig. 2. Eight phases for four-way intersections.

As for safety, a risk score is utilized to indicate the relative risk level. The score is calculated using a real time crash risk model calibrated based on the local historical crashes and traffic data. As the score is site-specific, its calculation process varies for different kinds of intersections, different locations, and different interests of the users. Other surrogate safety measures that imply the future crash risk could also be used as the "risk score".

In this study, the reward for safety is defined as the adjusted risk score by a baseline:

$$r_{ts} = \begin{cases} riskscore_t - riskscore_{base} \ when \ riskscore \ is \ generated \\ 0 \ otherwise \end{cases} \quad (12)$$

where $riskscore_t$ is the risk score at timestamp t and $riskscore_{base}$ is a baseline risk score calculated during the pre-training. It should be noted that the risk score might not be generated at every control step. In this case, the risk reward acts as a "delay" reward.

The reward could be interpreted as an "advantage": when the reward is positive, the safety performance is better than the baseline, which means that the agent will be rewarded; otherwise, the agent will be penalized. Defining the reward as an advantage term accelerates the learning process as it helps the agent to find the direction.

It should be also noted that the rewards for traffic efficiency and safety are used to direct the training process. Once the control agent is well-trained, such rewards are no longer needed during the operation.

4.4. Weighted sum approach for single policy MORL

Weighted sum approach (Karlsson, 1997), one of the single policy MORL algorithms, is selected due to its computational efficiency since the multi-policy algorithms are computational intractable for the ATSC problems that require real-time decision making. It computes a linearly weighted sum of Q-values for all the objectives to obtain a synthetic Q function:

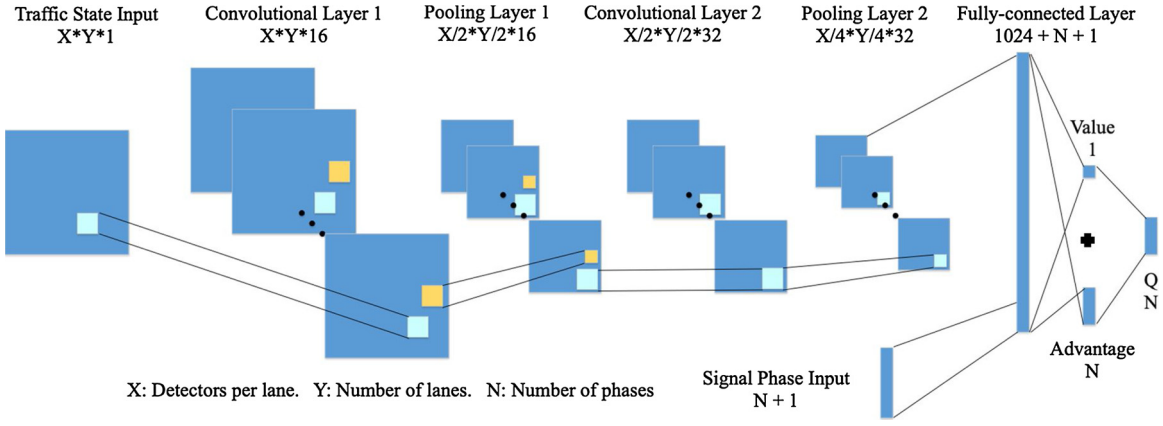

Fig. 3. The structure of the neural network used in the learning algorithm (adopted from Gong et al., 2019, with permission).

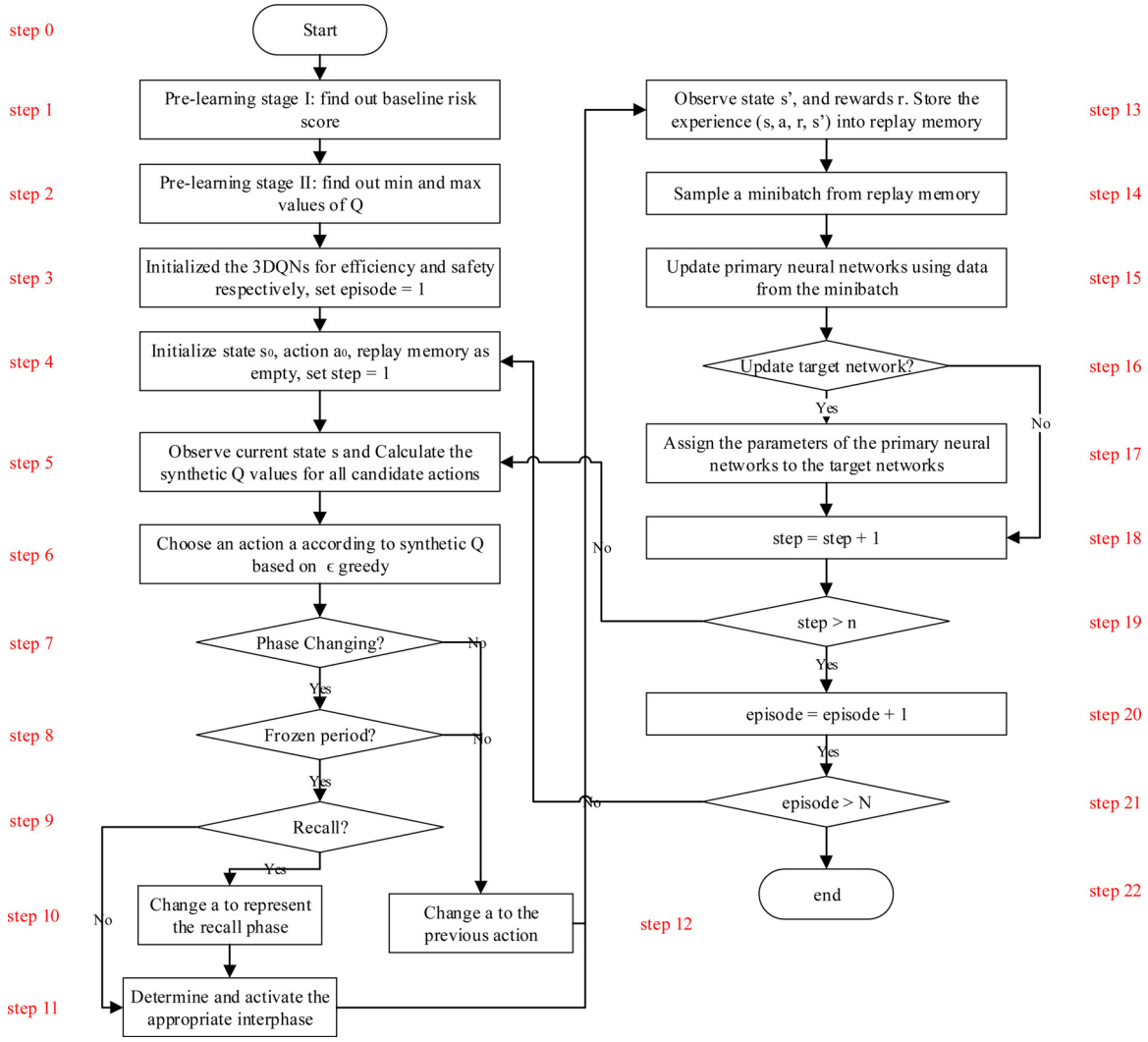

Fig. 4. ATSC Algorithm flow chart.

$$SQ(s, a) = \sum_{i=1}^{N} w_i Q_i(s, a) \tag{13}$$

where $SQ(s, a)$ is the synthetic Q value; $Q_i(s, a)$ is the Q value of the ith objective; w_i is the weight, it implies the relative importance of the specific ith objective. The weights could be pre-configured by the algorithm developers or be determined by the users.

In this study, the synthetic Q value is the weighted sum of the normalized Q values:

$$Q(s, a) = w_e \frac{Q_e(s, a) - Q_{e,min}(s, a)}{Q_{e,max}(s, a) - Q_{e,min}(s, a)} + w_s \frac{Q_s(s, a) - Q_{s,min}(s, a)}{Q_{s,max}(s, a) - Q_{s,min}(s, a)} \tag{13}$$

where $Q(s, a)$ is the synthetic Q value used to evaluate the actions; $Q_e(s, a)$ and $Q_s(s, a)$ is the Q value of traffic efficiency and traffic safety, respectively. Two Q values are normalized to the same magnitude by the min-max method since the rewards and Q values of the two objectives have a huge difference in terms of the magnitude. $Q_{i,min}(s, a)$ and $Q_{s,min}(s, a)$, $i \in [e, s]$ are the minimum and maximum values of the two Q values estimated from the pre-learning. w_e and w_s are the weights.

4.5. Backend learning algorithm

The proposed ATSC utilizes Double Dueling Deep Q Network (3DQN), one of the advanced Q-value-based deep learning algorithms, as its backend algorithm. Readers are encouraged to refer to Wang et al. (2015) for the technical details. 3DQN uses DNNs as its functional approximator. In this study, the convolutional neural network (CNN), one of the DNNs widely used in pattern recognition, is employed to construct the functional approximator. The structure of CNN used in the proposed algorithm is similar to the previous study by the authors (Gong et al., 2019). The CNN firstly takes the traffic state as the input. Then the traffic state is processed into a vector and connected with the signal phase input. Finally, the Q values of all actions are outputs. CNN takes the state as the input and outputs the Q values of all actions. It should be noted that the functional approximators of two Q values have the exact same structure but are trained separately.

4.6. Pre-Training

According to the design of the algorithm, there are two pre-requests for learning: first, getting the baseline risk score to derive the safety reward; second, getting the estimated range of the Q values to obtain normalized Q-values. Therefore, a two-phase pre-training is designed.

The objective of the first phase of the pre-training is to find out the baseline risk score. Since the risk score is a relative measure and site-specific, it is impossible to find a "best" risk score. Thus, the hourly-average risk score of a benchmark scenario, which is used to evaluate the proposed algorithm, is utilized. It means that the control agent is

"directed" to perform better than the benchmark signal controller does.

To estimate the range of Q-values, at the second phase of the pre-training, the control agent is asked to learn the exact policy of the benchmark signal controller or another reference controller if necessary. During the pre-learning, the agent observes the action taken by the benchmark controller rather than taking actions based upon its own policy and update the hyper-parameters of the functional approximator. The minimum and maximum of Q values generated by functional approximator during the learning course are recorded to get the estimated range of the Q values. It should be noted that as the baseline controller is not exactly the same as the optimal controller, the range of Q values of those two controllers might have a subtle difference. Therefore, there exists a dilemma that while it is impossible to know the range of Q values of the optimal policy before learning, without knowing the range of the Q values, it is impossible for the MORL agent to learn the optimal policy. Thus, a compromise could be achieved by using the reference controller if the baseline controller is not close enough to the optimal controller.

4.7. Overall algorithm

The flow chart of the proposed algorithm is presented in Fig. 4. Step 1 and 2 are the pre-learning steps to figure out the baseline risk score and min/max Q values. Step 3 and 4 are the initialization of the algorithm. The control process, which is basically activating the appropriate phase with certain constraints, is illustrated from step 5 to step 12. And the rest is the simplified learning process of the standard deep Q network model. The algorithm is coded by Python programming language using deep learning package Tensorflow (Abadi et al., 2015).

5. Case study

The proposed algorithm was tested in a simulated isolated intersection using a commercial traffic simulator Aimsun Next 8.3.0. The algorithm obtains the information from the simulator and implements its control policy to the simulated signal controllers. The simulated MORL agents were trained extensively. The performance of the well-trained agent is evaluated by the real-world signal timings and compared with an RL based ATSC optimizing only traffic efficiency (ATSC-SORL).

5.1. Simulation set up

The simulation scenario was built based on a real-world signalized intersection of North French Avenue (major arterial) and West 1st Street (minor arterial) in Seminole County, Florida. The intersection is a typical mid-size four-way intersection with moderate traffic volume. Fig. 5 shows the lane configuration of the approaches of the intersection. Right turns are permitted on red after a complete stop at the stop line, and the left-turns of the east-west approaches are configured as permitted-protected. In this study, the lane-based counts of all Tuesdays, Wednesdays, and Thursdays from January 2018 to March 2018 were extracted from the Automated Traffic Signal Performance Measures (ATSPM) system. Then the average counts for every 15 min are used to approximate the turning movement counts for a "normal weekday". Then they serve as the travel demands of the intersection. It should be noted that for the shared through-right-turning lane, the percentage of right turning is set as 30 %. As for the eastbound, as there exists a dedicated right-turning lane, the actual right turning volume is used.

Fig. 6 shows the 15-minutes counts for all four approaches. It shows that the

North-South approach is the major approach. The large volume of eastbound is caused by the right-turning vehicles using the dedicated right-turning lane.

In the field, the intersection is controlled by a coordinated actuated signal controller. In this study, a simulated controller using the signal timing provided by Seminole County is employed to replicate the field controller and serves as the benchmark (BC). The up-to-date timing plan, which was retimed on 03/07/2017, was used based on the travel demand study period. The benchmark signal timing includes three Time of Day (TOD) plans for coordination and the signal runs fully actuated during the nighttime. Fig. 7 shows the splits of TOD plans and max/min green time when the signal is fully actuated. The yellow time is five seconds and the all-red clearance time is two seconds. Another ATSC controller developed using a single objective RL algorithm (ATSC-SORL) (Gong et al., 2019), which aims at only optimizing traffic efficiency, is also used for comparison.

5.2. Real-time crash risk model

In this study, a real-time crash risk model is developed to forecast the crash odd in the next 5 − 10 min. The forecasted crash odds are used as the "risk score" to generate "safety reward".

As mentioned earlier, the real-time crash risk model is a binary classification problem, thus a binary logistic model is naturally preferred. If a crash occurred under certain conditions, the condition is classified as "crash" and vice versa. Suppose the "crash" case has the outcomes $y_i = 1$ and $y_i = 0$ with the respective probabilities of p_i and $1 - p_i$, $i = 1, 2, ...M$. M represents the total number of samples. The binary logistic regression can be expressed as:

$$y_i \sim Bernoulli(p_i) \tag{14}$$

$$logit(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_K X_{Ki} \tag{15}$$

where β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_K)$ is the coefficients vector, and $\boldsymbol{X_i} = (x_1, x_2, ..., x_K)$ is the independent variable vector for the ith observation.

Similar to the previous study by Yuan et al. (2019), the direct output of the binary logistic model is the predicted log crash odd of vehicles entering the intersection from a specific approach. As the number of crashes that have occurred at the test intersection is not sufficient to develop the model, crashes that have occurred in Seminole County, the same jurisdiction as the test intersection, were used. In total, data of 349 crashes from January 2017 to April 2018 were collected from Signal Four Analytics (S4A). These crashes occurring within the intersection area and the at-fault drivers were not under the influence of alcohol and drugs. Traffic data and signal timing logs of the intersections where crashes have occurred were extracted from ATSPM for a period of 10 min (divided into two 5-minutes time slices: slice 1 is 0 − 5 minutes and slice 2 is 5 − 10 min) prior to the crash occurrence. The data of different approaches were labeled using the same nomenclature as a previous study (Yuan and Abdel-Aty, 2018). The predicted approach is named as "A" approach, which is the traveling approach of the at-fault driver. "B", "C" and "D" approaches are labeled following a clockwise sequence (please refer to Yuan and Abdel-aty, 2018b for more details). Since the crashes are rare events, the "non-crash" events are randomly sampled to generate a balanced dataset. In this study, 3215 "non-crash" events and 349 "crash" events were collected to calibrate the final model.

Table 1 shows the modeling results. The model estimation results show that the future crash potential at signalized intersections is affected by various factors that are collected from the four approaches, including the green ratio from A approach, through volume from C approach, arrive on green from D approach, etc. These results indicate that the crash odd is represented by the complicated interactions between signal timing and vehicle arrivals.

As the value outputted by the model is the predicted "risk score" for one approach, the final "risk score" of the whole intersection is defined as the average "risk score" of four approaches.

The risk score is calculated every minute using a rolling horizontal approach. For example, at 19:00, the model uses data from 18:50 to

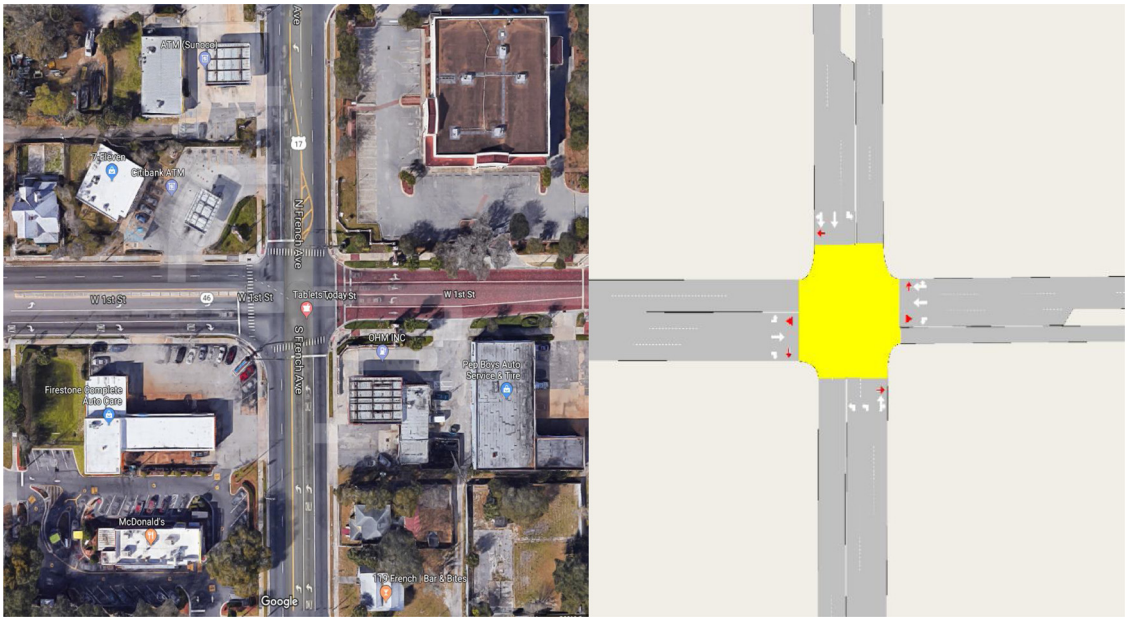

Fig. 5. Lane configuration of the simulated intersection.

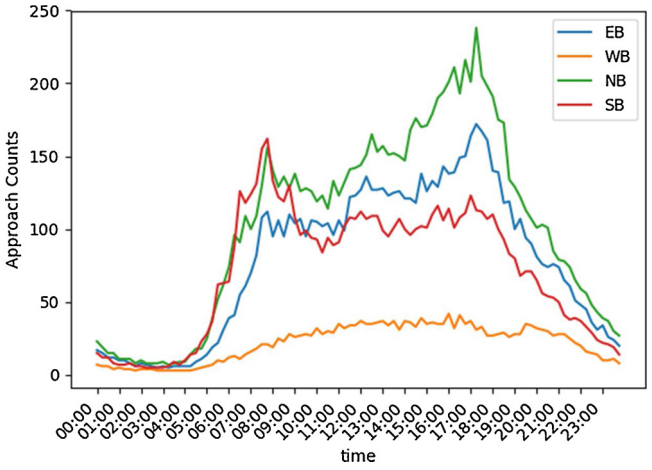

Fig. 6. 15-minutes counts of all approaches (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 1
Modeling Results of the Real-Time Crash Risk.

Variable	Coefficient	Standard Error	P-Values	Nomenclature
Intercept	−2.095	0.410	0.000*	*Prefix(approach label): A/B/C/D*
A_LT_GR_S1	1.790	1.048	0.087**	*Turning Movement*
A_TH_AOG_S1	0.003	0.002	0.076**	*LT: Left turning; TH: Through*
A_TH_GR_S2	−2.302	0.702	0.001*	*Variable Type:*
B_TH_GR_S2	−2.085	0.705	0.003*	*AOG: Number of vehicles arrived at the intersection on green*
C_LT_AOG_S2	−0.018	0.010	0.068**	*GR: Ratio of the green time within 5-minute*
C_TH_AOG_S1	0.007	0.002	0.000*	*Suffix(time slice): S1/S2*
C_TH_GR_S1	−2.167	0.624	0.001*	*Example: (A_ TH _AOG_S1): Number of through vehicles arrived at the intersection on green of approach A (predicted approach) at the time slice 1(0−5 minutes before the crash occurrence)*
C_TH_GR_S2	2.938	0.720	0.000*	
D_LT_AOG_S1	0.014	0.007	0.039*	
D_TH_AOG_S1	0.008	0.003	0.007*	
D_TH_GR_S1	−2.229	0.972	0.022*	
D_TH_GR_S2	1.786	0.920	0.052**	

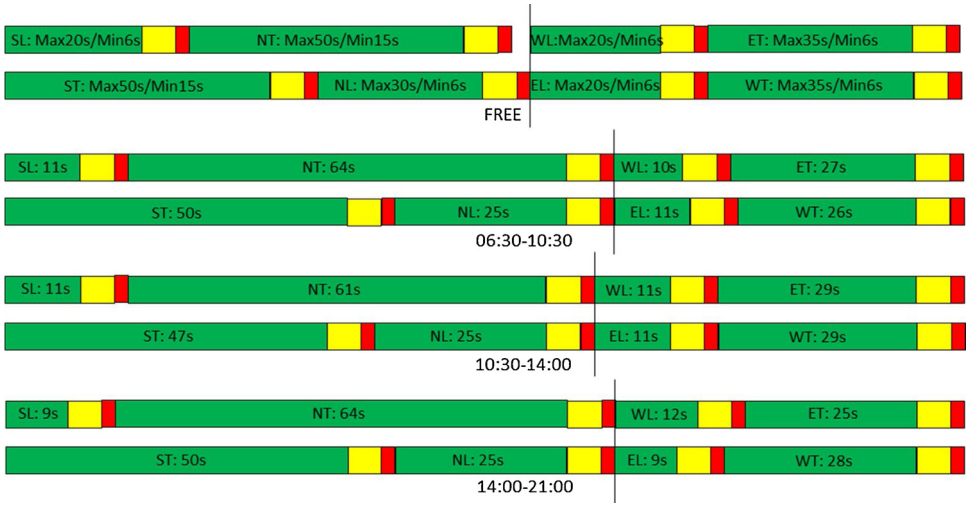

Fig. 7. Benchmark signal timing.

Table 2
Algorithm Setting.

Parameter	Value	Description	Note
N_a	8	Number of actions	Number of phases
g_{min}	10 s (6:30 − 21:00) 6 s(otherwise)	Minimum green time	
t_y	5 s	Yellow time	Same as benchmark
t_{ar}	2 s	All-red clearance time	
p_0	NT/ST	Default phase	Major approach through
$t_{maxwaiting}$	180 s	Maximum allowed waiting time	
w_e	0.5	Weight of the efficiency objective	
w_s	0.5	Weight of the safety objective	
γ	0.99	Discount factor	
ε_e	0.9999	Ending greedy	To avoid oscillation
M	20,000	Replay memory size	Roughly tuned
B	64	Minibatch size	
lr	0.00005	Learning rate	
t_{train}	1 s	length of training step	Same as control step

19:00 to forecast the crash likelihood from 19:05 to 19:10; at 19:01, data from 18:51 − 19:01 is used to forecast the crash likelihood from 19:06 to 19:11. Because the control step is typically several seconds, the risk score might not change between two control steps. In this case, the safety reward derived by the risk score act as a "delayed" reward.

5.3. Algorithm setting

Table 2 provides the algorithm setting. In general, the algorithm imitates the safety-related setting of the benchmark signal as much as possible. Moreover, the importance of traffic safety and efficiency was set to be equal.

The first phase of the pre-training was conducted using the benchmark signal controller, while the reference controller in the second phase of the pre-training is the ATSC-SORL controller.

5.4. Results

To evaluate the performance of the proposed multi-objective ATSC algorithm, the well-trained control agent (ATSC-MORL), the benchmark controller (BC) and the single objective controller (ATSC-SORL) were implemented for 30 simulation days. Three kinds of performance measures were observed: average daily delay per vehicle, average daily number of stops per vehicle and the average daily intersection crash risk score. Table 3 shows the average daily performance of the 30 simulated days.

According to the Table 3, compared to the BC controller, ATSC-MORL controller reduced average daily delay of by 25.93 % (26.395 s versus 19.550 s), average daily number of stops by 12.52 % (0.703 versus 0.615) and the average daily crash risk score by 8.89 % (0.045 versus 0.041). Compared with the ATSC-SORL controller, the ATSC-MORL did improve traffic safety and reduced the number of stops while increasing travel time. Interestingly, while the ATSC-SORL reduces the delay dramatically (49.1 %) comparing with the benchmark, it does increase the crash likelihood.

Table 3
Average Daily Performance of the Controllers.

Controller	Efficiency		Safety (Risk Score)
	Average Delay (sec)	Number of Stops	
BC	26.395	0.703	0.045
ATSC-SORL	13.434	0.691	0.072
ATSC-MORL	19.550	0.615	0.041

The performance of the three controllers at different times of day was also investigated. Fig. 8 shows the change of performance measures in 15-min aggregation intervals. Although ATSC-MORL performs well in most situations, there exist certain conditions that ATSC-MORL performs worse than the benchmark. First, ATSC-MORL tends to increase the average delay per vehicle dramatically when the traffic demand is extremely low (23:00 − 06:00, please refer to Fig. 3 for the demand). However, ATSC-MORL is able to reduce the delay when the traffic demand is medium to high. This is completely opposite to the benchmark controller. It is not supersizing as the goal of the RL-based ATSC is optimizing the total delay throughout the day; therefore, increasing the delay of a small number of vehicles while reducing the delay of a large number of increases the average number of stops per vehicle when the traffic demand is low.

Second, the ATSC-MORL failed to reduce the crash likelihood when the volume is close to zero (01:30 − 04:30). While admittedly, its objective is optimizing the risk score throughout the day, the causation needs to be further investigated.

In conclusion, the proposed ATSC-MORL based ATSC is bale to improve both traffic efficiency and safety compared with the existing field controller. As traffic safety and efficiency are likely to be competing objectives, if the ATSC does not consider traffic safety, it might lead to potential safety issues.

6. Discussion

6.1. Control policy analysis: opening the "Black Box"

Machine learning algorithms are criticized for their lack of interpretability, which is often referred to as the "Black Box" metaphor. While the vast majority of studies on RL-based ATSC showed their superior performance than traditional signal controllers, little attention has been given to illustrate how they achieve it. We would like to open the "Black Box" by analyzing the "optimal policy" of RL-based controllers on the test intersection. Especially there exist certain conditions that RL-based ATSC performs worse than the benchmark in terms of traffic safety. The analysis might not be comprehensive, but rather provides some insights for researchers and practitioners.

Several terms were defined to help control policy analysis:

Signal group: The set of turning movements that are controlled by the same traffic signal indications. For example, in this study, the northbound through movement and the northbound right-turning movement are controlled by the same set of signal indications. These two turning movements belong to the same signal group NT. Each phase could have a set of non-conflicting signal groups.

Signal group green interval length: The length of the time interval that the indication of a signal group is green (short for length in Table 4)

Green ratio: Ratio of the total signal group green interval length within a specific time interval (e.g. 15 min).

Table 4 presents the average length, the average green ratio, and the activated times of each turning movement. The daily traffic flow per lane is also provided as a reference. Fig. 9 illustrates the average 15-minutes-aggregated green ratio for each signal group. In terms of efficiency, RL-based ATSCs (ATSC-SORL and ATSC-MORL) which have better performance exhibits shorter green intervals. In other words, they change the phase more frequently. For an isolated intersection, given that the queue is cleared, shorter green intervals reduce the waiting times of vehicles on approaches whose signal indications are red. This might lead to the delay reduction of the RL-based ATSCs.

As for traffic safety, ATSC-MORL and BC whose risk score are less than that of ATSC-SORL favor the major approach through movement (NT−ST). More specifically, for ATSC-MORL, the green ratio of the NT signal group (with the highest flow) is significantly larger while the green ratio of the SL signal group (with the lowest flow) is significantly smaller. One possible explanation is that such a policy reduces the

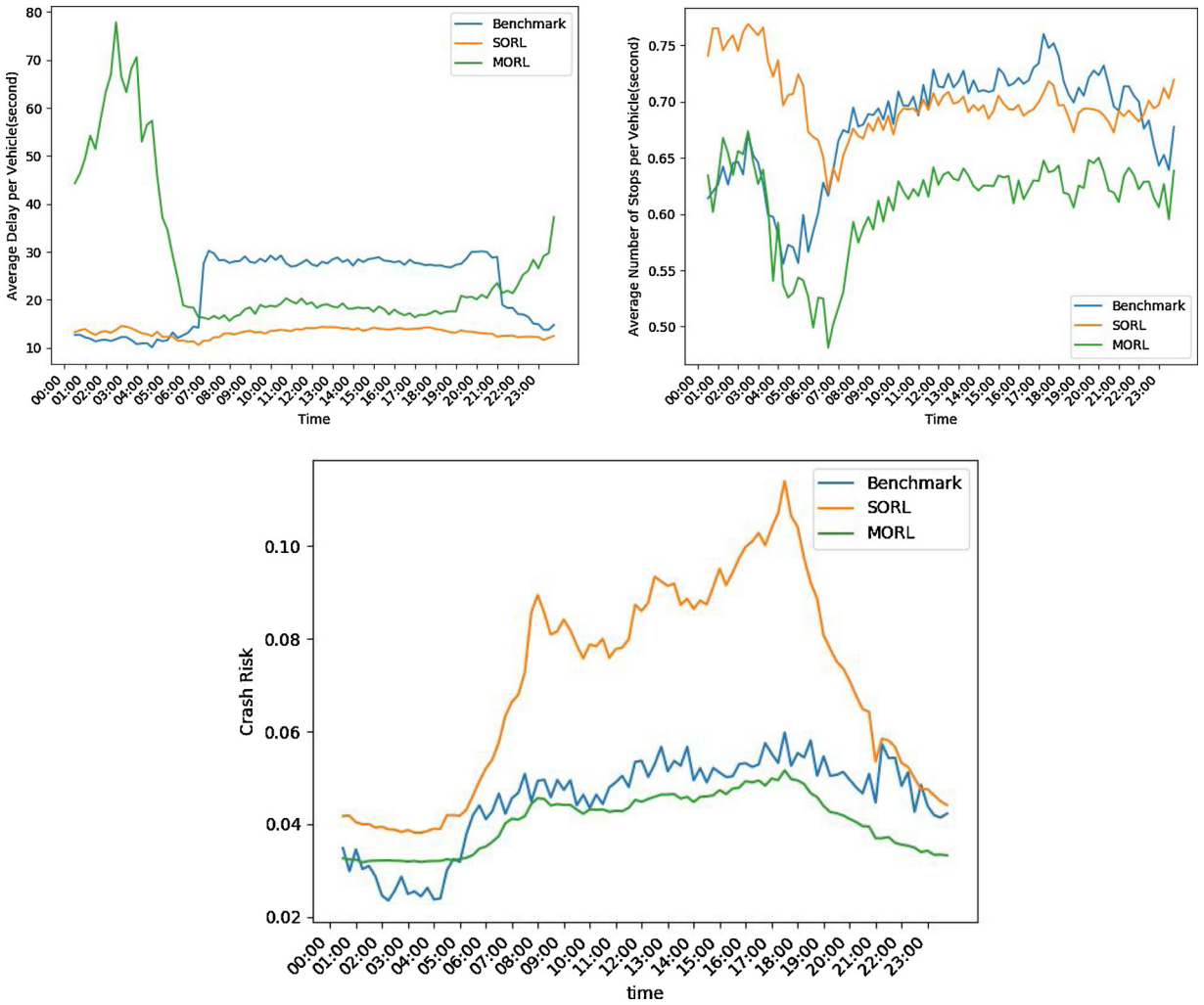

Fig. 8. 15-minutes aggregated performance measures of the controllers (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 4
Statistics of Green Interval Length, Activated Times and Green Ratio of Signal Groups.

Controller	NL				NT			
	Length	Activated Times	Green Ratio	Flow (pdpl)	Length	Activated Times	Green Ratio	Flow (pdpl)
BC	9.55	702	7.8%	3379	45.30	1153	56.8%	6431
ATSC-SORL	8.49	1443	12.4%		17.48	2024	40.9%	
ATSC-MORL	6.39	1133	8.2%		37.02	1662	70.5%	
Controller	SL				ST			
	Length	Activated Times	Green Ratio	Flow (pdpl)	Length	Activated Times	Green Ratio	Flow (pdpl)
BC	7.15	248	2.2%	340	29.68	1272	41.6%	3251
ATSC-SORL	11.63	427	6.8%		7.94	2442	20.9%	
ATSC-MORL	6.41	136	1.2%		25.22	1881	52.1%	
Controller	EL				ET			
	Length	Activated Times	Green Ratio	Flow (pdpl)	Length	Activated Times	Green Ratio	Flow (pdpl)
BC	10.41	655	9.0%	632	9.91	912	10.2%	1614
ATSC-SORL	8.21	1137	9.0%		7.90	1178	8.5%	
ATSC-MORL	7.12	730	6.6%		6.27	851	6.5%	
Controller	WL				WT			
	Length	Activated Times	Green Ratio	Flow (pdpl)	Length	Activated Times	Green Ratio	Flow (pdpl)
BC	8.49	521	5.5%	532	8.66	763	7.6%	787
ATSC-SORL	10.38	1567	17.6%		10.26	1540	17.7%	
ATSC-MORL	7.20	725	7.6%		6.96	752	6.5%	

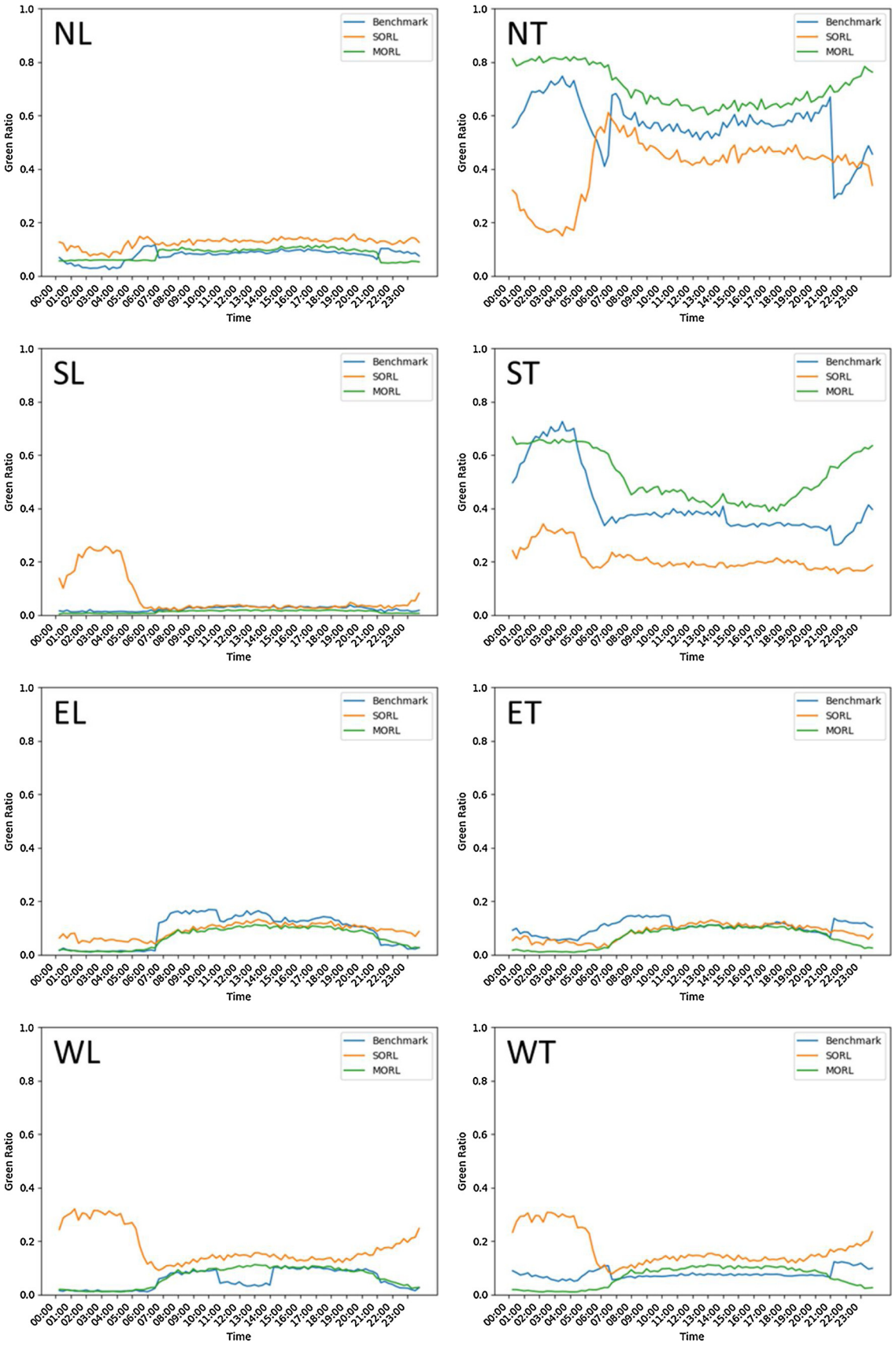

Fig. 9. 15-minutes aggregated green ratios of the controllers (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 5
Average Daily Performance of the Adjusted Controller.

Controller	Efficiency		Safety (Risk Score)
	Average Delay (sec)	Number of Stops	
BC	26.395	0.703	0.045
ATSC-MORL	19.550	0.615	0.041
AD	31.423	0.648	0.039

probability of the occurrence of conflicts with the least sacrificing of traffic efficiency. Consider an extreme case. If the southbound left-turning is completely prohibited, the conflict points with northbound through movement and east-west approaches are eliminated. Since the flow of southbound left-turning is lowest among eight signal groups, prohibiting it would not lead to a huge increase of delay. However, a legal turning movement could not be prohibited if there exists demand, therefore, a compromise is achieved by reducing the activation of the SL signal group. The assumption could also reveal the reason why the ATSC-MORL performs worse when the volume is close to zero. In such conditions, the green ratio of ATSC-MORL is actually less than that of the benchmark.

To support the assumption, a hypothetical coordinated actuated signal controller (AD) is created by adjusting the benchmark controller. From 6:30 to 21:00, the length of SL phase was reduced to the minimum green time (six seconds) while the length of NT phase was increased accordingly. Table 5 shows the average daily performance of the AD controller during 30 test simulated days. It is not surprising that the average delay is higher than the BC as the southbound left-turning movement was intentionally delayed without the remedy of ATSC. This might also imply that traffic safety and efficiency are competing objectives. The risk score is significantly less than the BC controller and even a little bit less than the ATSC-MORL controller. The green ratio of SL signal group of AD controller is actually 1.1 %, which is less than that of ATSC-MORL controller. Therefore, the aforementioned assumption could be regarded as the abstracted knowledge learned by the RL-agent and it could be transferable to a different type of signal controller.

The result of the control policy analysis of simulated RL-based ATSCs could also serve as a reference for improving the existing signal control system if the infrastructure in the field is not ready to adopt RL-based ATSCs or the practitioners are concerned with their acyclic nature. However, as the proposed RL-based ATSCs are site-specific, if it is transferred to other locations, it is recommended to re-train it in a traffic simulation that replicates local traffic conditions.

6.2. Other considerations of the signal settings

ATSCs improve traffic efficiency and/or safety by dynamically adjusting the signal timings. However, other signal settings beyond the timing parameters are also known to have impact on either efficiency or safety. Two tests were conducted to investigate how these factors influence traffic efficiency and/or safety.

6.2.1. Coordinated versus non-coordinated

As there are no universally accepted rules to select the benchmarking controller, this study chooses to use a signal controller that replicates the field one. One might find that the field controller is designed for coordination, yet this study focuses on an isolated intersection. Therefore, the performance of another hypothetical controller (HAC) that runs fully actuated throughout the day was compared with BC and ATSC-MORL (Table 6). The timing of the HAC was set as the same as BC when it runs fully actuated to avoid reevaluating safety-related timing parameters (such as minimum green time and passage time). According to Table 6, the performance of ATSC-MORL is better than HAC in terms of both traffic efficiency and safety, which further confirms the superiority of ATSC-MORL. Compared with coordinated BC, HAC reduces delay yet increases the number of stops. It is expected as the objective of coordination is to reduce the stops of the major approach through movement.

Table 6
Average Daily Performance of the Other Tested Controllers.

Controller	Efficiency		Safety (Risk Score)
	Average Delay (sec)	Number of Stops	
BC	26.395	0.703	0.045
ATSC-MORL	19.550	0.615	0.041
HAC	20.045	0.729	0.051
ATSC-MORLP	28.068	0.672	0.040[a]

[a] A paired T-test was conducted using the data of 30 simulated days to investigate whether the average risk scores are statistically significantly different between ATSC-MORL controlled scenarios and ATSC-MORLP controlled scenarios. The results showed that the difference is statistically significant at 0.0001 level.

6.2.2. Permissive versus protected left-turn

It is well known that the protected left-turning is safer than permissive/permissive-protected left-turning yet is more detrimental to traffic efficiency. An interesting test scenario was developed to investigate the outcome if the permissive-protected left-turning of the east-west approach was changed to protected. Another multi-objective RL-agent (ATSC-MORLP) was trained under such conditions. According to Table 6, ATSC-MORLP created excessive delay as it prohibits permissive left-tuning. However, it did reduce the risk score comparing with ATSC-MORL (see also Fig. 10), especially when the risk score is relatively high (from 15:00 − 19:00). This might imply that if the current crash risk is high, prohibiting permissive left-turning temporarily might be a potential solution.

6.3. Hybrid controller: a better solution

For any practical problems, there is always a trade-off between computational efficiency and algorithm's performance. While the weighted sum approach used in this study is computationally inexpensive, it is not guaranteed to be Pareto-optimal (Vamplew et al., 2008). Specifically, the ATSC-MORL controller performs worse than the BC controller does when the travel volume is extremely low. Therefore, a hybrid controller that changes its backend algorithm based on the traffic volume might have better performance.

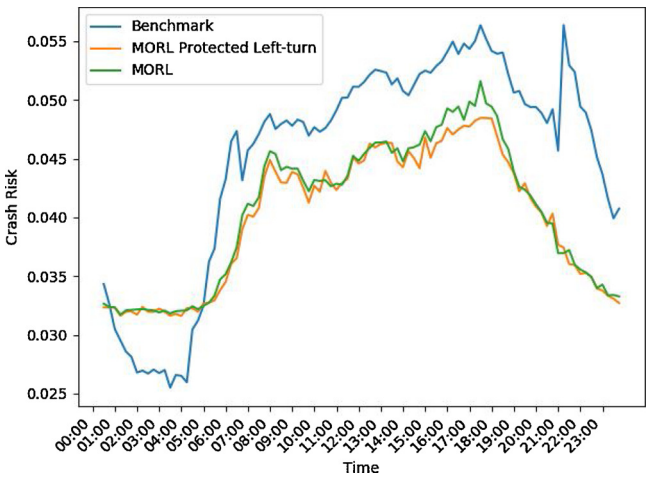

Fig. 10. 15-minutes aggregated risk score of the controllers (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 7
Average Daily Performance of the Hybrid Controller Compared with the Benchmark and ATSC-MORL Controller.

Controller	Efficiency		Safety (Risk Score)
	Average Delay (sec)	Number of Stops	
BC	26.395	0.703	0.045
ATSC-MORL	19.550	0.615	0.041
HS	18.538	0.615[a]	0.040[a]

[a] A paired T-test was conducted using the data of 30 simulated days to investigate whether the number of stops and average safety scores are statistically significantly different between ATSC-MORL controlled scenarios and HS controlled scenarios. The results showed that the difference of number of the number stops is not statistically significant but the difference of average safety scores is statistically significant at 0.0001 level.

A simple hybrid controller (HS) was proposed based on the local condition to test the feasibility of aforementioned concept. When the sum of 15-minute-flow-rates of all turning movements are below 150 vehicles per hour per lane, HS employs BC algorithm. Otherwise, HS employs ATSC-MORL algorithm. This eventually leads to a time-of-day-plan-like controller. From 0:00 to 5:00, BC is activated while ATSC-MORL is activated from 5:00 – 24:00. Table 7 shows the performance of HS controller compared with BC and ATSC-MORL controller. The HS controller slightly reduces the delay by 5.1 % and reduces the average risk score by 2.5 % compared with ATSC-MORL. The 15-minutes performance curve is not provided as the performance curve of HS is identical to the activated backend algorithm.

Other types of hybrid controller such as the hybrid of ATSC-MORL and ATSC-MORLP could also be employed to improve traffic safety if necessary.

7. Summary and conclusions

To improve the traffic safety of the signalized intersection, this study proposes a safety-oriented adaptive signal control algorithm to simultaneously optimize traffic efficiency and safety. The control agent takes high-resolution real-time traffic data as its input and selects appropriate signal phases every second to reduce vehicles' delay and the crash risk of the intersection. A multi-objective reinforcement learning framework using double dueling deep neural network is utilized as the backend algorithm to solve the discrete optimization problem. The weighted sum approach, one of the single policy multi-objective reinforcement learning algorithms, is employed to deal with the trade-off between traffic safety and efficiency.

The proposed algorithm was trained and evaluated in a simulated isolated intersection in Seminole County, Florida, built based on field observed traffic data. A real-time crash prediction model is calibrated using local crash data to provide the crash risk in the near future. The performance of the well-trained algorithm was evaluated by the real-world signal timings provided by the local jurisdiction. The evaluation results showed that the algorithm improves both traffic efficiency and safety compared with the benchmark. In addition, compared with an adaptive traffic signal optimizing only traffic efficiency, it did improve traffic safety significantly but with a slight deterioration of traffic efficiency. This might imply the traffic safety and efficiency are two competing objectives. Practitioners should take the trade-off into consideration.

A brief analysis of control policies of different signal controller reveals how the RL-based ATSCs are able to improve traffic efficiency and safety. The abstracted control rules from the analysis could serve as a reference for improving existing signal control systems if the infrastructure in the field is not ready to adopt RL-based ATSCs or the practitioners were concerned with their acyclic nature. However, as the proposed RL-based ATSCs are site-specific, it is recommended to train the RL-based ATSC in a traffic simulation that replicates the local condition. A hybrid controller that changes its backend algorithm based on traffic volume is also proposed to improve the performance of MORL controlling algorithm if the well-trained MORL is not Pareto-optimal.

Admittedly, there are several limitations. As the weighted sum approach is not guaranteed to be Pareto-optimal, the study could be improved by calculating the Pareto-front using more computationally efficient algorithms. Meanwhile, other kinds of safety measures such as traffic conflicts could be tested as the safety objective using the proposed algorithm. Moreover, as vehicles' operation speeds are correlated with both efficiency and safety, controlling vehicles' speed directly may provide additional safety and operational benefits (Li et al., 2018; Ma et al., 2017; Qu et al., 2020; Zhou et al., 2020). With the rapid development of the connected and automated vehicles (CAV), a safety-oriented control system that jointly controls of traffic signals and CAV would be a valuable future research direction.

CRediT authorship contribution statement

Yaobang Gong: Conceptualization, Methodology, Software, Investigation, Visualization, Writing - original draft. **Mohamed Abdel-Aty:** Conceptualization, Supervision, Writing - review & editing. **Jinghui Yuan:** Methodology, Software. **Qing Cai:** Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors also appreciate the data provided by FDOT and Seminole County.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jozefowicz, R., Jia, Y., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Schuster, M., Monga, R., Moore, S., Murray, D., Olah, C., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, M.F., Hsia, L., 2004. Predicting freeway crashes from loop detector data by matched case-control logistic regression. Transp. Res. Rec. 1897 (1), 88–95. https://doi.org/10.3141/1897-12.

Abdel-Aty, M., Dilmore, J., Dhindsa, A., 2006. Evaluation of variable speed limits for real-time freeway safety improvement. Accid. Anal. Prev. 38 (2), 335–345. https://doi.org/10.1016/j.aap.2005.10.010.

Arroyo, V.A., Bennett, S.E., Butler, D.H., Dougherty, M., Stewart Fotheringham, A., Halikowski, J.S., Dot, A., Michael Hancock, P.W., Hanson, S., Heminger, S., Hendrickson, C.T., Knatz, G., Osterberg, D.A., Rosenbloom, S., Schwartz, H.G., Sinha, K.C., Steudle, K.T., Dot, M., Gary Thomas, L.C., Bostick, T.P., Wallerstein, B.R., Winfree, G.D., Wright, F.G., Director, E., Zukunft, P.F., 2015. NCHRP Report 812 – Signal Timing Manual. second edition. Nchrp.

Fink, J., Kwigizile, V., Oh, J.-S., 2016. Quantifying the impact of adaptive traffic control systems on crash frequency and severity: evidence from Oakland County, Michigan. J. Safety Res. 57, 1–7. https://doi.org/10.1016/J.JSR.2016.01.001.

Gong, Y., Abdel-Aty, M., Cai, Q., Rahman, M.S., 2019. Decentralized network level adaptive signal control by multi-agent deep reinforcement learning. Transp. Res. Interdiscip. Perspect. 1, 100020. https://doi.org/10.1016/j.trip.2019.100020.

Houli, D., Zhiheng, L., Yi, Z., 2010. Multiobjective reinforcement learning for traffic signal control using vehicular ad hoc network. EURASIP J. Adv. Signal Process. 2010, 7. https://doi.org/10.1155/2010/724035. 1–7:7.

Jin, J., Ma, X., 2015. Adaptive group-based signal control by reinforcement learning. Transp. Res. Procedia 10, 207–216. https://doi.org/10.1016/j.trpro.2015.09.070.

Karlsson, J., 1997. Learning to Solve Multiple Goals. University of Rochester.

Khamis, M.A., Gomaa, W., 2014. Adaptive multi-objective reinforcement learning with hybrid exploration for traffic signal control based on cooperative multi-agent framework. Eng. Appl. Artif. Intell. 29, 134–151.

Khattak, Z.H., Magalotti, M.J., Fontaine, M.D., 2018. Estimating safety effects of adaptive signal control technology using the empirical Bayes method. J. Safety Res. 64,

121–128. https://doi.org/10.1016/J.JSR.2017.12.016.
Li, X., Ghiasi, A., Xu, Z., Qu, X., 2018. A piecewise trajectory optimization model for connected automated vehicles: exact optimization algorithm and queue propagation analysis. Transp. Res. Part B Methodol. 118, 429–456. https://doi.org/10.1016/j.trb.2018.11.002.
Liu, C., Xu, X., Hu, D., 2015. Multiobjective reinforcement learning: a comprehensive overview. IEEE Trans. Syst. Man Cybern. Syst. 45 (3), 385–398. https://doi.org/10.1109/TSMC.2014.2358639.
Ma, J., Fontaine, M.D., Zhou, F., Hu, J., Hale, D.K., Clements, M.O., 2016. Estimation of crash modification factors for an adaptive traffic-signal control system. J. Transp. Eng. 142 (12), 4016061. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000890.
Ma, J., Li, X., Zhou, F., Hu, J., Park, B.B., 2017. Parsimonious shooting heuristic for trajectory design of connected automated traffic part II: computational issues and optimization. Transp. Res. Part B Methodol. 95, 421–441. https://doi.org/10.1016/j.trb.2016.06.010.
Muresan, M., Fu, L., Pan, G., 2018. Adaptive traffic Signal control with deep reinforcement learning – an exploratory investigation. In: 97th Annual Meeting of the Transportation Research Board. Washington, D.C..
Qu, X., Yu, Y., Zhou, M., Lin, C.-T., Wang, X., 2020. Jointly dampening traffic oscillations and improving energy consumption with electric, connected and automated vehicles: a reinforcement learning based approach. Appl. Energy 257, 114030. https://doi.org/10.1016/j.apenergy.2019.114030.
Sabra, Z.A., Gettman, D., Nallamothu, Venkata, Pecker, C., 2013. Enhancing Safety and Capacity in an Adaptive Signal Control System (Phase 2). Sabra, Wang & Associates, Inc.https://doi.org/10.13140/RG.2.2.16217.83044.
Stevanovic, A., Stevanovic, J., Kergaye, C., 2013. Optimization of traffic signal timings based on surrogate measures of safety. Transp. Res. Part C Emerg. Technol. 32, 159–178. https://doi.org/10.1016/J.TRC.2013.02.009.
Stevanovic, A., Stevanovic, J., So, J., Ostojic, M., 2015. Multi-criteria optimization of traffic signals: mobility, safety, and environment. Transp. Res. Part C Emerg. Technol. 55, 46–68. https://doi.org/10.1016/j.trc.2015.03.013.
Sutton, R.S., Barto, A.G., 2018. Reinforcement Learning: an Introduction. MIT press.
Tageldin, A., Sayed, T., Zaki, M.H., Azab, M., 2014. A safety evaluation of an Adaptive Traffic Signal Control system using computer vision. Adv. Transp. Stud. Special 2, 83–98.
Vamplew, P., Yearwood, J., Dazeley, R., Berry, A., 2008. In: Wobcke, W., Zhang, M. (Eds.), On the Limitations of Scalarisation for Multi-Objective Reinforcement Learning of Pareto Fronts BT - AI 2008: Advances in Artificial Intelligence. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 372–378.
Van Der Pol, E., Oliehoek, F.A., 2016. Coordinated deep reinforcement learners for traffic light control. NIPS'16 Work. Learn. Inference Control Multi-Agent Syst.
Vidhate, D.A., Kulkarni, P., 2017. Cooperative multi-agent reinforcement learning models (CMRLM) for intelligent traffic control. 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM) 325–331. https://doi.org/10.1109/ICISIM.2017.8122193.
Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., De Freitas, N., 2015. Dueling network architectures for deep reinforcement learning. arXiv Prepr arXiv1511.06581.
Wang, L., Abdel-Aty, M., Lee, J., 2017. Implementation of active traffic management strategies for safety on congested expressway weaving segments. Transp. Res. Rec. 2635 (1), 28–35. https://doi.org/10.3141/2635-04.
Yau, K.-L.A., Qadir, J., Khoo, H.L., Ling, M.H., Komisarczuk, P., 2017. A survey on reinforcement learning models and algorithms for traffic signal control. ACM Comput. Surv. 50 (3). https://doi.org/10.1145/3068287. 34:1–34:38.
Yu, R., Abdel-Aty, M., 2014. An optimal variable speed limits system to ameliorate traffic safety risk. Transp. Res. Part C Emerg. Technol. 46, 235–246. https://doi.org/10.1016/j.trc.2014.05.016.
Yuan, J., Abdel-aty, M., 2018a. Approach-level real-time crash risk analysis for signalized intersections. Accid. Anal. Prev. 119 (April), 274–289. https://doi.org/10.1016/j.aap.2018.07.031.
Yuan, J., Abdel-Aty, M., 2018b. Approach-level real-time crash risk analysis for signalized intersections. Accid. Anal. Prev. 119, 274–289. https://doi.org/10.1016/j.aap.2018.07.031.
Yuan, J., Abdel-aty, M., Wang, L., Lee, J., Yu, R., Wang, X., 2018. Utilizing bluetooth and adaptive signal control data for real-time safety analysis on urban arterials. Transp. Res. (97 October), 114–127. https://doi.org/10.1016/j.trc.2018.10.009. Part C.
Yuan, J., Abdel-Aty, M., Gong, Y., Cai, Q., 2019. Real-Time crash risk prediction using long short-term memory recurrent neural network. Transp. Res. Rec. 2673 (4), 314–326. https://doi.org/10.1177/0361198119840611.
Zhou, M., Yu, Y., Qu, X., 2020. Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: a reinforcement learning approach. IEEE trans. Intell. Transp. Syst. 21 (1), 433–443. https://doi.org/10.1109/TITS.2019.2942014.
Zhu, L., Li, K., Liu, Z., Wang, F., Tang, K., 2019. A group-based signal timing optimization model considering safety for signalized intersections with mixed traffic flows. J. Adv. Transp. 2019, 1–13. https://doi.org/10.1155/2019/2747569.