運輸計劃季刊
第五十五卷 第二期
民國一一五年六月
頁 167～頁 192

Transportation Planning Journal
Vo1. 55 No. 2
June 2026
PP. 167～192

強化學習框架應用於高速公路都會瓶頸壅塞區匝道控制

RAMP METERING CONTROL FOR URBAN BOTTLENECK CONGESTED EXPRESSWAY AREAS VIA REINFORCEMENT LEARNING FRAMEWORK

嚴國基 Kuo-Chi Yen[1]
施武樵 Wu-Chiao Shih[2]
賀盛志 Sheng-Chih Ho[3]
趙世祺 Shih-Chi Chao[4]

(111 年 11 月 24 日收稿，112 年 6 月 27 日第一次修正，112 年 12 月 29 日第二次修正，113 年 5 月 9 日第三次修正，113 年 9 月 13 日接受)

摘 要

高速公路於臺北都會區瓶頸路段在通勤時間的壅塞狀況益趨嚴峻。為驗證人工智慧技術之應用，本研究結合深度強化學習方法與符合實際車流特性的模擬環境進行互動。利用妥適的獎勵函數，刺激號誌代理人於學習過程中，面對變動的環境狀態選取適當動作 (調整匝道號誌時間) 以追求累積折扣獎勵最大。訓練後的號誌代理人，利用號誌時間與動作轉換函數，以多代理人操作的特性達成控制多匝道號誌的目標。同時亦藉一般化測試，充分說明本文提出之框架方法確實可在高速公路都會區瓶頸壅塞環境進行號誌控制，深具未來發展潛力。

1. 國防大學運籌管理學系副教授。
2. 國防大學運籌管理學系助理教授 (聯絡地址：11258 臺北市北投區中央北路二段 70 號，電話：03-3801126 轉 604889，E-mail: yenkuochi0706@gmail.com)。
3. 德明財經科技大學資訊管理學系助理教授。
4. 國防大學運籌管理學系碩士。

關鍵詞： 瓶頸路段壅塞、匝道號誌、強化學習

ABSTRACT

The recurrent congestion of expressway sections in Taipei metropolitan area during commuting time is becoming more serious. It is proposed that this study uses a predefined signal program as the action space of an intelligent signal agent and combines a deep reinforcement learning algorithm to interact with a simulated environment that matches the actual traffic flow characteristics. To achieve the maximum cumulative discount reward expectation, the signal agents adjust the ramp signal time depending on the environmental state at different times during the learning process under the action of the reward function with different emphasis directions. By using the action transfer function and signal time, our signal agents achieve the goal of simultaneous control of multi-ramp signals while maintaining the characteristics of multi-agent. Through the general test of random traffic flow, it is shown that the artificial intelligence method proposed in this paper has great potential for the future because it can alleviate the problem of bottleneck sections of expressways in urban areas.

Key Words: *Bottleneck Congestion ; Ramp Signal ; Deep Reinforcement Learning*

一、前　言

隨著都市化開發、人口集中，行經都會區之高速公路不僅具備城際運輸的功能亦擔負起區域運輸的角色。這同時也衍生了日趨惡化的道路壅塞、環境汙染及能源消耗等問題。高速公路行經都會區之交通瓶頸路段，常透過交通控制策略予以改善，而匝道控制即為主要因應策略之一。由於匝道控制能夠縮限主線車流量，使車流量不至超過其容量，提高主線運轉效率及行車安全。因此，高速公路局 (以下簡稱高公局) 以匝道封閉與精進式匝道儀控方式調節進入高速公路之車流量。高公局藉由上述方法於長假日期間，在高速公路各嚴重壅塞路段實施管制。雖然 2021 年平均交通量為 110.2 百萬車公里 (mvk)，較 2020 年和平紀念日連假增加 1%。但是，管制策略實施後 2021 年連假整體總壅塞路段與時段數量較 2020 年同期減少 22% (https://www.freeway.gov.tw/Print.aspx?cnid=195&p=22177)。由此可知，匝道控制方法仍為目前高速公路舒緩壅塞現象的有效手段。

匝道號誌控制的方式約可分為 2 類。第一類為定時式預設式 (pre-timed metering)，此類是基於歷史數據建立預設的儀控時制，具有操作簡單、成本低等優點，但是其靈活性較弱，無法針對即時的交通狀態進行調節。第二類為交通感應式 (responsive metering)，此類是根據即時的交通資訊以及相應的演算邏輯進行計算，據以對匝道進入主線的車流進行調

節。其具備靈活性，目前諸多研究均以此類為基礎進行發展，惟其控制績效相當依賴交通狀態預測的準確性。前述兩類方法仍存在一定的局限性，例如，號誌控制的規則與方法需要具備經驗的工程人員介入進行調整，可能影響車流控制的彈性與穩定性 (Wu et al. [1])。隨著人工智慧的發展，部分研究已利用機器學習發展號誌控制的方法。其中，以強化學習 (Reinforcement learning) 為基礎的方法不僅能利用歷史數據，還可以與車流環境互動循環迭代更新模型的參數，將更具潛力。

高公局目前控制匝道及主線的儀控方法，是依據車流回饋資訊，再以動態查表法進行精準式匝道儀控。本研究以高公局 1968 網站上統計之平日經常壅塞路段「國道一號北上三重至圓山間路段」為研究場域，在現有條件與環境導入多代理人強化學習模型的人工智慧方法控制不同入口匝道的號誌進行協同作用，以達到整體控制的目標。為方便測試驗證，本研究以 SUMO (Simulation of Urban MObility) 微觀車流模擬軟體，建構符合具代表性的車流模擬環境，進而建立多代理人強化學習框架 (Multi-Agent Reinforcement Learning Framework, MARLF)，以完全合作方式(fully-cooperative)改善研究場域之壅塞問題。

本文組成如下：首先回顧高速公路壅塞問題與匝道控制策略；其次說明深度強化學習理論及模擬環境的建立與校驗；接著建立符合仿真條件的模擬環境，並使其與代理人之間進行互動與學習。其中將針對強化學習中優化器、網路結構以及獎勵函數進行設定，以取得較優的學習效率；再以完成訓練的多代理人以實際數值為基礎進行模型框架一般化能力驗證；最後依研究結果提出具體的結論與建議。

二、文獻回顧

2.1 匝道號誌控制

國內外常用的定時匝道儀控，主要依據匝道歷史車流資料計算出不同時段之最佳儀控率，現場設定後即可離線執行。現行高公局則採透過大數據技術結合車流理論的「精進式匝道儀控」。其原理是針對電子收費 (ETC) 車流資料，分析壅塞路段交通量來自各交流道的比例，針對交通量比例較高的交流道實施較嚴格的匝道儀控，再依期望通過壅塞路段的交通量比例調整各匝道儀控時制。然而，精進式匝道儀控仍是以查表方式為基礎仍係定時儀控，非屬人工智慧或完全適應式的控制方法。

國內外在獨立匝道儀控方面，多以交通反應式策略為主。其中包含需求容量式策略、佔有率策略 (Masher et al. [2])，與 ALINEA (Asservissement Linéaire d'Entrée Autoroutière)。其中，需求容量式策略與 佔有率策略皆為前饋型 (feedforward) 架構；ALINEA 策略與其後續發展則為回饋型 (feedback) 架構，除目標設定與量測標的與前饋型相異外，各時間所計算之值均將回饋至下個時間，以做為每次運算之依據(Papageorgiou et al. [3])。在區域匝道儀控方面，Wattleworth [4] 運用線性規劃求解最佳儀控率，目的在符合主線容量限制下求進入主線車流的車輛數最大。然而，線性規劃法僅考量進入流量與其旅次分配，未將車流

行進行為納入考慮，亦未考量下游主線路段之流量總和。Papageorgiou [5] 則在時間因素下，同時考量匝道上游主線流量與匝道流量間關係，及匝道區域的時間延滯及在匝道下游主線的擾動問題，提出以維持匝道下游主線容量為基礎的線性規劃模式，並提出分時控制演算法。該模式可依據旅行時間之長短，計算匝道流量到達主線下游各時段之時間，再加總同一路段與時段之車流量後求解最佳儀控率。該研究成果證明所提出的方法能有效舒緩高速公路在尖峰時段出現重複性壅塞的現象。至於在適應式交通系統 (adaptive traffic system) 方面，根據世界道路協會網頁 (https://www.piarc.org/en/) 顯示，倫敦附近的 M25 高速公路，以類似的系統將壅塞的形成時間推遲 10-20 分鐘，並改善了車道間的交通分佈狀態。此外，亦有利用模糊理論控制策略(Meldrum and Taylor [6])、類神經網路方法 (Zhang and Ritchie [7]) 以及分區演算法 (Stephanedes [8]) 等方法進行匝道儀控。

2.2 多代理人強化學習

多代理人強化學習 (Multi-Agent Reinforcement Learning, MARL) 是透過多代理人與環境的互動，在獎勵值的激勵下，持續修正決策並調整動作以獲取最佳的解決方案(Gronauer and Diepold [9])。相關研究領域廣及機械手臂、無人車等，似可將其方法應用於舒緩高速公路壅塞問題 (Yu and Stubberad [10])。強化學習主要包含兩部分，其一為以馬可夫決策過程 (Markov Decision Processes, MDP) 為基礎的強化學習演算法；其二則為由神經網路所建構而成的機器學習技術。MDP 最直接的目的是透過不斷的試誤過程 (trial and error) 來對陌生的事物進行學習。以本研究的區域「國道一號北上三重至圓山間路段」為例，包含三重、臺北、圓山等入口匝道號誌，由於各號誌分屬不同代理人目標各異，代理人之間如果有一個所採取的策略較差，將會導致整體任務的失敗。因此，各代理人應追求全局最大化為目標的完全合作關係。

代理人在環境中學習、決策並採取不盡相同的動作 (action, A^i)，在此 A^i 為第 i 個代理人可能的動作為隨機變數，如果將動作落實為實際值則記為 a^i。環境則會因代理人所採取的動作而改變其狀態 (state, S) 亦為隨機變數，而 s=$[o^1, o^2, \dots, o^m]$ 則表示為狀態值，由代理人所在處的觀測 o^i 組成。當狀態值確定後，則環境將回饋代理人獎勵值 r^i。不同的代理人將彙整狀態、採取的動作、動作後的狀態 (state', s_{t+1}) 進行學習。各代理人在目標一致之下 (如式(1))，亦需改進屬於個別代理人的策略網路。網路中的參數θ^i的更新將會增加價值函數的期望值，促使目標 J 極大化，如下示式 (2)：

$$J(\theta^1, \dots, \theta^m) = \mathrm{E}_S[V_\pi(S)] \tag{1}$$

$$\max_{\theta_1,\dots,\theta_m} J(\theta^1, \dots, \theta^m) \tag{2}$$

式 (1) 中，$V_\pi(S)$為狀態價值函數，其意義是評估系統狀態的優劣，以第 i 個代理人的 $V_\pi^i(s_t)$ 為例，是對於狀態 s_t 進行價值評估，亦是關於隨機變數動作 A_t 的期望。而動作價值函數 $Q_\pi^i(s_t, a_t)$ 是已知狀態與與動作下代理人的累計折扣獎勵，其意義是在評估動作的

優劣價值。有關 Q_π^i、V_π^i 如下示式 (3)、(4) 所定義。

$$V_\pi^i(s_t) = \mathrm{E}_A\left[Q_\pi^i(s_t, A_t)\right] \tag{3}$$

$$Q_\pi^i(s_t, a_t) = \mathrm{E}(U_t^i | S_t = s_t, A_t = a_t) \tag{4}$$

在式 (3)、(4)中可知道，代理人 i 的 Q_π^i 與 V_π^i 均依賴於其他所有代理人的策略。即便代理人 i 單獨改變自己的策略，也未必能使得自己的 V_π^i 與 Q_π^i 變得更好。在式 (4) 中的 U_t^i 稱為累積折扣獎勵 (discounted return) 它的定義是第 i 個代理人不同時間獎勵的加權和，如下示式 (5)：

$$U_t^i = R_t^i + \gamma \cdot R_{t+1}^i + \gamma \cdot R_{t+2}^i + \gamma \cdot R_{t+3}^i + \dots \tag{5}$$

式 (5) 中的 $\gamma \in [0,1]$ 為折扣率 (discount factor)，代表未來行動價值函數的重要性。亦可視之為代理人是否有足夠的 「遠見」，以預見未來的情況並隨之調整行動價值函數。當值愈大，代表代理人 愈重視後續的發展；反之則以當前的狀況為主。有關多代理人強化學習的網路與訓練之說明概述如下：

(1) 價值網路：其目的在近似狀態價值函數，本研究的代理人屬完全合作關係，所有代理人共有一個價值網路，在實際計算 V_π^i 時則以一個神經網路 $v(s; w^i)$ 進行近似。網路的輸入是所有代理人的觀測，綜合後成為狀態，如 $s = [o^{11}, \dots, o^{ij}]$，網路學習的輸出 $v(s; w)$ 為一實數，是對於狀態s的評分。

(2) 目標網路：其網路結構、輸入、輸出數量，與價值網路相同，其目的在於緩解價值網路在自舉 (bootstrap) 過程中所造成的偏差，其網路參數 w^-與價值網路 w 進行加權平均後，再進行更新。

(3) 策略網路：每個代理人均有專屬的策略網路，網路結構相同，但是網路參數 $\theta^1, \dots, \theta^m$不同，網路的輸入是各自隸屬代理人的不完全觀測，$s = [o^{11}, \dots, o^{ij}]$，網路學習的輸出將近似策略函數 $\pi(a|s; \theta^i)$，為一個向量，係各代理人離散動作空間上的機率分布。

(4) 價值網路與策略網路的意義：訓練策略網路相當於訓練第 i 個匝道代理人 (actor) 基於狀態做出動作，而價值網路 $v(s; \mathbf{w})$ 則相當於評審 (critic)，對於代理人表現進行量化評分。

(5) 網路訓練：價值網路是利用時序差分學習 (Temporal difference learning，TD learning) 進行優化，透過對模擬環境進行取樣 (觀測的 s_t, s_{t+1}與獎勵 r_t) 由當前價值函數的估計值中自舉的方式進行學習，更新網路參數w^i。目標網路則需藉由價值網路進行更新參數以優化輸出值，而價值網路與目標網路的輸出值可計算 TD 誤差，藉由誤差的減少則會使得價值網路的輸出不斷接近目標網路輸出。在策略網路的訓練則是利用觀測的狀態、動作、獎勵以及近似策略梯度更新網路參數θ^i。再藉由θ^i優化輸出近似 $\pi(a|s; \theta^i)$，能有更高的機率使得代理人選出最適合環境的行動。

多代理人強化學習藉由持續與環境的互動來進行學習、更新狀態並採取動作的特性，與高速公路匝道儀控的狀況類似，即透過觀察即時車流環境狀態及事先預設的號誌秒數來

調整儀控，再依據號誌預設邏輯應用於實際環境的結果，漸次修正以達到滿足高速公路車流的目標。

2.3 交通模擬與模型校驗

交通管理政策攸關民眾生活，由於政策稍有誤差將影響所涉及的不同層面，因此在新的政策推行前，務須經過再三的研究與驗證，以確保所訂策略是安全、有效率且符合民眾通行需求，而驗證交通管理政策，最經濟與直接的方式，莫過於運用交通模擬軟體來進行。美國聯邦公路總署 (Federal HighWay Administration, FHWA) 依據模擬對象的細緻程度，將交通模擬軟體區分為巨觀、微觀以及中觀三種類型 (Alexiadis et al. [11])。此外，為了使模擬過程貼近真實情況，除了依研究目的選用適當的交通模擬軟體外，還需要交通數據的導入來形成模擬場景。交通數據的種類相當繁多，最主要可分為道路實體資料、交通量資料以及交通控制設施資料等三種 (Lopez et al. [12])。

在交通量的蒐集方面，可藉由架 (埋) 設在車道上的車輛偵測器 (VD)、電子收費系統 (ETC) 或是攝影機即時影像來獲得，資料獲得後經過清整、分類等工作後即可導入模擬模型中產生車流。然而模擬模型所產生的車流是否能夠切實反映真實車流，則有賴於模型的確認、校估與驗證。近年來，受惠於偵測器、資通訊技術及數據科學的持續發展，交通資料蒐集已可由傳統人力計算演變為偵測器執行，資料量也倍增至數以百萬計。然而，如欲透過模擬軟體還原真實交通型態，數據的取得就不應只著重在量，質的部分亦為相當重要。Ozbay et al. [13] 即針對大數據資料應用於交通模擬模型提出數項重點：

1. 模型確認 (Model Verification)：

模型確認是要確定所設計的模型符合設計者的本意。模型需要接近真實的車流行為，亦需詳細蒐集真實數據及參數據以建立。所需資料包含駕駛資料 (年齡族群、旅次目的、感知能力及對實驗場域熟悉程度)、車輛資料 (車流組成情況、車種及車齡)、行人及單車資料等；環境參數則為車道數、道路容量、自由流車速、壅塞密度、速限等；路徑成本 (收費價格及可能替代路徑旅行時間)、交通控制 (號誌秒數、可變標誌及施工區域等)、天候 (雨、雪或霧等)、駕駛行為 (車間距離及車道選擇模式) 以及路徑選擇行為 (駕駛受路徑參數影響產生之行為改變) 等。這些資料及參數，甚而行為模式，均需經過逐一檢查確認，以建立符合設計者本意的模型。

2. 參數校估 (Parameters Calibration)：

參數校估是在輸入符合實際的資料下，經由參數調校，使模擬模型的產出能接近實際系統，是常用來配合模型驗證的方法。模型的產出包含流量、速度、車隊等候長度、軌跡資料、車禍資料以及廢氣排放資料等。產出的數據需合理、符合預設之信賴程度且趨近於實際系統，如果與實際系統有落差，則需再反覆的修正模型參數，直到符合預設目標為止。

3. 模型驗證 (Model Validation)：

模型驗證是要確定所設計模型的運轉特性能符合實際系統。交通模擬可參考美國聯邦

公路總署，以 GEH 值作為衡量標準。GEH 值類似統計中的卡方檢定，為 Geoffrey E. Havers 於 1970 年代所提出的一個經驗公式，如式 (6) 所示。

$$GEH_l^t = \sqrt{\frac{2\times\left(D_{obs,l}^t - D_{sim,l}^t\right)^2}{D_{obs,l}^t + D_{sim,l}^t}} \tag{6}$$

其中，

l　：路段指標

t　：觀測時間

$D_{obs,l}^t$：第 t 時間道路 ETC 或 VD 之觀測值

$D_{sim,l}^t$：模擬環境於第 t 時間產生之數值

目前常見的微觀車流模擬軟體如 PARAMICS (Gordon and Cameron [14]), VISSIM (Fellendorf [15]), AIMSUM (Casas et al. [16]), MATSIM (W Axhausen et al. [17]) 及 SUMO (Lopez et al. [11]) 等數種。前三項屬商用軟體、後二項則為開源軟體。本研究使用德國太空中心開發的 SUMO 軟體。其優點為免費、開源、高度靈活，更可透過 SUMO 的 TraCI (Traffic Control Interface) 工具包擷取並匯出模擬過程的交通數據。經過邏輯判斷或運算，動態調整模擬軟體中的交通參數，例如交通號誌秒數或車道變換機制等，對以人工智慧進行交通管理等研究，助益甚大。

目前已有許多應用 SUMO 模擬進行環境驗證之研究。Geistefeldt et al. [18] 以德國高速公路、具雙車道出口匝道的路段為模擬路網，使用該路段所觀測到車流量之 1.2 倍投入 VISSIM、AIMSUN、PARAMICS、BABSIM 及 SUMO 等五種微觀車流模擬軟體，針對進入高速公路之車流對於主線車流的影響進行模擬，並分析比較不同模擬軟體所產生之臨界流量 (Breakdown Volume) 與服務水準。結果發現前四種模擬軟體模擬結果相似；SUMO 則與其他四者有極大的差異，其所模擬的車輛在鄰近車道車輛間只有最小安全距離下仍會進行超車，導致服務水準較佳，與實際之超車行為邏輯不符，可見 SUMO 0.16.0 版本在跟車及變換車道等車流互動行為的模化仍不符實際。接續在 SUMO 0.21.0 的更新版本則增加 Krauss 跟車模型，Song et al. [19] 的研究發現，SUMO 新增的 Krauss 跟車模型，已考慮汽車在運動中逐漸減速的過程，並在等待時間、行駛時間、換道次數和安全速度方面取得了良好的效果。

由於強化學習需與環境相互作用，代理人們需要在模擬環境中進行有效學習，因此有許多研究 (Kheterpal et al. [20] ; Maiti and Chilukuri [21]) 結合 SUMO 模擬模式來建立深度強化學習模型。Li et al. [22] 即在可變車速限制的研究中廣泛使用 SUMO 模擬環境，以更加多樣化和細緻的設定以校估車輛的各種跟車和變換車道的駕駛微觀模型。本研究亦採用 SUMO 0.21.0 版本進行後續的研究。在模擬環境驗證部分，則是以 VD 偵測資料作為模擬輸入，並藉由模擬輸出值與 ETC 的車流資料，以美國聯邦公路總署建議的 GEH 值進行模型的校準與驗證，使得模擬模型的產出符合實際車流的特性。

三、模擬環境與深度強化學習框架建構

3.1 模擬環境

本研究選定高公局 1968 網站所列管之路段進行研究。該路段起自中山速公路北上 27 公里 (三重北上入口匝道) 起，至北上 23 公里 (圓山-濱江街出口) 止，全長 4 公里，由南向北計通過三重、臺北及圓山等進、出口匝道以及編號 01F0256N (三重-臺北)、01F0233N (臺北-圓山)、01F0213N (圓山-內湖) 等四處 ETC 門架，路段容量反映在不同車道數上，為使模擬過程貼近真實情況，本研究實際勘查研究該路段車道數情況，整理如表 1 所示 (不含路肩)：

表 1　研究路段諸元列表

北上里程數	地理位置	車道數
27~26.5	三重交流道 ~ 臺北交流道 (重慶北路出口) 間路段	4
26.5~25	臺北交流道 (重慶北路出口) 間路段 ~ 臺北交流道 (重慶北路出口) 間路段	4
25	臺北交流道 (重慶北路出口) ~ 臺北交流道 (重慶北路入口)	3
25~23	臺北交流道 (重慶北路入口) ~ 圓山交流道 (松江路、建國北路出口)	3
23	圓山交流道 (松江路、建國北路出口) ~ 主線內湖-濱江出口	2

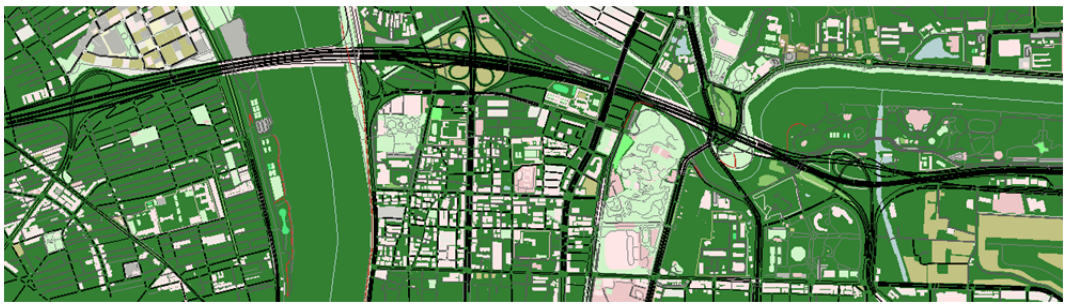
資料來源：本研究整理。

依據前述路網狀態，由 OpenStreetMap 框定研究路段，如下圖 1(a) 所示。另將該地域之電子地圖轉入 SUMO 建構模擬路網 (如下圖 1(b))。有關該路段之模擬環境設定以及參數狀態，均依高速公路局提供現況予以設定。此外，號誌時間則是以實際目前該研究路段於尖峰時區的秒數進行設定 (以上均彙整如表 2)。

本研究統計行經該區域門架的車流量以及 01F0256N ~01F0213N 門架之間的空間平均車速，彙整成研究路段工作日每小時的平均車輛數與平均車速，如圖 2 所示。由圖 2(a)知，自上午 0500 時起，車輛開始增加；0700 時至 1700 時為車流量高峰期，平均每小時流量均大於 2500 (veh/hr) ，且不能明顯區別晨峰與昏峰。自 1800 時，車流量開始遞減。由圖 2(b)知，平均車速自 0600 時開始降速；0800 時至 1200 時及 1500 至 1700 時，平均車速均低於 60 (km/hr)。



(a) OpenStreetMap 地圖路段



(b) SUMO 模擬路段

圖 1　研究路段路網圖

表 2　研究路段各項道路與控制設施參數彙整表

道路參數		
類別	寬度	速限
高速公路主線	3.65 公尺	120 公里/小時
高速公路匝道	4.5 公尺	50 公里/小時
環境號誌參數		
三重北上入口匝道 臺北北上入口匝道 圓山北上入口匝道	G:45, Y:3, R: 5 (06-07 時) G:14, Y:3, R:11 (07-10 時)	
濱江街出口平面號誌	G:45, Y:3, R: 5 (06-10 時)	
註：G:綠燈時間；Y:黃燈時間；R:紅燈時間		

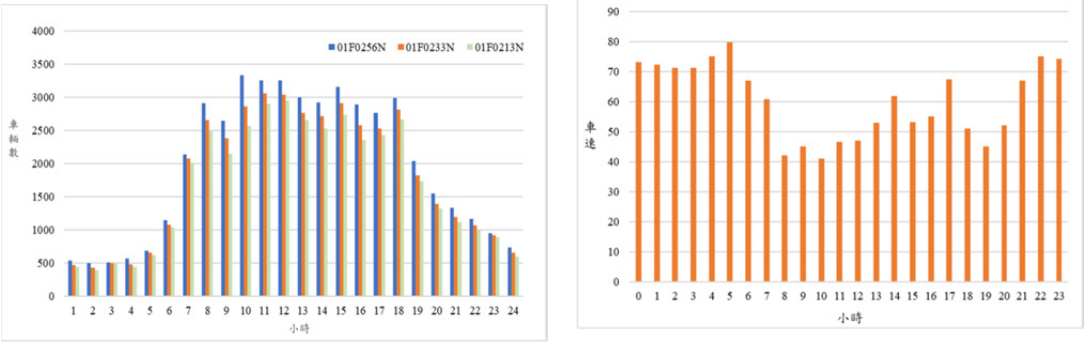
資料來源：本研究整理。

3.2 車流模擬的輸入及模型的校估與驗證

1. 車流模擬的輸入

本研究依 SUMO 要求的格式 (route.xml) 產製模擬模型輸入資料，包含車輛型式、加

速度、減速度、車長、跟車模式參數、最大車速、跟車模式以及車道變換模式等車輛參數資料。亦包含交通量資料，包括每小時車流量、車輛起迄與路徑軌跡 (車輛出發處與到達處) 以及車輛發生點的放車時隔等。其中，本研究將車輛分成小型車、大型車以及聯結車 3 類，模擬環境各項車輛參數設定如表 3 所示。



(a) 每小時平均車流變化　　(b) 01F0256N~01F0213N 間平均車速

圖 2　01F0256N 至 01F0213N 工作日每小時車流狀態

表 3　模擬環境車輛參數設定表

車種	車長 (m)	最大車速 (m/s)	最小車間距 (m)	跟車模型	車道變換模式	加速度 (m/s^2)	減速度 (m/s^2)
passenger veh.	4.70	27.78	2.5	Wiedemann99	DK2008	2.6	4.5
Truck/bus	8.00	25.00					
Trailer	13.50						

*註：最小車間距、加(減)速度為 SUMO 默認值。

交通量資料方面，本研究下載的旅次資料以高公局交通資料庫 (https://tisvcloud.freeway.gov.tw/) 中各旅次路徑原始資料 (M06A) 為依據，包含 01F0256N (三重-臺北)、01F0233N (臺北-圓山)、01F0213N (圓山-內湖) 等門架偵測的數據。將資料整併後並刪除相同出發時間且出現重複門架的旅次予以刪除，可得到每工作日 0600~1000 時平均約 25,790 筆資料。其中，0600~0700 時之旅次資料是為了暖機而蒐集，以利於模擬環境進行機器學習時為平穩狀態。(工作日為週二~週四是參考高公局日交通量參考值資料數據對於工作日與假日的區分)。

研究中模擬模式以五股-三重、三重北上入口、臺北北上入口、圓山北上入口為發車點 (可對應至前項所述的各 ETC 門架偵測的旅次軌跡資料)，將測試工作日 0600~1000 時的數據以每 60 分鐘為間距予以區分，即可獲得每 60 分鐘各發車點進入的旅次軌跡。根據旅次

軌跡資料可以得到各發車點旅次發生數、起迄交通量以及車種比例等資料。首先依據偵測資料，將連續車輛出現時間相減，即可以得到車輛發生的間隔時間。經 Kolmogorov-Smirnov (KS) 檢定，車輛發生的時隔呈現的分配狀態如下表 4 所示。本研究路段的最上游是中山高速公路五股-三重北向主線路段，其車輛到達時隔呈均一分配，其他三個入口匝道 (三重、臺北、圓山) 的車輛到達間隔均呈指數分配。依據表 4 的旅次發生的統計分佈型態即可隨機生成車流，具體作法是將 SUMO 中屬性 period 進行設定。例如，三重匝道入口車輛發生數時隔為指數分佈，因此將 period="exp(λ)"。其中 λ 是每小時交通量參數，其參數將使得車輛發生的時隔遵循 λ (每小時交通量) 的指數分佈。其次起迄交通量上，則藉由車輛軌跡可判斷相同入口發生的旅次到達迄點匝道的分佈比例，將其於 routeDistribution 中設定旅次到達匝道的比例。最後是車種比例，亦依照旅次軌跡資料不同車種比例分佈，利用 createVehTypeDistribution.py 套件於各旅次發生時隨機賦予車種。至此，前述的方法將可驅動模擬環境車流。

表 4　車輛發生時隔 KS 檢定

區域或匝道	P-value	結果分析	配適說明	檢定方法
五股-三重	0.231	Don't reject H_0	均一分配	KS 檢定
三重入口	0.397	Don't reject H_0	指數分配	
臺北入口	0.267	Don't reject H_0	指數分配	
圓山入口	0.113	Don't reject H_0	指數分配	

此外，藉由高公局的交通資料庫可進一步與 VD 動態資訊、日交通量參考值 (匝道、主線) 等資料進行關聯，共可獲得 156 個工作日 06-10 時的車流資料，經以 60 分鐘進行區分並計算平均值，可得到各區域及出入口處 06-10 時每 60 分鐘實際的平均流量，如表 5 所示。

2. 模擬模型的校估與驗證

本研究以研究路段上 3 座 ETC 門架 (編號 01F0256N、01F0233N 及 01F0213N) 及其鄰近 VD 偵測得的車流量分別作為模擬模型校估驗證的基準。模擬的產出，則係在上述門架處設置虛擬門架，以獲取模擬數據再依據式 (6) 對每 60 分鐘數據逐筆校估驗證。

目前研究中模擬及驗證的資料，是任意選取一日的數據進行模擬。對於 156 個工作日任取一日的代表性問題可以分成二部分進行驗證，第一是對於 2019 年每工作日旅次到達時隔資料進行隨機性鏈檢定 (run test)，確認 156 筆工作日數據的隨機性質；第二部分是以統計方法說明其中任一個工作日的旅次樣本資料的到達時隔與其他未選取工作日的旅次樣本之間進行 Kolmogorov-Smirnov 檢定 (KS 檢定)，以確認樣本資料之間的同質性。如所有工作日的旅次樣本皆具備隨機且彼此之間具有同質性，則任意挑選的工作日進行模擬均能滿足代表性問題。

表 5　研究範圍各路段與出入口工作日 06-10 每 60 分鐘平均流量

時間	五股–三重	三重入口	三重–臺北	臺北出口	臺北入口	臺北–圓山	圓山出口	圓山入口	濱江街出口	圓山續行
0600~0700	2981	641	3511	521	822	3813	755	522	498	3117
0700~0800	3247	1342	4561	1011	1551	5024	961	807	751	4351
0800~0900	3398	1377	4688	964	1461	5133	942	779	709	4510
0900~1000	3244	1228	4412	992	1232	4622	895	722	559	3988

(1) 隨機性檢定：

有關 2019 年工作日計 156 筆資料分別進行鏈檢定，由於數據量大，本研究以 Python 引入 statsmodels.sandbox.stats.runs 套件撰寫鏈檢定，直接將 156 筆資料逐次匯入進行運算。有關假設如下：

H_0 (虛無假設)：工作日旅次到達時隔具有隨機性

H_1 (對立假設)：工作日旅次到達時隔不具備隨機性

藉由大樣本鏈檢定的計算，156 筆資料均可得到 z 值所對應的 p-value 均大於 0.05 的臨界值無法拒絕虛無假設 H_0。因此，2019 年工作日 156 筆資料均具有隨機性。

(2) 同質性檢定：

有關同質性檢定，由於數據量大，本研究以 Python 引入 scipy.stats.ks_2samp() 套件撰寫 KS 檢定，將資料逐次匯入進行兩兩檢定。有關假設如下：

H_0 (虛無假設)：工作日旅次到達時隔具有同質性

H_1 (對立假設)：工作日旅次到達時隔不具備同質性

藉由 KS 檢定的計算，156 筆資料成對檢定後可得到統計值所對應的 p-value 成對檢定的統計值均大於 0.05 的臨界值，均無法拒絕虛無假設 H_0。因此，2019 年工作日 156 筆資料彼此之間均具有同質性。

此外，本研究的模擬模型具隨機性，需模擬數次以其平均值代表模擬的產出來避免隨機的不確定性。本研究以式 (7) 來估算最少需重複模擬的次數，以取得經濟有效的模擬產出。

$$n^*(\gamma) = \min\left\{ i \geq n : \frac{t_{i-1,1-\frac{\alpha}{2}}\left(\sqrt{S^2(n)/i}\right)}{\bar{X}(n)} \leq \gamma' \right\} \tag{7}$$

在式 (7) 中，γ為在信賴水準α下滿足相對誤差水準所需的最少重複模擬次數；$n^*(\gamma)$為最少重複模擬次數；$t_{i-1,1-\alpha/2}$ 為在 $i-1$ 自由度與信賴水準 α下的 t 值；$S^2(n)$為已模擬 n 次的車流量標準差平方；$\bar{X}(n)$為模擬 n 次的車流量平均值；其中，$\gamma' = \gamma/(1+\gamma)$。經

逐次進行確認，在給定相對誤差 γ=0.1 以及信賴水準 α 為 0.9 的前提下可知 γ'=0.09，並藉由式 (7) 可以得到表 6 (a) 至表 6 (c) 中各時間的最少模擬次數值。由於表 6 (a) 至表 6 (c) 所有最少模擬次數值至多為 151 次，故以模擬 151 次所得到的模擬平均車流再與實際偵測車流計算 GEH 值。校驗結果各時間點 ETC 門架偵測的車流與 VD 偵測的車流雖存在差異，但與模擬的車流間均能滿足 GEH≦5 標準，亦即本研究建立的模擬環境對實際車流環境已具備仿真的能力，可在後續強化學習的架構中應用。

表 6 (a) 模擬車流校驗 (01F0256N)

時間	VD 編號 VD-N1-N-27.100-N-LOOP 偵測車流	ETC 偵測車流	最少重複 模擬次數	模擬 平均車流	GEH 值 (VD-模擬)	GEH 值 (門架-模擬)
0700~ 0800	4321	4327	151	4255	0.51	0.61
0800~ 0900	4688	4648	149	4657	0.11	0.01
0800~ 1000	4412	4398	147	4378	0.13	0.05

表 6 (b) 模擬車流校驗 (01F0233N)

時間	VD 編號 VD-N1-N-25.800-M-LOOP 偵測車流	ETC 偵測車流	最少重複 模擬次數	模擬 平均車流	GEH 值 (VD-模擬)	GEH 值 (門架-模擬)
0700~ 0800	5024	4991	148	5111	0.75	1.43
0800~ 0900	5133	5046	150	5099	0.12	0.27
0800~ 1000	4622	4607	151	4581	0.18	0.07

表 6 (c) 模擬車流校驗 (01F0213N)

時間	VD 編號 VD-N1-N-23-O-SE-1A 偵測車流	ETC 門架編號 01F0213N 偵測車流	最少 重複模擬次數	模擬 平均車流	GEH 值 (VD-模擬)	GEH 值 (門架-模擬)
0700~ 0800	4351	4299	146	4311	0.18	0.01
0800~ 0900	4510	4388	150	4417	0.969	0.10
0800~ 1000	3988	3971	144	3959	0.11	0.02

3.3 建立多代理人強化學習框架

本研究以 SUMO、輸入的模擬數值以及多代理人強化學習模型等，建立強化學習框架，如圖 3 所示。面對變動的車流環境，三重、臺北、與圓山入口匝道設置號誌控制代理人，各代理人均賦予策略網路，其輸入數值為第 i 個代理人鄰近該區域 ETC 門架位置的 VD 環境偵測值o_t^i；其網路輸出為 $\pi(A^i|s;\theta^i)$，是各個代理人離散動作空間上的機率分布，代理人可依據隨機所對應的機率值選取機率值較高的動作。經執行動作a_t^i後，環境狀態由 s_t 移轉至狀態 s_{t+1}。而藉由s_t可獲得環境反饋的獎勵 r_t，而s_t、s_{t+1}將分別作為價值網路與目標網路的輸入進行學習訓練。

在價值網路與目標網路的輸出為 $\hat{v}_t$以及$\hat{v}_{t+1}^-$，分別為 t 時刻的狀態與 t+1 時刻狀態的價值評估。$\hat{v}_t$、$\hat{v}_{t+1}^-$以及獎勵 r_t將進行 TD 誤差的計算，其誤差將分別更新價值網路與目標網路的參數 ω、ω^-。藉由學習的過程使得價值網路的輸出將不斷接近目標網路輸出，降低 TD 誤差。前述意義可直觀理解為 t 時刻的狀態價值逐漸貼近 t+1 時刻狀態價值。此外，TD 誤差亦將回饋至代理人的策略網路，更新網路參數θ^i，策略網路將據此輸出更新後的 $\pi(A^i|s;\theta^i)$ 以採取動作 a_{t+1} 施加於環境中，如此迭代循環將逐步提升環境狀態價值，以降低 TD 誤差。有關多代理人強化學習框架中的各網路設定與結構，如下表 7 所示。

表 7　強化學習各網路設定與結構

	(1)3 層，各 32 個神經元 (2)激勵函數：ReLU	(1)替選動作數量的神經元 (2)激勵函數：Softmax
$s=[o^1,\dots,o^m]$	(1)3 層，各 32 個神經元 (2)激勵函數：ReLU	(1)一個神經元 (2)激勵函數：Sigmoid
$s=[o^1,\dots,o^m]$	(1)3 層，各 32 個神經元 (2)激勵函數：ReLU	(1)一個神經元 (2)激勵函數：Sigmoid

表中各網路隱藏的神經元以整流線性單位函數 (Rectified Linear Unit, ReLU) 作為激勵函數，其常用的函數表示型態為 f(x)=max (0,x)，其中 x 為神經元的輸入值，其代表的意義就是當 x 的輸入值大於 0 的時候，函數將進行完整的輸出，否則以 0 作為這個函數的輸出。相較於 sigmoid (二焦點曲線函數)、tanh (雙曲正切函數) 等激勵函數，ReLU 函數較能夠避免深度學習中梯度消失問題。另外，在各個網路的輸出層的激勵函數，策略網路為 Softmax 函數 (歸一化指數函數)，適用於多元輸出問題；而價值網路與目標網路的輸出為一實數，則激勵函數以 Sigmoid 函數較為妥適。

本研究每次均進行 1000 回合的學習，每一回合均進行一次 0600-1000 的模擬，每模擬 15 分鐘進行環境狀態取值並變更一次行動 (匝道號誌時制)，藉由每回合與環境的互動，代理人逐漸學會如何控制環境。本研究每回合模擬期間以每 15 分鐘進行切割並變換號誌時

制，主要是考慮代理人將行動發佈後，環境的反應時間以及狀態取值是否有效。如切割時間過短，代理人動作尚未對環境造成影響；反之，切割時間過長則無法以適當動作即時回應環境的變化，目前所採取 15 分鐘的切割時間是經過多次試驗後的較佳結果。

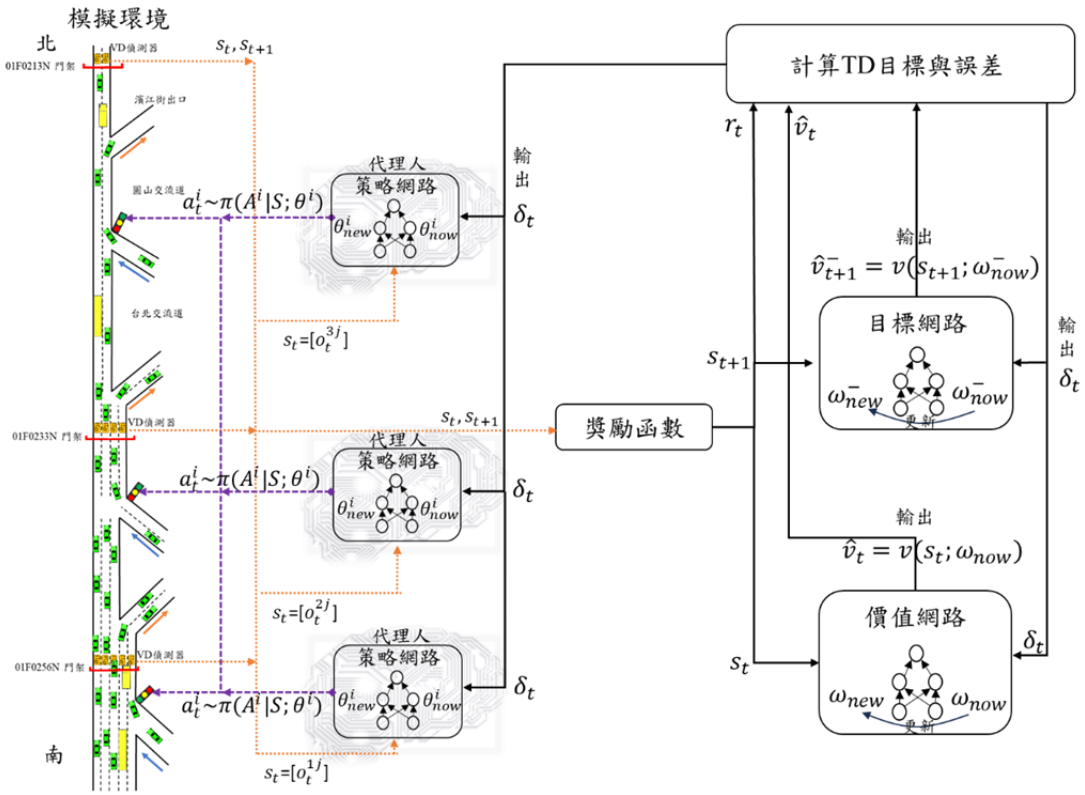


圖 3　多代理人強化學習框架示意圖

3.4 建構馬可夫決策過程

強化學習是代理人與環境互動的試誤過程，而環境-儀控時制-強化學習的程序應可構建為馬可夫決策過程 (MDP)。馬可夫決策過程包含了 (S, A^i, P, R, γ) 等元素，其中，S 為狀態空間、A^i為動作的集合、P 為狀態轉移函數、R 為獎勵函數、$\gamma \in [0,1)$為折扣率，將模擬環境的時間狀態轉換為離散狀態 t=1,2,3,…。號誌代理人 (agent) 將根據策略網路學習後，可得到在狀態 s 中每個代理人 i 可能動作a^i作為最佳動作的機率。

當所有的代理人執行完動作後，環境依據狀態移轉 (state transition function) 確定下一時刻的狀態。以本研究的環境為例，是指當前車流狀態s_t (SUMO 在時間 t 的車流狀態)，所有代理人執行了動作 a_t^i 一單位時間後，SUMO 模擬環境就會因為行動的執行而得到新車流狀態s_{t+1} 的機率，如下式 (8)：

$$p_t(s_{t+1}|s_t;a_t) = P(S = s_{t+1}|S_t = s_t, A_t = a_t) \quad (8)$$

狀態移轉函數是條件機率，該式的意義是下一時刻的狀態S_{t+1}取決於當前時刻狀態S_t以及環境中所有代理人的動作 $A_t = [A_t^1, A_t^2, A_t^3]$。以下將就號誌控制代理人、環境狀態、代理人動作以及獎勵函數進行定義。

1. 號誌控制代理人

本研究將高速公路匝道號誌控制器視作代理人，在強化學習的框架下，代理人的學習即是與環境互動試探與評價的過程。在此，環境中代理人們將依當前的車流模擬環境的狀態以策略網路學習的結果近似策略函數 (π)，做出相對應的行動控制各匝道號誌。本研究路段有多個號誌，是多代理人問題，代理人們將分別由行動列表選擇具體行動進行控制。代理人學習的目標在追求最大化的累積折扣獎勵，而好的折扣獎勵設計則可據以舒緩本路段壅塞現象，提高運作效率。

2. 環境狀態

環境狀態 ($s_t \in S$) 是即時衡量交通環境變化的指標。車流狀態變化的數值可以通過偵測器 (如 VD) 取得，環境模擬時間為上午 0600~1000，0600~0700 為暖機時間，模擬一次即為 1 回合 (episode)。其中，參照 高速公路實際 ETC 門架設置地點，在其相對位置於 SUMO 環境中所設置虛擬 VD，包括 01F0256N 門架計 5 車道 (含減速車道)、01F0233N 門架計 4 個車道 (含減速車道)、01F0213N 門架計 2 個車道，如圖 4 所示。多代理人強化學習的狀態由不同代理人所在附近的觀測組成，以每 15 分鐘一次分別擷取不同位置所偵測的平均車速與流率作為環境觀測，如下表 8。

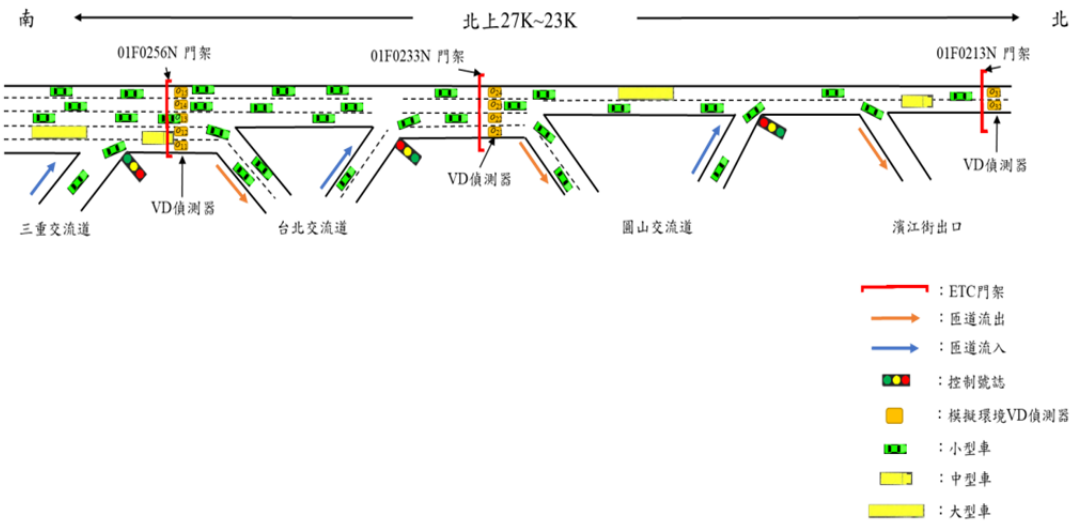


圖 4　模擬環境示意圖

表 8　深度強化學習的狀態值

編號	所屬代理人	門架編號	偵測器數量	觀測
1	三重入口號誌	01F0256N	5	o^1=[$ave_s^{11}, ave_s^{12}, ave_s^{13}, ave_s^{14}, ave_s^{15}, q^{11}, q^{12}, q^{13}, q^{14}, q^{15}$]
2	臺北入口號誌	01F0233N	4	o^2=[$ave_s^{21}, ave_s^{22}, ave_s^{23}, ave_s^{24}, q^{11}, q^{12}, q^{13}, q^{14}$]
3	圓山入口號誌	01F0213N	2	o^3=[$ave_s^{31}, ave_s^{32}, q^{11}, q^{12}$]

註：其中o^i表第 i 區域觀測的集合。

3. 獎勵

獎勵 (r) 所反映的是對當前環境狀態的評價，其設計的良窳將直接影響代理人的學習。R^i可記為第 i 個代理人的獎勵的隨機變數，在完全合作的設定下，第 t 時間的獎勵r_t將由狀態 S_t 與所有代理人的動作$A = [A^1, \dots, A^m]$ 所共同決定。當第 i 個代理人所屬的當地狀態被實際觀測後，經由下式 (9) 計算後其值記 $\tilde{r}_t^i$。由於環境中代理人之間需相互協作，使得狀態績效能夠提升，因此總體獎勵可設計為所有代理人所在各區域獎勵的和 (如下式 (10))。本研究關注之焦點在於舒緩測試路段壅塞程度，而壅塞時段具有流量低車速亦低的特性。所以，應同時提升該區域各個車道的平均速度與車流，故而將 SUMO 環境中觀測區所偵測的平均速度與車流列為當前獎勵中狀態的評價較為妥適。

$$\tilde{r}_t^i = \sum_j ave_s_{ij}^t \times q_{ij}^t \tag{9}$$

$ave_s_{ij}^t$：時間 t，第 i 個代理人偵測區的第 j 車道觀測的平均速度

q_{ij}^t：時間 t，第 i 個代理人偵測區的第 j 車道觀測的車流

$$r_t = \tilde{r}_t^1 + \tilde{r}_t^2 + \cdots + \tilde{r}_t^m \tag{10}$$

4. 代理人所採取的行動

代理人所採取的動作 ($a_t \in A$) 係在所有動作選項中，依策略網路的輸出作為選取動作的依據，代理人各自執行一個動作與當前模擬環境的交通狀態相互作用。本研究以 3 個代理人分別控制圓山北上入口、三重北上入口、臺北北上入口等 3 個北上入口號誌。同時，參照高公局交通工程手冊對於匝道流入率對應的儀控時制為基礎進行簡化設定，如表 9 所示。該表所列之儀控時制代表前述三個入口匝道不同號誌代理人的行動方案，行動編號為 0,1,2,…,15，匝道的流入率增加 100 (vph) 的遞增變化。

本研究提出動作轉換函數，如下式 (11)。其中，$flowrate_t^i$為第 i 個代理人所在的入口匝道於時間 t 的流入率； a_t^l為第 l 入口匝道代理人在時間 t 時所採取的動作；I 為匝道入口進入車流的增量，設為 100 (vhp)。可利用策略網路對於所有動作機率輸出值以隨機抽樣選

取方式挑選行動，進而對於匝道的流入率進行控制。

$$flowrate_t^l = a_t^l \times I \tag{11}$$

表 9　代理人行動 (儀控時制) 列表

行動編號	流入率(vph)	週期時間(sec)	紅燈時間(sec)	綠燈時間(sec)	黃燈時間(sec)
0	<100	長紅	長紅	--	--
1	100	72	65	4	3
2	200	48	38	7	3
3	300	48	35	10	3
4	400	36	23	10	3
5	500	36	16	17	3
6	600	28	11	14	3
7	700	21	8	10	3
8	800	24	8	13	3
9	900	25	7	15	3
10	1000	28	7	18	3
11	1100	33	7	23	3
12	1200	36	6	27	3
13	1300	45	6	36	3
14	1400	53	5	45	3
15	>1400	長綠	長綠	--	--

資料來源：本研究整理。

代理人選擇動作的具體方式敘述如下：

a. 目前環境中計有三重北上入口　(I)、臺北北上入口　(II)、圓山北上入口　(III)，各個代理人的動作依表 9 進行編號為 0,1,2,…,15。動作的選擇以臺北北上入口匝道(II)為例，在 t 時刻，經由該策略網路輸出各個動作機率，進而依機率分布狀態隨機選取動作a_t^{II}，輸出機率較高動作被選取的機會較高。所選取的動作假設為 2，將$a_t^{II} = 2$帶入上式　(11)，得到在時間 t 匝道的流入率為　$\text{flowrate}_t^{II} = 2 \times 100 = 200$(vph)。

b. 其他三重、圓山入口匝道代理人亦可利用前述 a 的方式獲得時間 t 的動作，進而對於環境進行控制。

3.5 多代理人強化學習演算過程

MAC-A2C 演算法 (Multi-Agent Challenge Advantage Actor-Critic) 是 Samvelyan et al.[23] 於 Actor-Critic 基礎上改進的演算法，通過引入多代理人的價值函數與策略函數以實現多

代理人在完全合作關係下的協同學習。本研究利用前述演算法概念，結合 SUMO 的車流環境建立強化學習框架，如前圖 3 所示，該框架主要是包含模擬環境、號誌控制代理人以及策略學習等。其中的模擬環境先前已完成校驗的模擬環境，代理人藉由模擬環境中埋設的 VD 偵測器讀取的狀態 (s_t，各車道速度與流量) 給出相應的號誌程序 (a_t^i動作)，並藉由獎勵函數得到獎勵值 (r_t^i) 以及時間 $t+1$ 的狀態 (s_{t+1})。

在網路參數的更新上，除了策略網路、價值網路與目標網路的各項超參數設置 (如前述之表 7)，另採用 Adam 優化器 (Kingma and Ba[24]) 進行運算，該優化器結合了動能 (momentum) 和學習率自適應的特性，適用於訓練學習網路。在每回合訓練步驟中，透過向前傳播計算損失函數，並透過反向傳播計算損失函數對網路參數的梯度，進而更新網路中的參數值，代理人依據策略網路輸出近似 $\pi(a|s;\theta^i)$，以隨機取樣的方式，在特定狀態選擇出最佳動作。以下為本研究之強化學習演算法：以隨機方式起始化策略網路、目標網路以及價值網路的各項參數值。

1. 設定學習回合 (episode) 數 T=1000；學習率 α =0.995；折扣因子 γ =0.8。
2. For episode = 1 to T
   a. 藉由 SUMO 開啟本研究路段之模擬環境，For t to 16 (t 為模擬時間由 0600 開始，每 15 分鐘 t+1)
   b. 由模擬環境中車道所設置的 VD 偵測器所偵測的狀態值作為環境初始狀態 ($s_t = [o_t^{11}, o_t^{12}, \dots, o_t^{ij}]$)。
   c. 當前狀態 s_t 作為策略網路的輸入，讓每個代理人 i 在參數為θ_{now}^i下，計算當前策略網路的輸出。各代理人獨立進行隨機抽樣 $a_t^i \sim \pi(A^i|S;\theta^i)$ 以挑選動作a_t^i，並依前式 (9)、(10) 計算獎勵 r_t。
   d. 模擬環境在所有代理人執行動作 a_t^i後，可得出與下一個環境狀態 s_{t+1}。
   e. 將狀態 s_t 作為價值網路的輸入，在參數為 ω_{now}下，計算當前價值網路的輸出 $\hat{v}_t = v(s_t;\omega_{now})$。
   f. 另將環境偵測 $t+1$ 的狀態s_{t+1}作為目標網路的輸入，在參數為ω_{now}^-下，計算當前目標網路的輸出 $\hat{v}_{t+1}^- = v(s_{t+1};\omega_{now}^-)$。
   g. 藉由價值網路輸出值 ($\hat{v}_t$)、目標網路輸出值 ($\hat{v}_{t+1}^-$)、折扣率 (γ) 以及計算 TD 目標 ($\hat{y}_t^-$) 與誤差 (δ_t)，分別如下式 12、13：

$$\hat{y}_t^- = r_t + \gamma \cdot \hat{v}_{t+1}^- \tag{12}$$

$$\delta_t = \hat{v}_t - \gamma \cdot \hat{y}_t^- \tag{13}$$

在此，將式 (12) 求得的 TD 目標 ($\hat{y}_t^-$) 視作常數，藉由式 (13) 與訓練價值網路可使得$\hat{v}_t$不斷接近 $\hat{y}_t^-$，減少 TD 誤差 (δ_t)。

   h. 更新價值網路參數$\boldsymbol{\omega}$：

$$\omega_{new} \leftarrow \omega_{now} - \alpha \cdot \delta_t \cdot \nabla_\omega v(s_t;\omega_{now}) \tag{14}$$

式 (14) 中α表示為價值網路參數值更新的學習率。

i. 更新各策略網路參數$\boldsymbol{\theta}^i$：

$$\omega_{\text{new}}^{-} \leftarrow \tau \cdot \omega_{\text{new}} + (1-\tau) \cdot \omega_{\text{now}}^{-} \tag{15}$$

將價值網路中更新後的參數$\boldsymbol{\omega}_{new}$用以更新目標參數$\boldsymbol{\omega}_{new}^{-}$，上示式 (15) 中的$\tau$是介於 0~1 之間的權重值，在此設定為 0.5。

j. 更新各策略網路參數 θ ^i：

$$\theta_{\text{new}}^{\text{i}} \leftarrow \theta_{\text{now}}^{\text{i}} - \beta \cdot \delta_t \cdot \nabla_{\theta^{\text{i}}} \ln\pi\left(a_t^{\text{i}} \middle| s_t; \theta_{\text{now}}^{\text{i}}\right) \quad \forall \text{i} = 1, \ldots, \text{m} \tag{16}$$

k. end for

3. end for

四、數值驗證與分析

4.1 深度強化學習框架訓練測試

在 MAC-A2C 的演算法下，模型中的策略網路、目標網路以及價值網路迭代更新，各匝道代理人依模擬環境中所屬區域的偵測器，每 15 分鐘擷取各車道的平均流率以及車速等觀測值計算獎勵值。此外，代理人在選擇行動上，係根據代理人自身的策略網路的輸出，以隨機取樣由表 10 中選擇動作，以因應當前狀態。雖然每一個入口匝道的動作是獨立產生，但是價值網路與目標網路在學習時的輸入分別是 *t* 及 *t*+1 時刻的全局狀態，再藉由網路輸出值與獎勵計算 TD 目標與誤差，即是對於所有的代理人的表現進行評分，而策略網路亦依據 TD 誤差進行更新，促使代理人之間進行合作協調，進而對於環境做出貢獻。

圖 5 中藍色曲線為多代理人學習 1000 次訓練下獎勵值的變動狀態，橘紅色曲線則為依現行高速公路匝道號誌時間模擬對應的獎勵值。由圖 5 可知，透過不斷地學習訓練，各代理人已逐漸能選擇動作因應變動的車流環境。在學習初期獎勵數值在 1400~1600 間震盪，經迭代學習後，各匝道入口號誌代理人由列表中隨機挑選動作，應用於測試環境以提升獎勵數值，經由學習獎勵值不斷提升，歷經約 6 百餘次迭代後，獎勵數值漸趨於穩定於 2486。而在現行高速公路匝道號誌時間下的模擬環境所反映的獎勵數值，因每次模擬隨機的結果，故而獎勵值於 1930 上下隨機波動。

各匝道入口代理人依據模擬環境的狀態所選取的動作將隨著學習過程而趨於穩定，使得累積折扣獎勵最大化。由表 10 可看出，在模擬初期階段，車流壅塞狀態尚不顯著，代理人選擇以較長的綠燈時間與較短的紅燈時間的動作進行控管 (如圓山北上入口匝道 0700~0730 的綠燈時間為 36 秒)。隨著環境時間的推移，車流開始增加，匝道的號誌開始縮短週期長度並且增加紅燈時間，以限制進入車流，充分說明號誌代理人在面對變動環境

以及獎勵函數設計下，已能夠於不同時間狀態下選取適切的動作策略提升環境獎勵數值。

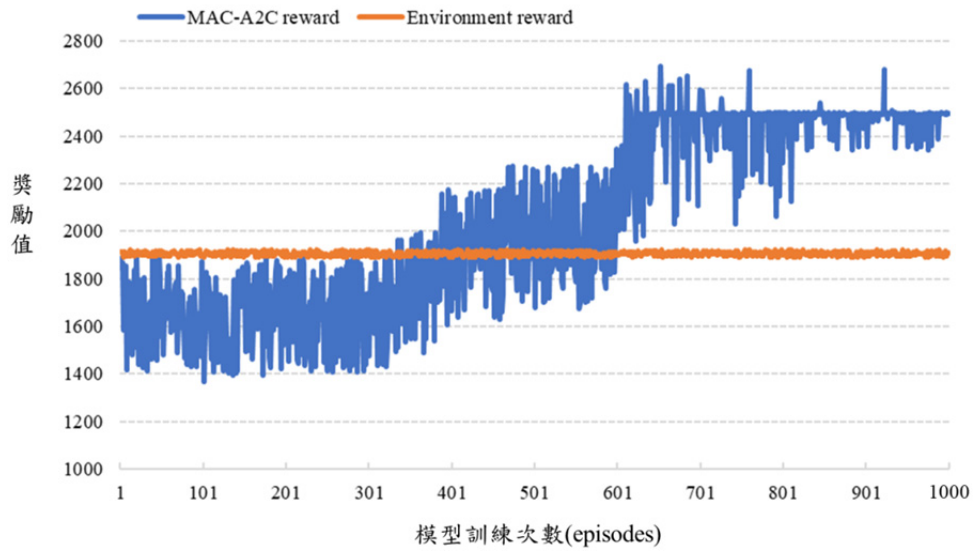


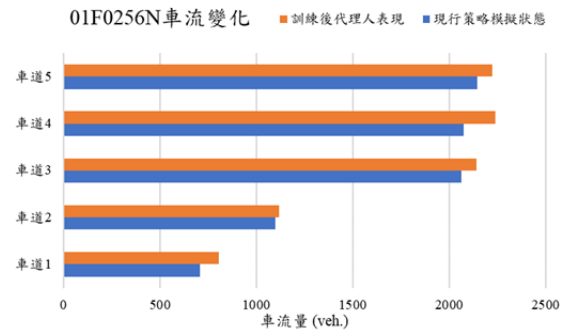
圖 5　獎勵數值變動

表 10　代理人學習後所取行動表

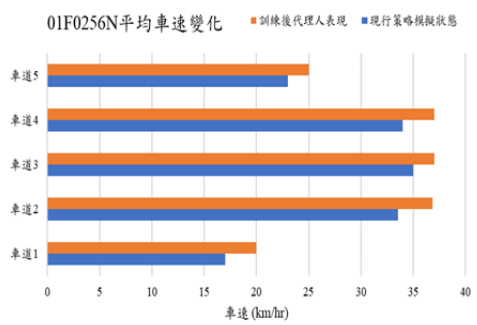
模擬時間	各匝道代理人行動		
	三重北上入口	臺北北上入口	圓山北上入口
0700-0715	12(G:27, Y:3, R:6)	12(G:27, Y:3, R:6)	13(G:36, Y:3, R:6)
0700-0730	12(G:27, Y:3, R:6)	12(G:27, Y:3, R:6)	13(G:36, Y:3, R:6)
0730-0745	8(G:13, Y:3, R:8)	9(G:15, Y:3, R:7)	5(G:16, Y:3, R:17)
0745-0800	8(G:13, Y:3, R:8)	9(G:15, Y:3, R:7)	5(G:16, Y:3, R:17)
0800-0815	4(G: 10, Y:3, R: 23)	5(G:16, Y:3, R:17)	6(G:14, Y:3, R:8)
0815-0830	4(G: 10, Y:3, R: 23)	5(G:16, Y:3, R:17)	6(G:14, Y:3, R:8)
0830-0845	3(G:10, Y:3, R:35)	4(G: 10, Y:3, R: 23)	4(G: 10, Y:3, R: 23)
0845-0900	3(G:10, Y:3, R:35)	4(G: 10, Y:3, R: 23)	4(G: 10, Y:3, R: 23)
0900-0915	4(G: 10, Y:3, R: 23)	4(G: 10, Y:3, R: 23)	5(G:16, Y:3, R:17)
0915-0930	4(G: 10, Y:3, R: 23)	4(G: 10, Y:3, R: 23)	5(G:16, Y:3, R:17)
0930-0945	9(G:15, Y:3, R:7)	9(G:15, Y:3, R:7)	10(G:18, Y:3, R:28)
0945-1000	9(G:15, Y:3, R:7)	9(G:15, Y:3, R:7)	10(G:18, Y:3, R:28)

註：範例說明：13(G:36, Y:3, R:6)-表示行動編號為 13，括號內為綠、黃、紅燈時間。

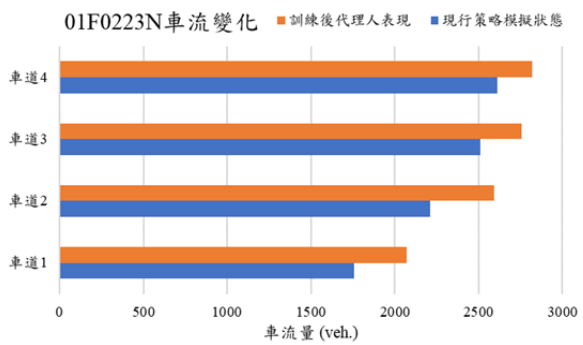
除前述的獎勵值與代理人行動分析外，亦須對於代理人在模擬環境的表現進行討論。由圖 6 可知，主線通過車流量及平均車速在號誌代理人控制下，相較於現行匝道號誌時間，績效均有不同程度的改善。其中模擬環境中於 01F0256N、01F0223N、01F0213N 門架相對位置偵測的各車道車速與車流量均有不同程度的提升。在車流變化方面，01F0256N 提升 5.21%、01F0223N 提升 9.75、01F0213N 提升 1.02%；在平均車速部分，01F0256N 提升 1.42%、01F0223N 提升 10.05%、01F0213N 提升 3.84%。如以獎勵值的計算角度將平均車速與車流量相乘後的提升則為 19.54 %。由此前述分析可以得知，當獎勵以主線的車流率與速度的提升，對於各代理人形成刺激。在與模擬環境的訓練互動後，各代理人因應狀態變化彼此合作下，給予不同的行動，確實可以提升主線各車道平均車速，並且相當程度的增加通過車流率，也同時說明多代理人強化學習在研究環境的適用性。



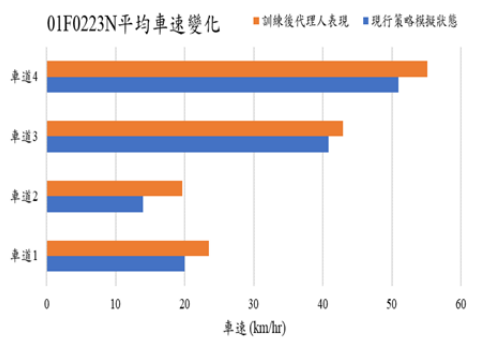
(a) 01F0256N 各車道車流



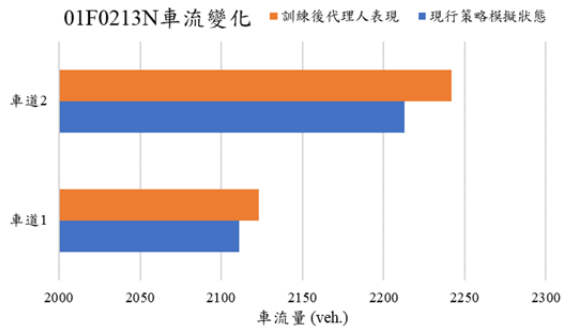
(b) 01F0256N 各車道平均車速



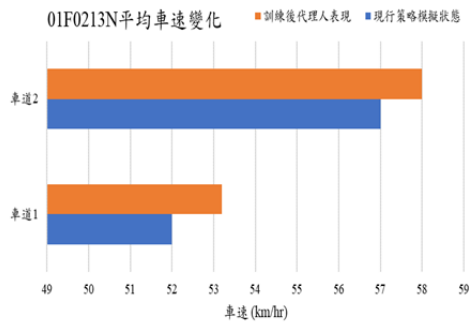
(c) 01F0223N 各車道車流



(d) 01F0223N 各車道平均車速



(e) 01F0213N 各車道車流



(f) 01F0213N 各車道平均車速

圖 6　代理人控制下環境之車流量及車速狀態

4.2　一般化能力測試

本研究採用一般化測試(又稱泛化能力測試, Generalization)，以驗證代理人在完成強化學習框架訓練後，對於研究區域路段在相同分布的隨機車流型態仍能有效的進行管控。此項測試類似監督式學習的驗證，主要是以訓練期間模型未見過的數據上仍能保有訓練時的表現，以避免過擬合的問題發生。但是，在強化學習並無法區分訓練環境與驗證環境，因此，Bellemare et al. [24]、Tobin et al. [26] 係以隨機改變環境初始化設定的方式來進行一般化測試。本研究亦採用此種作法，即改變車輛發車時隔，以表 4 車輛發生時隔配適分配表為基礎，例如，五股-三重車輛發生時隔為均一分布是將分布的上界值增加一常數值，同時下界值減去一常數值，其餘三重、臺北、圓山等入口為指數分布，則是將其參數值 λ 控制在縮小與放大 1/2 以生成亂數，使車輛到達的時隔的變化程度增加，以驗證模型框架訓練後的一般化能力。

有關一般化測試結果如圖 7 所示，由圖可觀察出車輛到達的時隔變化增加，將使得原本的道路車流產生變化，經測試後結果如圖所示。由圖可觀察出，當各區域或入口匝道車輛發生時隔的變化程度增加，所得到的獎勵值亦有所改變，環境狀態所得到的獎勵值整體表現較原訓練階段的獎勵值(2480)為低，但是整體狀態環境的獎勵值能穩定在 2390 上下變動，說明研究的模型框架在變化程度較大的車流環境下，亦能保持穩定的能力與績效，充分展示本研究之深度強化學習框架具備一般化能力以及穩定性。

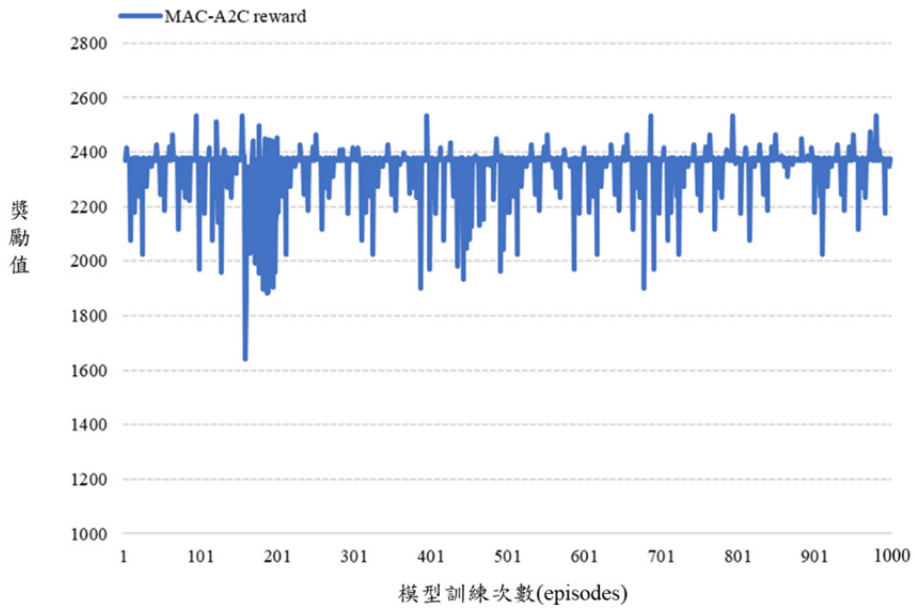


圖 7　一般化能力驗證獎勵值變動圖

五　結論與建議

由於資通訊技術及人工智慧的發展，使得 ITS 以及強化學習技術得以在交通管理領域發展與運用。本研究提出之結論與建議詳述如次：

1. 本研究針對高公局列管壅塞路段，利用網路爬蟲技術，將國道 ETC 以及 VD 車流資訊進行有效的收集，並藉由進行資料清整與統計獲得實際路段的實況，並經實地踏勘再次檢視數據資料與車流壅塞狀況是否相符。此外，利用前述交通特性資料及設施參數(如：入口匝道號誌時間)，藉由 SUMO 微觀車流模擬軟體建構該路段場景以及車流狀態。經以 FHWA 提供之校估方式與 ETC、VD 等車流資料進行驗證，結果說明本研究所構建之車流模擬環境合乎檢驗標準，即本研究採用的模擬模型能充分詮釋實際路段所反映的車流特性。對此，有利於深度強化的代理人於可信賴的環境下進行互動與學習。
2. 本研究提出的多代理人強化學習框架，是以多代理人為基礎，透過代理人對於模擬環境所反映的獎勵值、動作轉換函數以及代理人動作表，控制多個入口匝道號誌。結果顯示，號誌代理人訓練後，在模擬的環境下能做出最佳動作策略，相較於既有的匝道儀控方式，本研究的方法確能提升匝道儀控管制路段的主線車流量，同時提高平均車速。
3. 獎勵函數的設計關乎代理人的學習成效以及所互動的環境是否能達到交通管理的目標，本研究以管制範圍各主線路段的流量乘以速率的和來設計的獎勵函數，在代理人成熟穩健的表現下確可很快趨於收斂，值得後續研究參考。
4. 本文所選用數據為 2019 年全年工作日上午 0700 時至 1000 時，避開 2020 年新冠肺炎後的因素影響。此外，對於突發性狀況，如交通事件所造成加重壅塞現象，受限於事件之

不確定性及歷史數據不足等問題，並未納入研析。深度強化學習框架在面對此類問題時，必須因時因地制宜，重新進行建模與訓練，方適用於新的車流狀態。

5. 本研究目前是以多代理人方式透過強化學習控制多個入口匝道共同協作。未來，應考慮以區域控制的角度進行多代理人的協控問題，但是區域控制的多代理人可能存在彼此目標不同的問題，可以整體環境並考慮代理人之間多目標關係，使得代理人之間以多目標競合的方式解決問題，使得未來的研究更具成效與彈性。

參考文獻

1. Wu, Y., Tan, H., Qin, L., and Ran, B., "Differential variable speed limits control for freeway recurrent bottlenecks via deep actor-critic algorithm", *Transportation research part C: Emerging Technologies*, Vol. 117, 2020, 102649.
2. Masher, D. P., Ross, D. W., Wong, P. J., Tuan, P. L., Zeidler, H. M., Petracek, S., *Guidelines for Design and Operation of Ramp Control Systems*, Stanford Research Institute, Menlo Park, California, 1975.
3. Papageorgiou, M., Hadj-Salem, H., and Blosseville, J.-M., "ALINEA: A local feedback control law for on-ramp metering", *Transportation Research Record*, Vol. 1320, No. 1, 1991, pp. 58-64.
4. Wattleworth, J. A., *Peak-period analysis and control of a freeway system* (Report No. 24-15/HPR 1/4), Texas Transportation Institute, San Antonio (Tex.), 1965.
5. Papageorgiou, M., "A new approach to time-of-day control based on a dynamic freeway traffic model", *Transportation Research Part B: Methodological*, Vol. 14, No. 4, 1980, pp. 349-360.
6. Meldrum, D. R., and Taylor, C. E., *Freeway traffic data prediction using artificial neural networks and development of a Fuzzy Logic Ramp Metering Algorithm*, Washington State Department of Transportation, Olympia (Wash.), 1995.
7. Zhang, H. M., and Ritchie, S. G., "Freeway ramp metering using artificial neural networks", *Transportation Research Part C: Emerging Technologies*, Vol. 5, No. 5, 1997, pp.273-286.
8. Stephanedes, Y. J., "Implementation of on-line Zone Control Strategies for optimal ramp metering in the Minneapolis Ring Road", 7th International Conference on Road Traffic Monitoring and Control, 1994, pp. 181-184.
9. Gronauer, S., and Diepold, K., "Multi-agent deep reinforcement learning: a survey", *Artificial Intelligence Review*, Vol. 55, Issue. 2, 2022, pp. 895-943.
10. Yu, X. H., and Stubberud, A. R., "Markovian decision control for traffic signal systems", Proceedings of 36th IEEE Conference on Decision and Control, Vol. 5, 1997, pp. 4782-4787.
11. Alexiadis, V., Colyar, J., Halkias, J., Hranac, R., and McHale, G., "The next generation simulation program", *Institute of Transportation Engineers*, Vol. 74, No. 8, 2004, pp. 22.
12. Lopez, P. A., Behrisch, M., Bieker-Walz, L., Erdmann, J., Flötteröd, Y.-P., Hilbrich, R., Lücken, L., Rummel, J. Wagner, P .and WieBner, E., "Microscopic traffic simulation using SUMO", 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2018, pp. 2575-2582.

13. Ozbay, K., Yang, H., Morgul, E. F., Mudigonda, S., and Bartin, B., “Big data and the calibration and validation of traffic simulation models”, In Traffic and Transportation Simulation-Looking Back and Looking Ahead: Celebrating 50 Years of Traffic Flow Theory, a Workshop, 2014, pp. 92-122.

14. Gordon D. B., and Cameron, G. I., “PARAMICS — Parallel microscopic simulation of road traffic”, *The Journal of Supercomputing*, Vol.10, No. 1, 1996,pp. 25-53.

15. Fellendorf, M., “VISSIM: A microscopic simulation tool to evaluate actuated signal control including bus priority”, 64th Institute of Transportation Engineers Annual Meeting, Vol. 32, 1994, pp. 1-9.

16. Casas, J., Ferrer, J. L., Garcia, D., Perarnau, J., and Torday, A., “Traffic simulation with aimsun”, *Fundamentals of traffic simulation*, Springer Science+Business Media , Berlin, 2010, pp. 173-232.

17. W Axhausen, K., Horni, A., and Nagel, K., *The multi-agent transport simulation MATSim*, Ubiquity Press, London, 2016.

18. Geistefeldt, J., Giuliani, S., Vortisch, P., Leyn, U., Trapp, R., Busch, F., Rascher A. and Celikkaya, N., “Assessment of level of service on freeways by microscopic traffic simulation,” *Transportation Research Record*, Vol. 2461, Iss. 1, 2014, pp. 41-49.

19. Song, J., Wu, Y., Xu, Z., and Lin, X., “Research on car-following model based on SUMO”, The 7th IEEE/International Conference on Advanced Infocomm Technology”, 2014, pp. 47-55.

20. Kheterpal, N., Parvate, K., Wu, C., Kreidieh, A., Vinitsky, E., and Bayen, A., “Flow: Deep reinforcement learning for control in sumo”, *EPiC Series in Engineering*, Vol. 2, 2018, pp. 134-151.

21. Maiti, N., and Chilukuri, B. R., “Traffic Signal Control for An Isolated Intersection Using Reinforcement Learning”, 2021 International Conference on COMmunication Systems & NETworkS (COMSNETS), 2021, pp. 629-633.

22. Li, Z., Liu, P., Xu, C., Duan, H., and Wang, W., “Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks”, *IEEE transactions on intelligent transportation systems*, Vol. 18, No. 11, 2017, pp. 3204-3217.

23. Samvelyan, M., Rashid, T., De Witt, C. S., Farquhar, G., Nardelli, N., Rudner, T. G., Chia-Man H., Philip H. S. T., Jakob F., and Whiteson, S., “The starcraft multi-agent challenge”, *arXiv*:1902.04043, 2019, pp. 1-14.

24. Kingma, D. P., and Ba, J., “Adam: A method for stochastic optimization”, *arXiv*:1412.6980, 2014, pp. 1-15.

25. Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M., “The arcade learning environment: An evaluation platform for general agents” , *Journal of Artificial Intelligence Research*, Vol.47 , 2013, pp. 253-279.

26. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., and Abbeel, P., “Domain randomization for transferring deep neural networks from simulation to the real world”, 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 23-30.