Expert Systems With Applications 260 (2025) 125405

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Evaluating driver-pedestrian interaction behavior in different environments via Markov-game-based inverse reinforcement learning

Gabriel Lanzaro *, Tarek Sayed

Department of Civil Engineering University of British Columbia 6250 Applied Science Lane, Vancouver, BC, Canada

ARTICLE INFO

Keywords:
Road user behavior
Vehicle-pedestrian interactions
Multi-agent policy
Inverse Reinforcement Learning
Traffic conflicts

ABSTRACT

The rapid advances in the technology of Autonomous Vehicles (AVs) requires effective collision avoidance systems that should be able to identify crash-risk situations and make AVs react fast. These systems can be modeled using Reinforcement Learning, where agents are assumed as rational decision-makers that take optimal decisions to achieve a goal (e.g., avoid a crash), which can be written in terms of a reward function. However, obtaining this reward function might be challenging as it deals with human behavior, and Inverse Reinforcement Learning can be implemented to recover these functions from actual road user trajectories in near-miss interactions. This approach provides reward functions that give insights into road user behavior and optimal policies that represent the best sequence of decisions to avoid a crash. However, road user behavior varies considerably depending on the traffic environment, and policies from one location might not be entirely transferable to different locations. This study utilizes Multi-agent Adversarial Inverse Reinforcement Learning (MA-AIRL) to simulate conflict trajectories of vehicle–pedestrian interactions in four different cities (i.e., Boston, Las Vegas, Pittsburgh, and Singapore). This model accounts for the competitive behavior in conflict interactions by explicitly considering that road users have an equilibrium between their intentions. Results show that the behavior is noticeably different depending on the environment. For example, the reward functions demonstrate that road users have various preferences when interacting with each other. Moreover, the MA-AIRL was reasonably able to replicate the evasive action mechanisms of drivers and pedestrians, but the accuracy of this prediction varied among the four cities, reflecting the difference in the environment. Finally, transferring agent behavior from one location to another led to increased risk levels. Therefore, to be implemented in AV collision systems, multi-agent policies should consider local behavioral characteristics as road user behavior plays a crucial role in safety.

1. Introduction

Technical advances have enabled the rapid development of autonomous vehicles (AVs), which are anticipated to provide considerable safety and mobility benefits (Duarte and Ratti, 2018; Talebpour et al., 2017; Zhu and Tasic, 2021). AVs require collision avoidance systems that are able to promptly identify crash-risk situations. These systems have much shorter perception-reaction times compared to human-driven vehicles which make AVs capable of reacting faster to risky interactions with other road users, which reduces the likelihood of crashes. Furthermore, several studies that investigated crash causes have noted that most crashes result from human-related factors (Pakgohar et al., 2011; Sayed et al., 1995), and AVs are expected to remove the unpredictability of human behavior and thus reduce the number of crashes caused by human drivers. However, caution should be taken as AVs may not entirely reflect the intricate nuances of human decision-making, especially under challenging and unpredicted scenarios. These vehicles have been developed under strict decision rules, which might be different from how humans actually make decisions. In addition, some recent incidents have contributed to increasing public distrust in AVs, such as a pedestrian fatality in Arizona, US (Penmetsa et al., 2021). In this situation, the vehicle only identified the pedestrian a few seconds prior to the crash, and the human driver, who was distracted and relying on the AV capabilities, did not take any action to avoid the crash. This situation raised two major issues regarding the safety of AV technologies. The first is related to the liability in AV crashes as AVs are programmed with algorithms based on black boxes that are difficult to interpret, which makes it challenging to attribute

* Corresponding author.
E-mail addresses: lanzaro@mail.ubc.ca (G. Lanzaro), tsayed@civil.ubc.ca (T. Sayed).

https://doi.org/10.1016/j.eswa.2024.125405
Received 12 July 2023; Received in revised form 9 September 2024; Accepted 14 September 2024
Available online 19 September 2024

responsibility for the crashes. Second, it showed that extensive research and more testing should be conducted to improve the safety of these technologies.

To maximize AV safety, manufacturers have put great effort into collision avoidance systems, which should be developed considering the optimal behavior in a situation where two or more users nearly collide. More specifically, these systems require a comprehensive understanding of crash avoidance mechanisms, i.e., the evasive actions that road users take to avoid a crash with each other. Traffic conflicts can then be used to properly demonstrate the different aspects of human behavior under risky situations. Conflicts have been used as surrogates for crash data, which have various limitations (Sayed and Zein, 1999; Zheng et al., 2021). First, crash data should be collected over a reasonable amount of time to provide accurate diagnostics, which represents an ethical dilemma as crashes should accrue first to be subsequently reduced. Second, crash data suffers from poor reliability as they tend to be underreported and may present low quality. Finally, crash data provides limited insights into crash failure mechanisms, as information regarding potential attempts to avoid the crash (e.g., evasive actions) is usually unavailable. Instead, conflicts have more frequent occurrence than crashes and enable understanding road users' behavior in situations where they are very close in time and space. Conflict indicators, such as the Post-encroachment Time (PET), can be utilized to differentiate between regular and unsafe interactions. Conflicts have normally been identified using on-site observation or by extracting road user trajectories from traffic cameras (Chai et al., 2023; Essa and Sayed, 2018; Saunier and Sayed, 2006). More recently, large-scale AV datasets (Caesar et al., 2021; Chang et al., 2019; Kamel et al., 2022) have been released, and they contain several microscopic road user behavioral information, such as positions and speeds, which enable obtaining conflict data for various road users.

This conflict data from human drivers can be used to model the interaction behavior of road users in conflict situations, and such models have the potential of being implemented in AVs. As AVs are expected to replicate human behavior, properly representing road users' actions in conflict interactions is important. For example, collision avoidance systems should define actions to be taken by AVs while interacting with other road users, and these actions must depend on several factors such as road users' speeds, orientations, and relative distances. This situation can be modeled in a Reinforcement Learning (RL) framework. In this technique, drivers and pedestrians can be assumed as agents who take logical decisions (i.e., actions) based on their current states with the objective to maximize their utilities, which can be represented by reward functions (Sutton and Barto, 2018). However, obtaining these functions might be difficult considering that human behavior cannot be easily quantified using simple mathematical formulations. Alternatively, Inverse Reinforcement Learning (IRL) enables estimating these reward functions from demonstrations (Abbeel and Ng, 2004; Ng and Russel, 2000), i.e., actual road user trajectories of conflict interactions. These reward functions can be utilized for two purposes. First, they can provide inferences into road user behavior in conflict situations, which is useful for understanding road user preferences while interacting with each other. Second, state-of-the-art RL algorithms (Song et al., 2018; Sutton and Barto, 2018; Yu et al., 2019) can be used to obtain optimal policies based on the reward functions, which represent the best actions to be taken in risky scenarios. These policies can then be implemented in AV collision avoidance systems. Additionally, they can be combined with other data-driven approaches to provide real-time safety estimates of the traffic environment. Particularly, the optimal policies can be used to predict road user trajectories at very short intervals, which makes it possible to obtain simulated traffic conflicts between vehicles and pedestrians. These conflicts can then be utilized to evaluate the risk of the environment and potentially estimate the number of crashes (Fu and Sayed, 2022; Zheng et al., 2021). Therefore, the IRL modeling framework can support proactive safety analysis by helping to anticipate potential crash situations before they actually occur.

However, road user behavior might vary considerably across different environments, making it difficult to directly apply modeling results from one location to another. To navigate diverse traffic environments, AVs must be trained considering datasets that reflect various infrastructure types (e.g., intersections, highways) and contexts (e.g., lane-based, non-lane-based). For example, for an effective collision avoidance system, AVs should understand what is considered to be a low enough relative minimum distance that would entail a sudden evasive action taken by either the vehicle itself or by the other road user (e.g., pedestrian, cyclist, other vehicles). This is especially relevant considering pedestrians, which have unpredictable behavior that is highly dependent on local conditions. Also, pedestrians can perceive interactions with regular human-driven vehicles with different degrees of severity (McIlroy et al., 2020; Tinella et al., 2022). Therefore, frameworks that consider the interactions between AVs and pedestrians, more specifically the optimal policy to be implemented in conflict situations, should be context-specific. This leads to the need for developing models considering vehicle–pedestrian interactions in various environments and evaluating how differently pedestrians react to standard human-driven vehicles.

This study aims to model vehicle–pedestrian conflict interactions in very different environments using Multi-agent Adversarial Inverse Reinforcement Learning (MA-AIRL). This technique enables considering that road users' interactions are based on an equilibrium concept, which is essential in conflict interactions as road users react to each other. This study uses data obtained from a large-scale AV dataset (Caesar et al., 2021) collected in four cities with very different driving characteristics: Boston, Las Vegas, Pittsburgh, and Singapore. This dataset provided accurate trajectory information about vehicles and pedestrians, such as their positions and speeds at very high frame rates. This study makes numerous contributions to the existing literature, such as (1) developing MA-AIRL models in different environments to evaluate how multi-agent policies can be applied in various locations (and how they can replicate road user behavior locally), (2) inferring differences in road user behavior using the reward functions by comparing them, (3) estimating the evasive action mechanisms of drivers and pedestrians in different environments, (4) using large-scale AV datasets to model road users, which shows the potential of utilizing these datasets for real-time policy implementation as they provide high-quality information at very short time intervals, and (5) evaluating the transferability of road user behavior from different locations and its influence on the severity of the interactions, which might impact AV deployment in various environments. Finally, this work shows that multi-agent policies should be developed for specific locations to account for local behavioral characteristics and thus properly incorporate the behavior of other road users (e.g., pedestrians, human drivers).

2. Literature review

Road user behavior plays a significant factor in road safety, but few studies have focused on comparing and evaluating the safety of different environments. Previous studies conducted in several countries found that different conflict indicators reflect different levels of severity (Tageldin et al., 2017; Tageldin and Sayed, 2019). Also, Guo et al. (2019) evaluated if microsimulation calibration parameters could be directly transferred from Canada to Australia, and results showed that this could result in inaccurate safety assessments. Moreover, Jiang et al. (2015) conducted a comparative analysis of pedestrian walking behavior in unsignalized midblock crosswalks in China and Germany, and significant differences were found in pedestrian speed and wait time. Multiple studies evaluated how safety performance functions are transferable to other jurisdictions (Barbosa et al., 2014; Feng et al., 2020; La Torre et al., 2022). However, these functions are mostly based on crash occurrence and rarely reflect behavior mechanisms.

Other studies investigated how pedestrians and drivers perceive safety in different countries. For example, Nordfjærn and Zavareh

(2016) noted that Iranians and Pakistanis report different levels of aggressive behaviors, which shows that traffic enforcement needs to consider cultural factors. Tinella et al. (2022) investigated how personality traits influenced distracted driving in Australia and Italy, and the results showed that mind-wandering tendencies (i.e., the tendency of having thoughts not linked to the current environment) varied across personality traits and countries. McIlroy et al. (2020) applied a questionnaire in six different countries to measure how attitudes toward traffic violations were associated with risky pedestrian behaviors. The study found that this association might be stronger depending on the country. Furthermore, pedestrians in more organized traffic environments tend to report safer behavior, whereas, in less organized and mixed traffic environments, pedestrians have shown greater risk awareness (Nordfjærn et al., 2014).

Pedestrian safety is highly associated with vehicle behavior and, in order to better represent the interactions between vehicles and pedestrians, several previous studies have made attempts to model their actions in conflict situations. For example, Nasernejad et al. (2021) used a Gaussian process to obtain pedestrians' utilities in conflict interactions with vehicles. This approach considered the pedestrian as an intelligent agent capable of taking decisions, which reacted to an environment that was fixed. The framework was later expanded to account for both road users taking actions simultaneously (Nasernejad et al., 2022). Waizman et al. (2015) created a simulation tool to mimic road user behavior in vehicle–pedestrian interactions, and this tool was based on three main actions: acceleration, deceleration, and steering. Chao et al. (2015) developed a model to simulate vehicle and pedestrian behavior using a gap acceptance model. Results showed that both road users took evasive actions while interacting with each other. Zhang and Fu (2022) used a social force method to represent vehicle–pedestrian behavior under complex evacuation scenarios, such as parking lots. Lu et al. (2016) introduced a cellular automata model for driver-pedestrian interactions, and the model parameters were calibrated using real-world data from unsignalized crosswalks. Zeng et al. (2017) used a hybrid approach to model pedestrian dynamics while interacting with vehicles. The model was calibrated using a genetic algorithm.

Properly understanding road users' behavior is important for predicting their trajectories, which can support crash prevention. However, the assumptions that govern vehicle–pedestrian interactions are considerably different given the unstructured environments that pedestrians navigate (Golchoubian et al., 2023; Tolksdorf et al., 2023). Still, various studies have used actual trajectory data to develop motion prediction algorithms that can estimate the likelihood of a traffic conflict or a crash (Formosa et al., 2020; Parada et al., 2021). In an emerging environment where AVs will interact with other road users (e.g., human-driven vehicles and pedestrians), trajectory prediction algorithms can also provide information to forecast future traffic conditions, such as traffic states depending on the approaching speeds of the vehicles. As previous studies have shown that the crash risk can be estimated in very short-time intervals, such as signal cycles (Fu and Sayed, 2022; Zheng and Sayed, 2020), an accurate prediction of road user intentions in conflict situations is essential to estimate the safety of an environment. Previous studies have obtained reasonably great crash estimates with low sample sizes, and simulation studies can be used to further improve the results without making actual implementations in the field (Farah and Azevedo, 2017; Wang et al., 2018).

Additionally, modeling road user behavior involves considering the actions of several agents simultaneously given that road users are expected to react to each other whilst maximizing their preferences. Multi-agent approaches can be used in this regard as they enable modeling different road users and consider the equilibrium between their intentions. For example, Markov Games (MGs) are a promising approach to model agents as rational decision-makers that take decisions based on their utilities. This approach has been employed in numerous applications, such as transportation systems (Nasernejad et al., 2022; Shou et al., 2022), business (Georgila et al., 2014; Wu et al., 2021), and games (Rossato et al., 2020; Silver et al., 2017). MGs can be integrated into IRL frameworks to extract inferences from agents' behavior. This technique has been successively used to understand drivers' intentions (Nasernejad et al., 2022; You et al., 2019) using reward functions, which can subsequently be utilized to estimate drivers' optimal policies (i.e., best sequences of decisions) and thus implement them in AV technologies. These reward functions are highly non-linear to represent road user behavior, and several previous studies used adversarial neural networks to describe road users' intentions while interacting with each other. For example, Alsaleh and Sayed (2021) introduced MA-AIRL models to represent the conflicting behavior between cyclists and pedestrians in shared spaces, where interactions between active transportation users become more frequent. Subsequently, Lanzaro et al. (2022a) extended the methodology to incorporate the behavior of motorcyclists, which take numerous actions such as swerving and decelerating to avoid crashes with pedestrians. Similarly, Nasernejad et al. (2022) conducted a study in Shanghai to evaluate the microscopic behavior of road users in vehicle–pedestrian interactions. These studies were able to accurately replicate road user behavior, considering their evasive actions, conflict indicators, and road user trajectories, especially when compared to single-agent models that assume that rational agents react to an environment that is fixed.

However, most of the existing work focused on understanding the microscopic behavior of road users were limited to a single environment (i.e., a single location). Previous research has shown that pedestrian behavior varies considerably across different locations (Jiang et al., 2015; Tageldin and Sayed, 2019). This is extremely important considering the emerging era of AVs, which have been developed using advanced algorithms that try to mimic actual drivers while interacting with different road users, such as pedestrians. As road user behavior is highly environment-dependent, directly implementing the policies of one AV in a new environment might present challenges related to transferability and therefore impact the safety of AV operations. In the IRL framework, the reward functions are context-specific and are expected to be different, which affects the decisions taken by AVs. This work bridges this gap by modeling road users with MA-AIRL in four different environments, which provides insights for the safe deployment of AVs in diverse locations with driving conditions that might not be related.

3. Data collection

3.1. The nuPlan dataset

This research uses a subset of the nuPlan dataset (Caesar et al., 2021), which is a large-scale autonomous driving dataset. It consists of approximately 1500 h of driving data from Boston, Las Vegas, Pittsburgh, and Singapore. These cities contain a wide variety of different driving scenarios, which are useful for motion prediction benchmarks. The Boston area is represented by an urban environment in South Boston with a high quantity of cyclists and pedestrians. The region contains several cycle paths and wide crosswalks, which encourage active transportation. The Las Vegas dataset was collected in the Las Vegas Strip, where many famous casinos, resort hotels, and sights are situated. In this location, vehicles travel through various casino pick-up and drop-off points, and the environment is considerably busy with multiple pedestrians and several lanes in each direction. The Pittsburgh area, including both South Side Flats and Hazelwood, is less dense than the other American cities. It has narrower lanes and many locations where on-street parking is permitted. Finally, Singapore presents left-hand traffic and vehicles travel in Queenstown or more specifically, in One North and UTown with many university residences and high-tech buildings. Data were collected on several days during daytime and under clear weather conditions. For the Las Vegas dataset, data were collected in May 2021. For the Boston, Pittsburgh, and Singapore datasets, data were collected from August 2021 to October 2021.

Table 1
Subset of the nuPlan dataset used in this study.

City	Detected pedestrians	Detected vehicles	Scenes	Hours (h)	Vehicle-pedestrian conflicts
Boston	237,994	551,230	1523	87.70	855
Las Vegas	870,706	532,365	1116	66.49	911
Pittsburgh	116,360	447,300	1494	119.25	208
Singapore	184,843	228,356	2394	147.99	631

Vehicle and pedestrian trajectories were obtained from instrumented vehicles, which were equipped with several sensors: lidar, cameras, a global navigation satellite system (GNSS), and an inertial measurement unit (IMU). The lidar sensors were able to collect data from the surrounding environment with a frequency of 20 Hz and a maximum range of 200 m. The identified objects were labeled into seven categories: vehicle, bicycle, pedestrian, traffic cone, barrier, construction zone sign, and other generic objects. The dataset used in this study considers one major source: the lidar sensor, which obtained information regarding the surrounding environment (i.e., other vehicles and pedestrians), including their coordinates and speeds. Therefore, the dataset contained accurate trajectory information about vehicles and pedestrians, which allowed identifying conflict interactions between them.

3.2. Data description

Table 1 presents the subset of the nuPlan dataset used in this study. Vehicle-pedestrian interactions with PET lower than 3 s were denoted as conflicts, as aligned with previous studies (Chen et al., 2019; Kassim et al., 2014). This conflict indicator is the time difference between the instant where the first road user (e.g., the vehicle) leaves the conflict point and the second road user (e.g., the pedestrian) reaches it. Lower PET values indicate that a crash was avoided by the smallest margin, which represents more severe interactions. The dataset contained several scenes, where the AV had a clear goal, i.e., travel from an origin to a destination without intermediate stops. On these trips, the AV collected data about other road users.

Fig. 1 displays the PET histograms for each city. In all cities, the figure shows that the PET distributions are left-skewed, i.e., the quantity of more severe interactions (lower PETs) is considerably lower than the quantity of less severe interactions (greater PETs).

Several variables were calculated from the road user trajectories, including speeds, accelerations, distances, and yaw rates. These variables were obtained considering different studies that modeled drivers and pedestrians in conflict situations (Nasernejad et al., 2022; Waizman et al., 2015; Zhang and Fu, 2022) and that used inverse reinforcement learning to model road users (Alsaleh and Sayed, 2020; Lanzaro et al., 2022b). Table 2 presents the means of these variables, along with their standard deviations, for each city. Fig. 2 depicts a graphical representation of these variables.

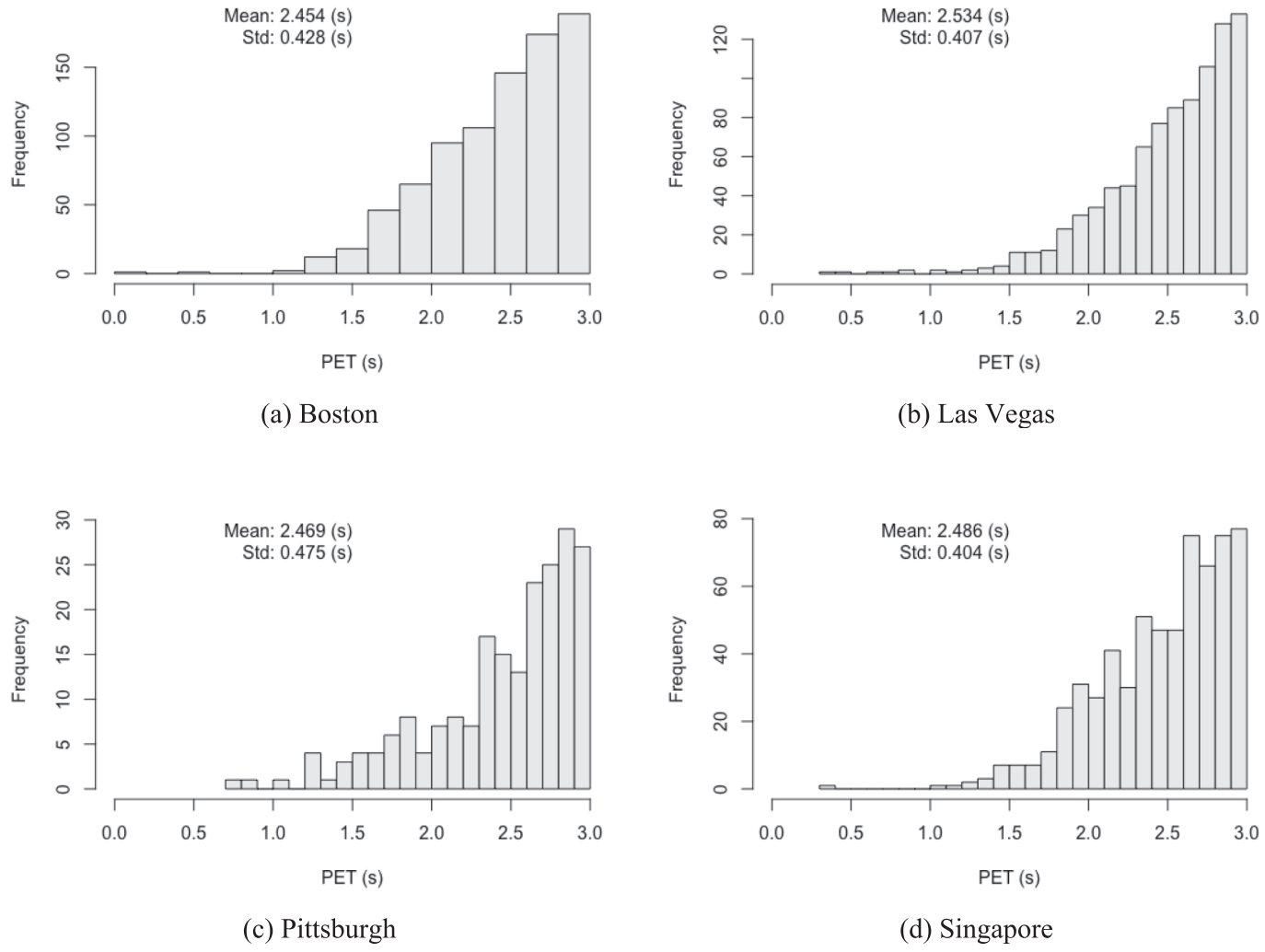


Fig. 1. PET histograms for vehicle-pedestrian interactions in 4 different cities.

Table 2
Mean variables for vehicle-pedestrian interactions in 4 different cities.

Variable	Boston		Las Vegas		Pittsburgh		Singapore	
	Pedestrian	Vehicle	Pedestrian	Vehicle	Pedestrian	Vehicle	Pedestrian	Vehicle
Speed(m/s)	1.454 [0.765]	6.236 [2.972]	1.246 [0.646]	4.692 [2.683]	1.478 [0.803]	5.846 [3.001]	1.182 [0.629]	6.957 [3.219]
Speed Difference (m/s)	4.782 [3.116]	4.782 [3.116]	3.446 [2.702]	3.446 [2.702]	4.368 [3.160]	4.368 [3.160]	5.774 [3.343]	5.774 [3.343]
Distance to Destination (m)	4.455 [3.028]	18.443 [13.581]	3.795 [2.481]	14.672 [10.340]	5.040 [4.766]	17.753 [12.947]	3.627 [2.250]	19.783 [15.008]
Acceleration (m/s^2)	0.114 [1.400]	0.317 [3.275]	0.076 [1.524]	0.511 [3.895]	0.100 [1.289]	0.279 [2.962]	0.107 [1.374]	0.083 [2.988]
Yaw Rate(rad/s)	0.001 [3.537]	−0.033 [2.342]	−0.008 [6.497]	0.053 [2.217]	−0.014 [3.790]	0.018 [2.308]	0.006 [4.296]	0.015 [3.252]
Longitudinal Distance (m)	0.921 [6.729]	−1.066 [11.406]	0.569 [5.272]	−1.221 [8.530]	1.287 [6.897]	−1.299 [10.924]	1.445 [7.296]	−1.047 [12.821]
Lateral Distance (m)	7.868 [6.456]	3.663 [2.251]	5.974 [5.042]	3.302 [2.020]	7.298 [6.061]	3.618 [2.262]	8.511 [7.149]	3.025 [2.060]
Interaction Angle (rad)	0.197 [1.717]	0.197 [1.717]	0.033 [1.713]	0.033 [1.713]	0.074 [1.724]	0.074 [1.724]	−0.158 [1.755]	−0.158 [1.755]

* The numbers in brackets indicate the standard deviations.

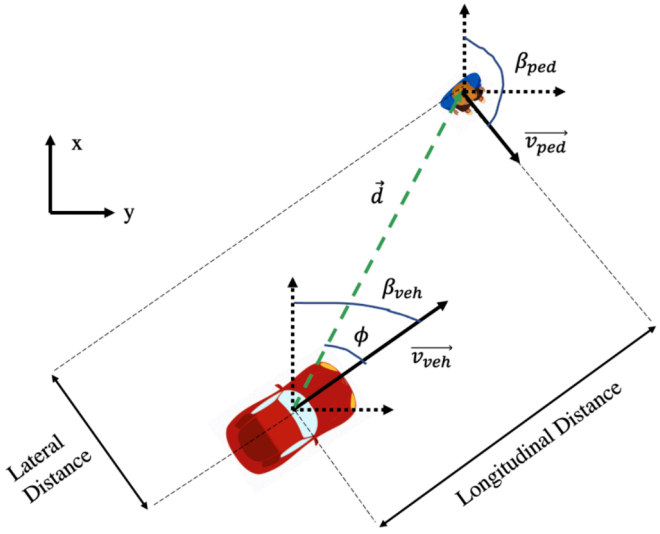


Fig. 2. Graphical representation of distances and angles between vehicles and pedestrians.

Three distance variables are used in this study: lateral distance, longitudinal distance, and distance to destination. The lateral and the longitudinal distances between vehicles and pedestrians are calculated with respect to an angle ϕ, as consistent with Fig. 2. This angle is defined between the vehicle speed vector $\overrightarrow{v_{veh}}$ and the distance between them $\vec{d}$, as indicated in Eq. (1). Thus, the longitudinal and the lateral distance are obtained as shown in Eq. (2) and Eq. (3), respectively. Using this reference system, positive longitudinal distance values imply that the vehicle is before the pedestrian, whereas negative values show that the vehicle has overtaken the pedestrian. Also, the distance to destination variable is indicates the road users' intention of motion and can be considered a guidance variable. Such variable is calculated as the distance between the current position and the expected destination.

$$\phi = \cos^{-1}\left(\frac{\overrightarrow{v_{veh}}\cdot\vec{d}}{\|\overrightarrow{v_{veh}}\|\cdot\|\vec{d}\|}\right) \tag{1}$$

$$d_{long} = \|\vec{d}\|\cdot\cos\phi \tag{2}$$

$$d_{lat} = \|\vec{d}\|\cdot\sin\phi \tag{3}$$

where d_{long} and d_{lat} are the longitudinal and lateral distances, respectively.

Speed variables are also considered. In addition to both road users' individual speeds, the speed difference is used, and positive values indicate that the vehicle is faster than the pedestrian. Moreover, the interaction angle is the absolute difference between the heading angles of the vehicle and of the pedestrian. Finally, the road user actions are defined in terms of the acceleration and the yaw rate, which are the differentiation of the speed profile and the variation of the road users' heading angle, as defined in Eq. (4) and Eq. (5), respectively.

$$a(t) = \frac{dV}{dt} \tag{4}$$

$$y(t) = \frac{d\beta}{dt} \tag{5}$$

where V is the speed profile, β is the road user heading angle, and $a(t)$ and $y(t)$ are the acceleration and the yaw rates at each time step t.

The road user trajectories were obtained at each time frame (1/20 s), and the Savitzky-Golay filter was applied to reduce data noise (Savitzky and Golay, 1964). This filter was used to smooth the positions, the speed profiles, and the acceleration profiles of each trajectory. The extracted conflict interactions were randomly categorized into two groups: the training dataset and the validation dataset. For each city, 75 % of the total number of conflict interactions were used for training, whereas 25 % were destinated for validation.

4. Methodology

This paper uses the multi-agent adversarial inverse reinforcement learning (MA-AIRL) approach to estimate driver and pedestrian reward functions and their decision sequences (i.e., optimal policies). The MA-AIRL framework employs the logistic stochastic best response equilibrium to account for conflicting preferences. In addition, the model uses adversarial neural networks to estimate the reward function, which enable obtaining behavioral functions that are highly nonlinear and complex in nature. The road users' policies are determined with the deep reinforcement learning-based multi-agent actor-critic algorithm with Kronecker factors (MACK DRL). Finally, simulation platforms were developed to simulate road user behavior using the estimated policies.

4.1. Markov game

In single-agent models, RL frameworks usually consider Markov Decision Processes (MDPs) to represent the sequences of decisions of an

agent that constantly interacts with the environment. A Markov Game (MG) can be used instead to account for various agents (e.g., drivers and pedestrians) that interact with each other and the environment. In this approach, road users have preferences, which are represented by reward functions, and learn their optimal decisions by maximizing these functions (Littman, 1994; Sutton and Barto, 2018). A MG can be characterized by $(\mathscr{S}, \mathscr{A}, \mathscr{P}, \zeta, \mathscr{N}, \gamma)$, where $\mathscr{S}$ and $\mathscr{A}$ stand for sets of states and actions, $\mathscr{P}$ are the probabilities of changing from states s^t to s^{t+1} if any action $a_1, \cdots, a_{\mathscr{N}}$ is taken, ζ represents the initial state distribution, $\mathscr{N}$ is the number of agents, and γ represents the discount factor, which is the influence of future rewards on the present state. It should be noted that, in this paper, bold variables are used to denote every agent of the MG framework, whereas the notation $-i$ indicates all agents except i. In this model, agents should maximize their expected returns, as shown in Eq. (6).

$$ExpRet_i^{\pi,\boldsymbol{\pi}_{-i}}(s_t, \boldsymbol{a}_t) = \mathbb{E}_{s^{t+1:T}, \boldsymbol{a}^{t+1:T}}\left[\sum_{l \geq t} \gamma^{l-t} r_i\left(s^l, \boldsymbol{a}^l\right) | s_t, \boldsymbol{a}_t, \boldsymbol{\pi}\right] \tag{6}$$

where r indicates the reward function, π represents the stochastic policy, and $l \in \{t, T\}$ is the time step.

In this study, pedestrians and drivers are modeled in a MG setting with a clear representation of states and actions. For both road users, their states are defined in terms of six variables: speed, speed difference, interaction angle, longitudinal distance, lateral distance, and distance to destination. In addition, the actions of the road users are characterized by the acceleration and yaw rate profiles, i.e., their changes in speed and orientation, respectively. The discount factor is presumed to be equal to 0.975, and this is consistent with previous research that modeled vulnerable road users in conflict situations (Alsaleh and Sayed, 2020; Lanzaro et al., 2022b; Nasernejad et al., 2021).

Modeling road user behavior using MGs is more realistic than using MDPs because it allows considering an equilibrium concept that relates both road users' preferences. For example, driver decisions are based not only on its interaction with a fixed environment (e.g., signalized intersection) but also on the decisions of other agents (e.g., pedestrians). This is vital in conflict situations as road users take decisions to avoid crashes with each other. In MGs, both agents aim to maximize their sums of cumulative discounted rewards; however, their intentions might be better represented in cooperative or competitive formulations. In a cooperative framework, both agents share a final objective and try to achieve it by collaborating with each other. Conversely, in a competitive framework, the agents possess conflicting intentions and prefer to maximize their individual expected rewards (Zhang et al., 2019). These scenarios can be easily modeled using the agents' reward functions. In the cooperative setting, the long-term reward sum of both agents is maximized, whereas, in the competitive setting, the sum of rewards is expected to approach zero.

This structure of multiple agents interacting simultaneously requires additional frameworks that account for the equilibrium between the agents' decisions. Standard equilibrium concepts, such as the Nash equilibrium and the correlated equilibrium, are not capable of replicating suboptimal behavior (Alsaleh and Sayed, 2022). These approaches may either consider that agents always take optimal decisions or prevents agents from achieving higher rewards while changing their policies (Yu et al., 2019). The logistic quantal response equilibrium (LQRE) enables agents to select stochastic policies. These policies are then obtained in situations where agents have higher probabilities of achieving higher returns. The formulation to obtain this policy is shown in Eq. (7).

$$\pi_i(a_i|s) = \frac{\exp\left(ExpRet_i^{\pi}(s, a_i, \boldsymbol{a}_{-i})\right)}{\sum_{a_i'} \exp\left(ExpRet_i^{\pi}(s, a_i', \boldsymbol{a}_{-i})\right)} \tag{7}$$

where $\boldsymbol{a}_{-i}$ is the action of all agents but i.

To allow the proper representation of multi-agent reward functions, the logistic stochastic best response equilibrium (LSBRE) can be introduced. The LSBRE uses a best-response approach based on entropy regularization (Yu et al., 2019). In this framework, the reward functions for each agent are defined as $r_i : \mathscr{A} \times \cdots \times \mathscr{A}_n$ for n agents and, considering $\{\pi_i^t\}_{t=1}^T$ policies dependent on time, the value function is described in Eq. (8).

$$\begin{aligned} Q_i^{\boldsymbol{\pi}^{t+1:T}}\left(s^t, a_i^t, \boldsymbol{a}_{-i}^T\right) = & r_i\left(s^t, a_i^t, \boldsymbol{a}_{-i}^T\right) + \mathbb{E}_{s^{t+1}\ P(.|s^t, \boldsymbol{a}^t)}\Big[\mathscr{H}\left(\pi_i^{t+1}(.|s^{t+1})\right) \\ & + \mathbb{E}_{\boldsymbol{a}^{t+1}\ \boldsymbol{\pi}^{t+1}(.|s^{t+1})}\left[Q_i^{\boldsymbol{\pi}^{t+2:T}}\left(s^{t+1}, \boldsymbol{a}^{t+1}\right)\right]\Big] \end{aligned} \tag{8}$$

where P_i are the transition probabilities and $\mathscr{H} = \mathbb{E}[-\log\pi(a|s)]$ represents the policy entropy.

A Markov chain is then considered to enable transitions of consecutive states. The state at a k th step is given by $z^{(k)} = (z_1, \cdots, z_N)^{(k)}$, where the random variables $z_i^{(k)}$ assume values from the actions $\mathscr{A}$, and the LSBRE is applied by considering T stochastic policies in the action space, as shown in Eq. (9). Then, the joint policies are expressed in Eq. (10).

$$z_i^{t,(k+1)}(s^t)\ P_i^t\left(a_i^t | \boldsymbol{a}_{-i}^t = \boldsymbol{z}_{-i}^{t,(k)}(s^t), s^t\right) = \frac{\exp\left(Q_i^{\boldsymbol{\pi}^{t+1:T}}\left(s^t, a_i^t, \boldsymbol{z}_{-i}^{t,(k)}(s^t)\right)\right)}{\sum_{a_i'} \exp\left(Q_i^{\boldsymbol{\pi}^{t+1:T}}\left(s^t, a_i^t, \boldsymbol{z}_{-i}^{t,(k)}(s^t)\right)\right)} \tag{9}$$

$$\boldsymbol{\pi}^t(a_1, \cdots, a_N | s^t) = P\left(\bigcap_i \{z_i^t(s^t) = a_i\}\right) \tag{10}$$

where $\boldsymbol{z}_{-i}$ are the transitions for every agent but i.

4.2. Multi-agent inverse reinforcement learning

In a MG, the reward functions represent the agents' utilities, which they maximize over time. These functions are formulated a priori based on the agents' intentions in most settings. However, specifying reward functions may be difficult in situations where many variables influence the agents' behavior and there is no clear mathematical formulation. Inverse Reinforcement Learning (IRL) (Abbeel and Ng, 2004; Ng and Russel, 2000) can be used to learn the reward functions from road user trajectories. This approach considers M demonstrations (trajectories) $\mathscr{D} = \{\tau_j\}_{j=1}^M$, and each trajectory represents a sequence of both states and actions $\tau_j = \left\{\left(s_j^t, a_j^t\right)\right\}_{t=1}^T$ of lifetime T. Furthermore, obtaining these reward functions from demonstrations without any additional considerations may be inappropriate as agents do not necessarily take optimal decisions. To account for potential suboptimal behavior, the maximum entropy framework, introduced by Ziebart et al. (2008), can be applied. In this technique, agents select state-action sequences based on the probability of choosing paths with higher expected rewards. Eq. (11) illustrates the likelihood function to be maximized in the maximum entropy IRL framework.

$$p_w(\tau) \propto \frac{1}{Z_w} \exp\left(\sum_{t=1}^T r_w(s^t, a^t)\right)\left[\zeta(s^1)\prod_{t=1}^T P(s^{t+1}|s^t, a^t)\right] \tag{11}$$

where $p_w(\tau)$ is the probability of selecting a path based on a set of feature weights w, $\zeta(s^1)$ is the state distribution for the initial state, and Z_w stands for the partition function.

Several different formulations of reward functions can be used to represent behavior, including a linear combination of future weights (Abbeel and Ng, 2004; Ziebart et al., 2008), Gaussian processes (Levine and Koltun, 2012), and neural networks (Fu et al., 2017; Yu et al., 2019). In this research, generative adversarial neural networks are utilized to model drivers' and pedestrians' behavior in conflict interactions. This approach considers two major structures: a discriminator and a generator. The generator samples data distributions, which should be input into the discriminator. In turn, the discriminator is responsible for

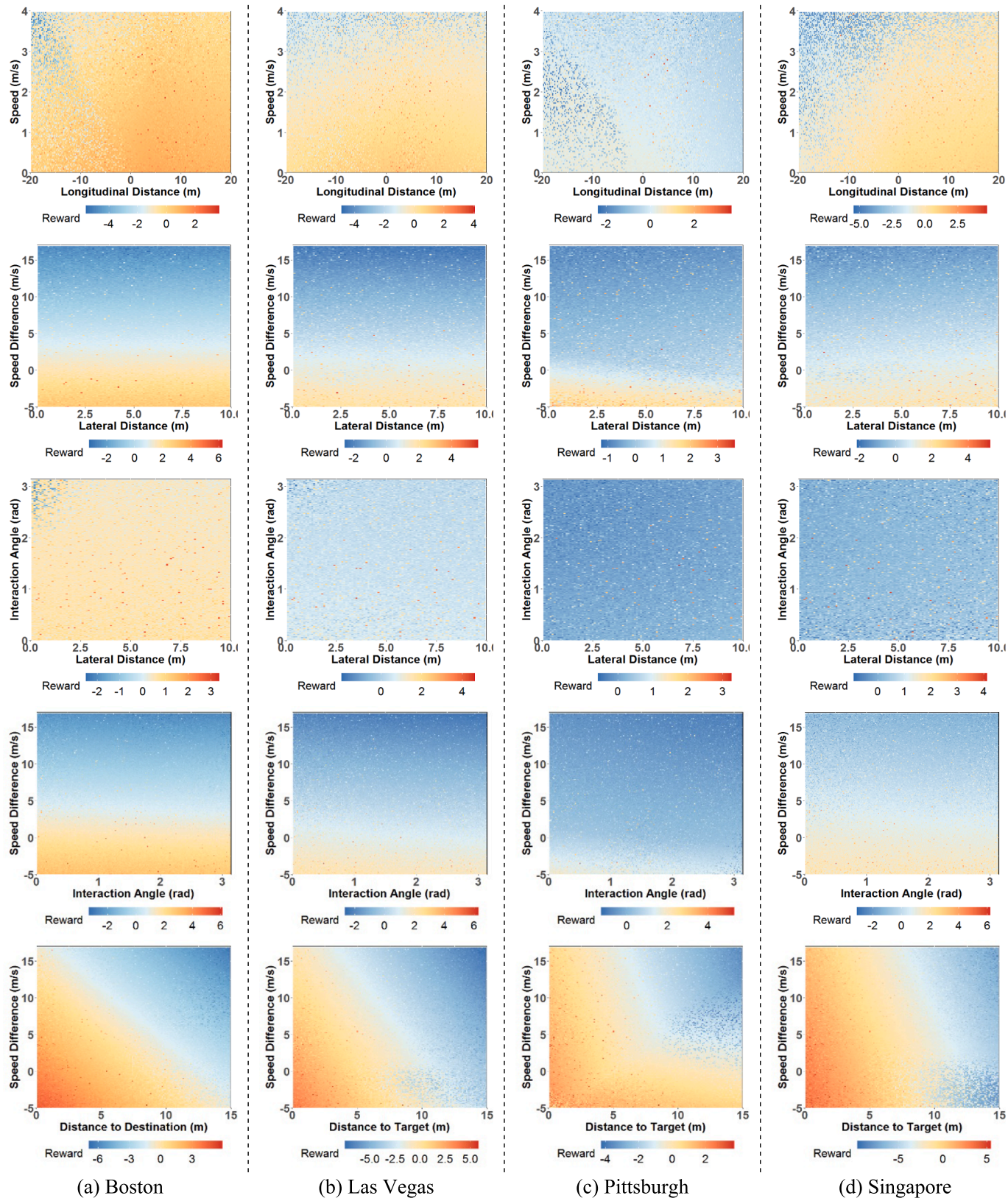


Fig. 3. Pedestrian Multi-Agent Reward Functions (PRF) for 4 different cities.

classifying if the data distributions were obtained from the expert demonstrations or were artificially generated. Both the discriminator and the generator are trained concurrently until the discriminator is not able to differentiate between the outputs of the generator and actual data. At this point, the generator would be able to sample data distributions that sufficiently approach the structure of the underlying data distribution (Goodfellow et al., 2014). The aim of this approach is related to maximizing the likelihood given by the discriminator

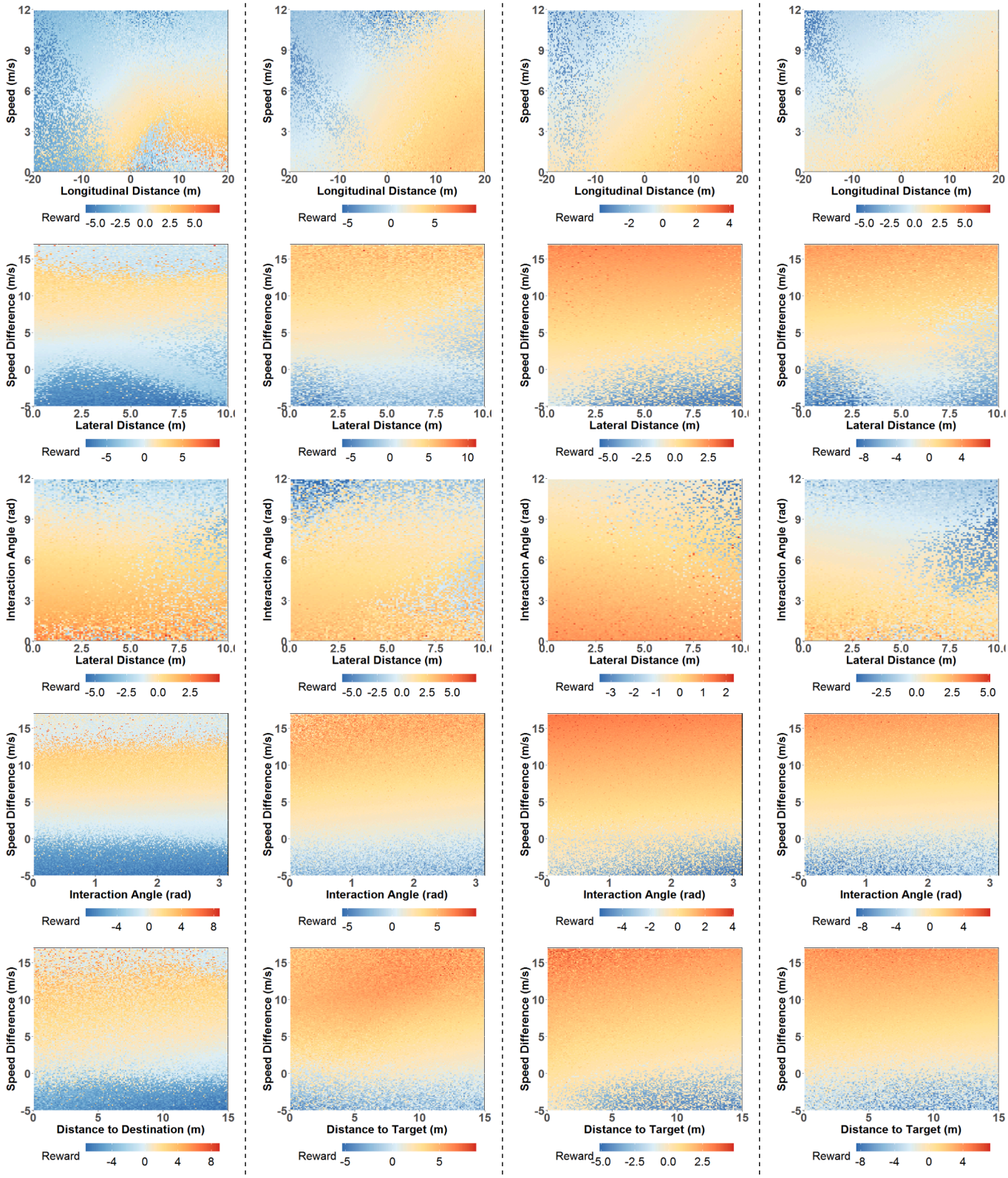


Fig. 4. Driver Multi-Agent Reward Functions (DRF) for 4 different cities.

function, as shown in Eq. (12), where the discriminator is expressed in Eq. (13).

$$\mathbb{E}_{\pi}\left[\sum_{i=1}^{N}\log(D_{wi}(s,\boldsymbol{a}))\right]+\mathbb{E}_{q_\theta}\left[\sum_{i=1}^{N}\log(1-D_{wi}(s,\boldsymbol{a}))\right] \tag{12}$$

$$D_w(s,a)=\frac{\exp\left(r_{w_i}(s,\boldsymbol{a})\right)}{\exp\left(r_{w_i}(s,\boldsymbol{a})+q_{\theta_i}(a_i|s)\right)} \tag{13}$$

where D_w is the discriminator, $r_w(s,a)$ represents the reward function, and $q_\theta(a|s)$ stands for the adaptive sampler (i.e., outputs of the generator and inputs of the discriminator), which is parametrized by weights θ. The partition function, which is difficult to be calculated in high-dimensional or continuous environments, can be computed by sampling q_θ effectively (Finn et al., 2016). Finally, the weights θ are obtained using the discriminator function, as given by Eq. (14).

$$\mathbb{E}_{q_\theta}\left[\sum_{i=1}^{N}\log(D_{wi}(s,\boldsymbol{a}))-\log(1-D_{w_i}(s,\boldsymbol{a}))\right] \tag{14}$$

The reward function may be subject to ambiguity, where various reward functions can be used to describe the same behavior. To overcome this limitation, robust rewards should be recovered (Fu et al., 2017), as shown in Eq. (15). This approach enables recovering reward functions that are much more associated with the ground-truth rewards.

$$r_{w,\varphi}\left(s^t,a^t,s^{t+1}\right)=g_w(s^t,a^t)+\gamma h_\varphi\left(s^{t+1}\right)-h_\varphi(s^t) \tag{15}$$

where g_w represents the reward estimator and h_φ stands for the shaping function.

In summary, the algorithm works as follows. The discriminator is updated with the reward function estimates. In addition, the policy is updated as the sampling distribution is continuously improved. The adaptive samplers should minimize the divergence between the trajectories induced by the reward function and the expert trajectory distribution (i.e., demonstrations).

4.3. Actor-critic deep reinforcement learning

The deep reinforcement learning-based multi-agent actor-critic algorithm with Kronecker factors (MACK DRL) (Song et al., 2018; Wu et al., 2017) can be considered to learn the optimal policies. The actor-critic framework is constituted by two major structures: an actor component and a critic component. The actor component estimates the policy, whereas the critic component evaluates the policy by applying the value function recursively. Neural networks can be used to represent both components as these structures can handle non-linear, complex, and uncertain policies. The gradient update for both critic and actor components is based on the Kronecker-factored approximation, which improves the efficiency of the algorithm. The policy is updated using an advantage function, as expressed in Eq. (16).

$$A^{\pi_i}_{\phi_i}(s,a_t)=\sum_{j=0}^{k-1}\gamma^j r\left(s_{t+j},\boldsymbol{a}_{t+j}\right)+\gamma^k V^{\pi_i}_{\phi_i}\left(s_{t+k},\boldsymbol{a}_{-i,t+k}\right)-V^{\pi_i}_{\phi_i}\left(s_t,\boldsymbol{a}_{-i,t}\right) \tag{16}$$

where $V^{\pi_i}_{\phi_i}$ is the value function baseline. The ϕ parameters, which are related to the critic component, are continuously updated with the objective to minimize the numerical difference between expected and predicted returns. Such function is used to estimate the θ parameters of the generator, which are associated with the actor component (Mnih et al., 2016), as given by Eq. (17). The Kronecker-factored curvature is then utilized to obtain the optimized weights of the neural networks.

$$\mathbb{E}\left[\sum_{t=0}^{\infty}\nabla_{\theta_i}\log\pi_{\theta_i}(a_t|s_t)A^{\pi_i}_{\phi_i}(s,a_t)\right] \tag{17}$$

4.4. Simulation platform

With the estimated road user policies, a simulation tool was developed. This tool was initialized with the initial positions, speeds, and heading angles of both road users, which enabled calculating their state variables (i.e., a combination of lateral distances, longitudinal distances, speed differences, speeds, interaction angles, and distances to destination). Then, the subsequent actions were obtained by sampling the optimal policy. These actions were utilized to change both road users' positions, speeds, and heading angles at each time frame. This procedure was conducted until the trajectories' original number of frames was reached.

This simulation tool was utilized for two different purposes. First, it was considered to evaluate how accurately the model replicated road user behavior depending on the environment, i.e., to verify if the model could learn the intrinsic behavioral nuances of each environment. Second, the tool was used to investigate how the behavior of agents in one environment could impact another environment if no additional considerations were taken. In other words, the tool was used to measure the effects of simply switching policies from one environment to another.

5. Results and discussion

Different MA-AIRL models were developed for each city (i.e., Boston, Las Vegas, Pittsburgh, and Singapore). The model outputs included the optimal policies and the reward functions for both interacting road users. The reward functions were first considered to extract inferences into driver and pedestrian behavior in conflict interactions for each city. Second, the optimal policies were used as inputs for the simulation tool to obtain simulated road user trajectories in different environments.

5.1. Reward function analysis

The reward functions were estimated for each city and for both road user types. As the MA-AIRL framework considers that road users behave according to an equilibrium concept, the algorithm estimates the reward functions for drivers and pedestrians simultaneously while the model is trained. Fig. 3 and Fig. 4 show the reward functions for pedestrians and for drivers, respectively, which are used to make inferences about their behaviors while interacting with each other. The reward functions are represented in terms of bivariate state variables with the remaining state features at their mean values. These functions are associated with stronger preferences; thus, higher rewards imply that road users prefer a particular state in comparison to others.

The pedestrian reward function (PRF), plotted with the speed and longitudinal distance variables, shows the trend of reducing the speed in the conflict surroundings. For example, in Boston, there is a stronger preference for reducing the speed immediately before the conflict (positive longitudinal distance values) when compared to Singapore and Las Vegas. In Pittsburgh, the higher rewards tend to be more dispersed, which shows that pedestrians in general keep their speeds with no sudden changes when approaching drivers. Furthermore, the PRF, plotted with the interaction angle and speed difference variables, shows that pedestrians prefer to be faster than drivers. This behavior is consistent across different interaction angles, except for Pittsburgh. In this city, pedestrians prefer even lower speeds as they lose sight of the vehicles (i.e., when the interaction angle diverges from zero). Moreover, the PRF, plotted with the speed difference and distance to destination variables, shows that road users have a higher preference to arrive at their destinations with negative speed differences (i.e.., with higher speeds than vehicles); however, pedestrians tend to be more tolerant of this speed difference depending on the location. For example, Singapore presents the highest variability in speed difference values when the distance to destination approaches zero. This is different from Boston and Las Vegas, which have different degrees of tolerance. Finally, in

Table 3
MA-AIRL model results for 4 different cities.

City	Explained variance		Loss		
	Driver	Pedestrian	Driver	Pedestrian	Total
Boston	0.803	0.850	1.643	1.241	2.883
Las Vegas	0.793	0.846	1.419	1.072	2.492
Pittsburgh	0.626	0.747	2.344	1.752	4.059
Singapore	0.553	0.812	2.369	1.545	3.915

Pittsburgh, this speed difference seems to be of lower relevance at low distance to destination values.

The driver reward function (DRF), plotted with the speed difference and interaction angle variables, shows that, in general, drivers prefer high speed differences (i.e., greater speed than pedestrians) regardless of the interaction angle between them. However, the preference for this speed difference varies depending on the city. For example, this difference is more pronounced in Pittsburgh when compared to Boston. In addition, the VRF, plotted with the speed and longitudinal distance variables, demonstrates that drivers prefer low speeds around the conflict point (both immediately before and after the conflict). However, driver speeds immediately before the conflict tend to be slightly higher in Boston. Moreover, the VRF, plotted with the lateral distance and interaction angle variables, shows a greater preference for lower lateral distance variables, and this preference seems to be mostly indifferent to the interaction angle. However, the effect is different depending on the city. For example, in Pittsburgh, the preference for low lateral distances at low interaction angles is stronger when compared to Singapore.

The reward functions can demonstrate the competitive behavior of the conflict interactions, which can be properly represented by the MA-AIRL model. For instance, the VRF, plotted with the lateral distance and the speed difference variables, shows that drivers prefer positive speed differences (i.e., to be faster than pedestrians). Conversely, the PRF, plotted with the same set of variables, indicates that pedestrians have preferences toward negative speed differences (i.e., to be faster than drivers). Therefore, in all cities, this behavior indicates that road users have conflicting preferences while interacting with each other. This is different from single-agent models that solely model a single road user (e.g., the pedestrian) (Lanzaro et al., 2022b; Nasernejad et al., 2021), which cannot deal with the equilibrium between both agents' intentions.

It should be noted that the reward plots show the variation of two states while keeping the other states at their mean variables, and additional behavioral influences can be made if the other state variables are changed. Furthermore, the plots may present various scales for different cities. Therefore, the magnitude of the reward in one city is not compared to the magnitude of the reward in another city, i.e., just descriptive behavioral trends are inferred.

5.2. Model results and trajectory simulation

The MA-AIRL framework was applied to four different cities, which presented datasets with very different characteristics (e.g., data size, state and action variables' distributions). Therefore, the models are expected to present different learning patterns to properly reflect road user behavior. Table 3 shows the explained variances and the losses for the models developed in each city. These metrics were calculated considering the final iterations before model convergence. The explained variance indicates the percentage of the variance from the actual dataset that can be accurately described by the predictor state variables, and values closer to 1 are preferred. Also, the losses show how well the model was able to predict actual behavior, and lower values are desired. Results demonstrate that the Boston and the Las Vegas models presented the best results in terms of explained variance and losses, which shows that the MA-AIRL was better at representing road user behavior in these cities. Conversely, the model results in both Pittsburgh and Singapore showed lower performance. For example, the explained variances for driver and pedestrian behavior were 45.2 % and 4.7 % greater in the Boston model, respectively, when compared to the Singapore model. Similar results were obtained for the loss metric results, where the total loss was 26.3 % lower in the Boston model when compared to the Singapore model.

Table 4
Normalized Root Mean Square Error (NRMSE) for 4 different cities.

City	Longitudinal distance	Lateral distance	Pedestrian speed	Vehicle speed
Boston	0.086	0.344	0.576	0.438
Las Vegas	0.147	0.378	0.584	0.417
Pittsburgh	0.139	0.329	0.728	0.502
Singapore	0.091	0.346	0.841	0.510

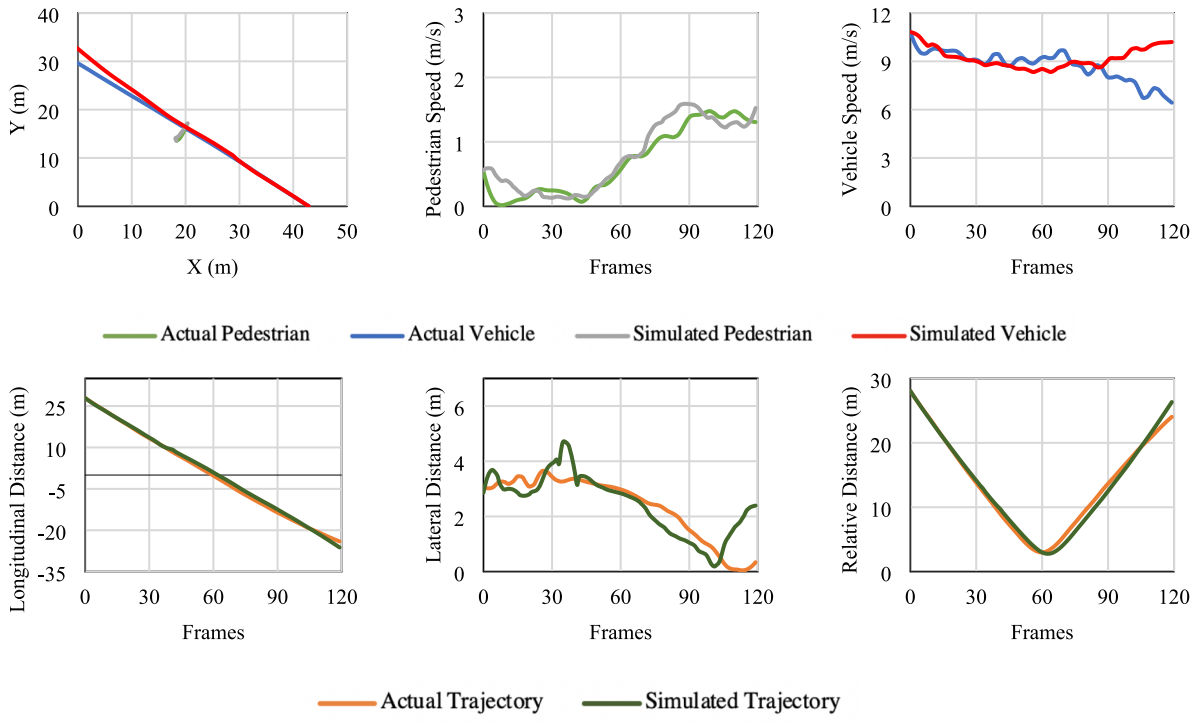
Table 5
Accuracy for predicting the evasive action mechanisms in 4 different cities.

City	Driver		Pedestrian	
	Change in Speed	Change in orientation	Change in Speed	Change in orientation
Boston	84.1 %	71.5 %	76.6 %	70.6 %
Las Vegas	76.8 %	70.2 %	71.5 %	65.4 %
Pittsburgh	76.9 %	61.5 %	76.9 %	61.5 %
Singapore	82.2 %	63.3 %	63.3 %	65.9 %

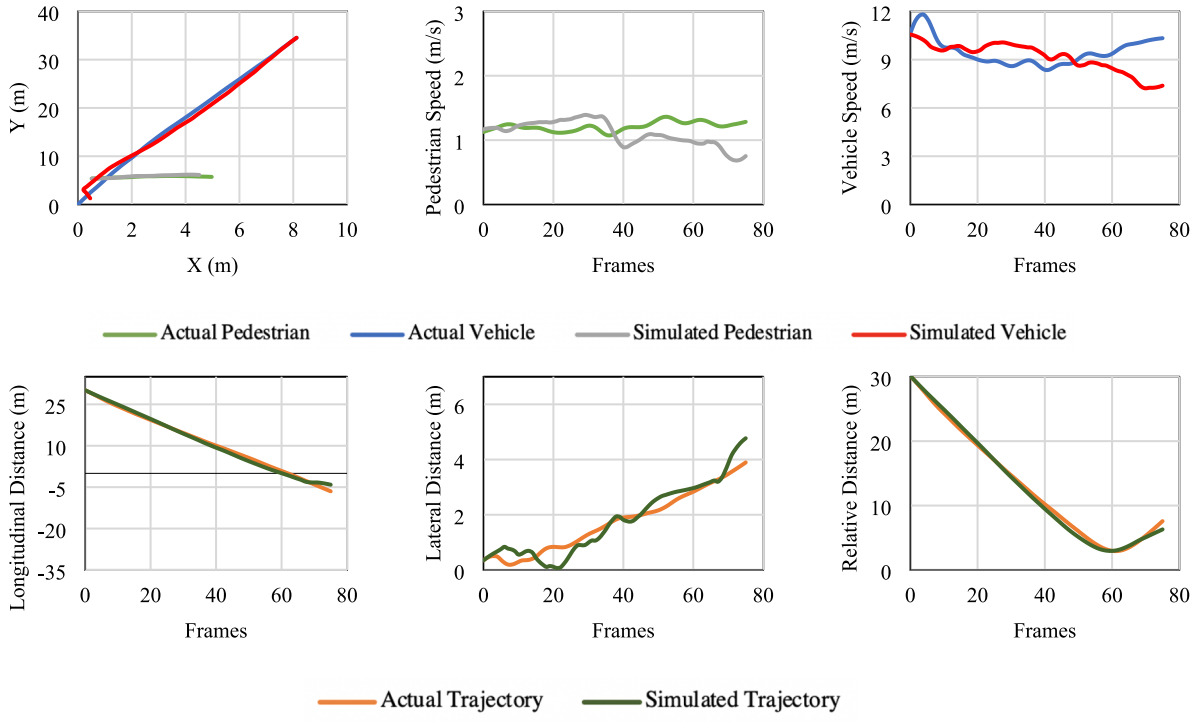
Simulated road user trajectories were obtained with the optimal policies, which were estimated with the MA-AIRL algorithm. The Normalized Root Mean Square Error (NRMSE) was employed to compute the differences between the actual road user trajectories and the simulated trajectories. Table 4 illustrates the NRMSE values for longitudinal distance, lateral distance, and speed of both road users. Results show that, in terms of speed prediction, the Boston and Las Vegas models presented the lowest errors, whereas the highest errors were found in the Singapore model. However, the distance errors for the Singapore model were comparable to the Boston model (i.e., 5.8 % greater for longitudinal distance and 0.6 % for lateral distance). Finally, the Las Vegas model presented the highest errors for the distance variables despite having obtained relatively low errors for the speed prediction when compared to the other cities.

Table 5 shows the model accuracy for predicting evasive actions in Boston, Las Vegas, Pittsburgh, and Singapore. The simulated trajectories were then compared to the actual road user trajectories to identify the evasive action mechanisms taken by both road users. For changing the speed, road users could accelerate, decelerate, or take no action. For changing the orientation, road users could either make yaw change maneuvers or keep their original intentions of motion. The table shows that predicting drivers' evasive actions is generally easier than predicting pedestrians' evasive actions. Also, the speed prediction performance tends to be better than the orientation prediction performance, except for pedestrian behavior in Singapore. In this city, the accuracy to predict pedestrians' change in orientation was greater than to predict pedestrians' change in speed. The Boston model presented the best accuracy results, and the Singapore model presented the lowest performance overall. Therefore, despite predicting the positions fairly well, the Singapore model was not able to capture the evasive action mechanisms as accurately as the models developed in the other cities.

Fig. 5 shows examples of simulated and actual conflict interactions in each city. The plots indicate that the different models predicted the road user trajectories with good accuracy and replicated the evasive action mechanisms. For example, in Boston, the example depicts a conflict interaction where the vehicle approached the pedestrian with minor speed reductions. The pedestrian was at a very low speed and only increased its speed more significantly after a few seconds, i.e., when the vehicle was distancing itself from the pedestrian. In the Las Vegas



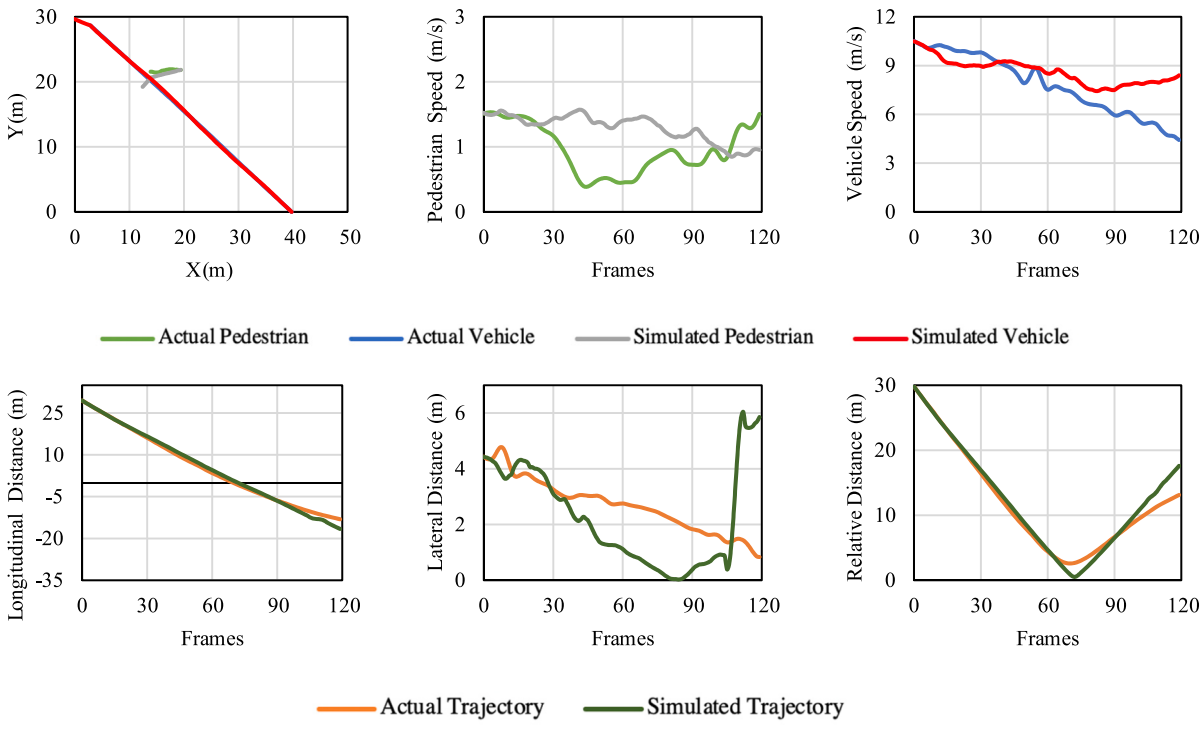
(a) Boston



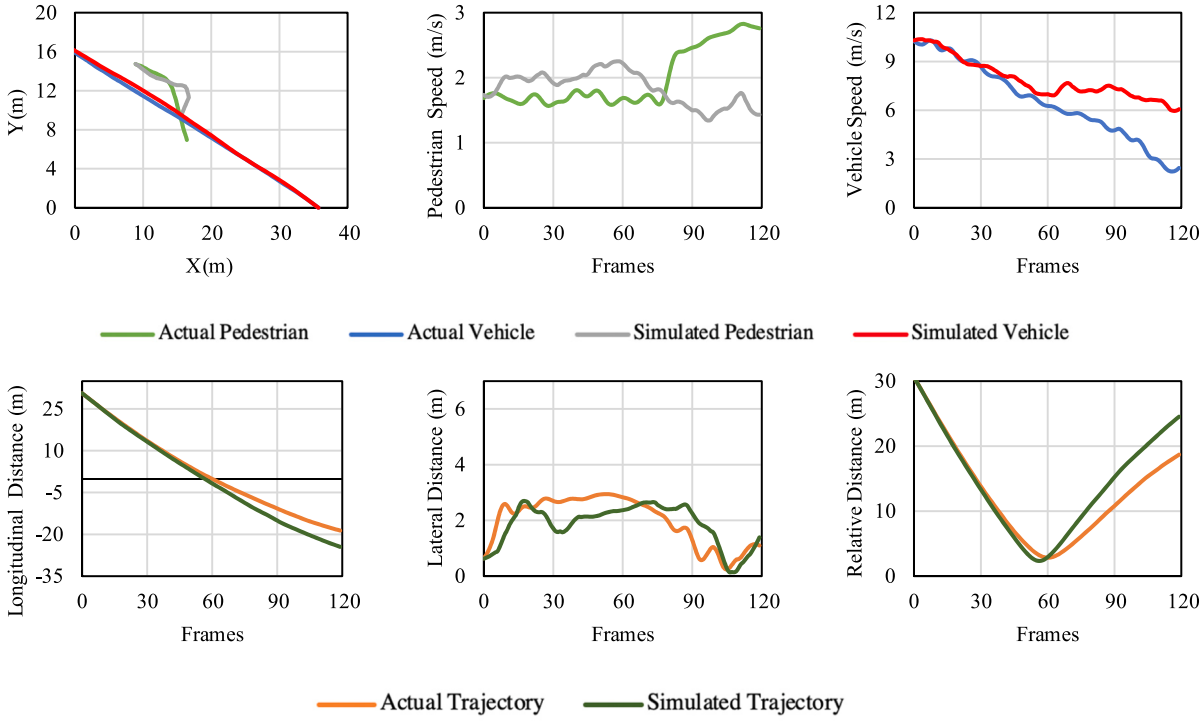
(b) Las Vegas

Fig. 5. Examples of Vehicle-Pedestrian Interactions in 4 different cities.

example, both road users reduced their speeds immediately before the conflict. A similar situation is encountered in the Pittsburgh model. However, the speed prediction accuracy in the Pittsburgh model is lower when compared to the Las Vegas model. Finally, in the Singapore model, the driver took an evasive action to decrease its speed while it was approaching the pedestrian. However, in the simulated model, the pedestrian took an additional evasive action maneuver (i.e., change in orientation) to prevent a vehicle–pedestrian crash. Therefore, the



(c) Pittsburgh



(d) Singapore

Fig. 5. (*continued*).

models were able to properly represent the evasive action mechanisms; nevertheless, the accuracy of the prediction was different depending on the environment.

5.3. Transferability analysis

Road user behavior is highly dependent on the environment, which plays a significant role in road safety. For example, the Boston dataset represents an active transportation friendly environment as it contains

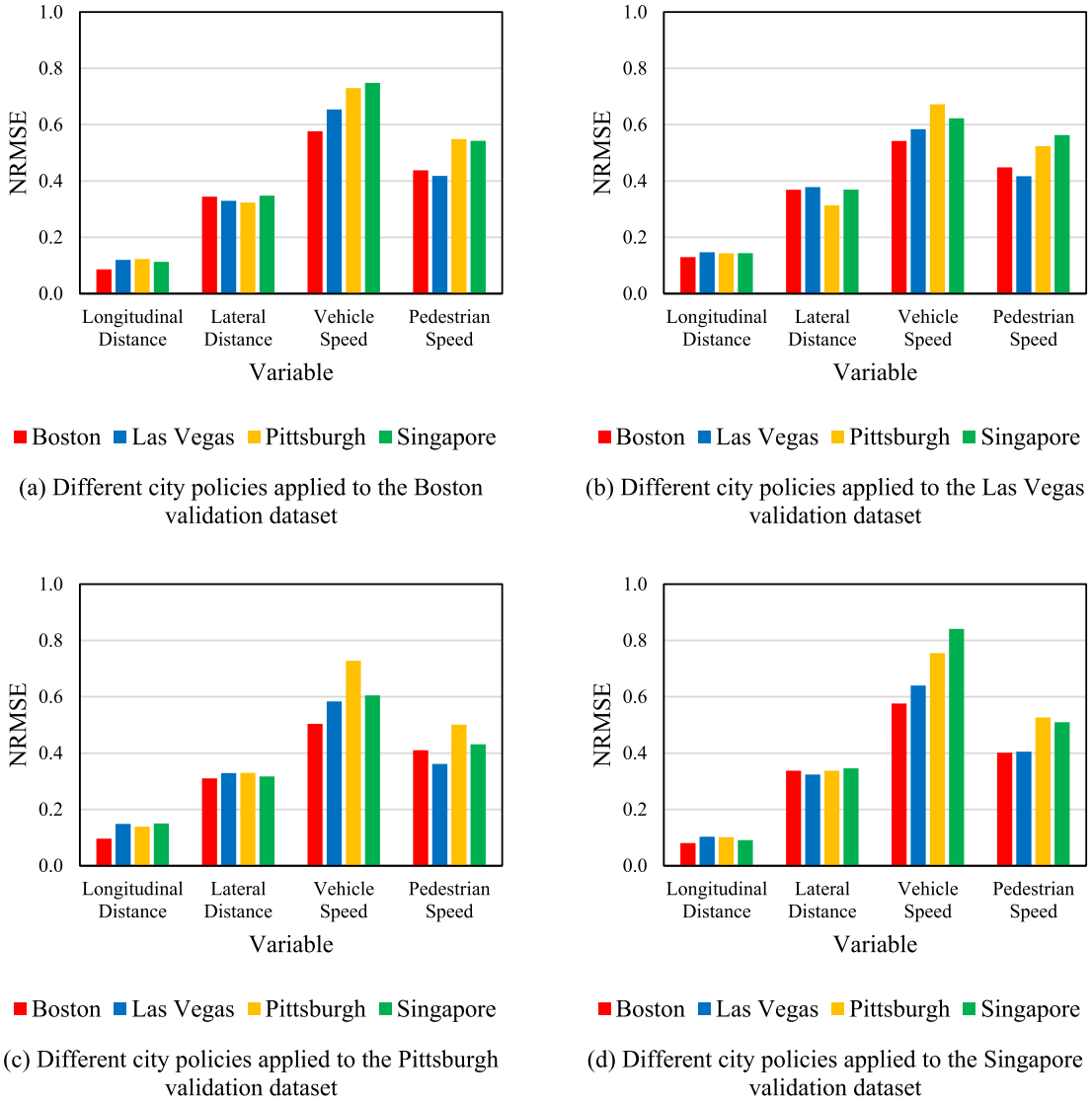


Fig. 6. NRMSE values obtained from applying different multi-agent policies in the Boston, Las Vegas, Pittsburgh, and Singapore validation datasets.

large sidewalks and multiple cycle paths, and drivers tend to travel at lower speeds. Alternately, the Las Vegas dataset contains a larger number of tourists, and the urban environment consists of wide roads with several lanes in each direction. Therefore, pedestrians and drivers are expected to behave differently because they are exposed to diverse conditions. For each city, a multi-agent policy that describes the optimal sequence of decisions of both drivers and pedestrians was obtained. In order to evaluate how agents of one location would behave in other environments, the multi-agent policies of each city were switched across different environments. Then, the multi-agent policy of one location was applied in (1) the validation dataset corresponding to the city where the model was trained and (2) the validation dataset of the other cities.

Fig. 6 shows the NRMSE values for longitudinal distance, lateral distance, and speeds of both road users after applying different multi-agent policies in the various validation datasets. This NRMSE results were calculated by comparing the simulated road user trajectories, considering different multi-agent polices, to the actual trajectories. Results indicate that, in terms of position prediction, the multi-agent policies can be interchanged without leading to major changes in NRMSE values. For example, considering the Las Vegas validation dataset, applying the Pittsburgh multi-agent policy resulted in a 2.1 % difference for longitudinal distance when compared to the Las Vegas multi-agent policy. In addition, in the same validation dataset, applying the Boston multi-agent policy led to a 2.4 % difference for lateral distance when compared to the Las Vegas multi-agent policy. However, the results for speed prediction are considerably different. For instance, considering the Boston validation dataset, applying the Singapore multi-agent policy resulted in a 29.9 % error increase in vehicle speed and a 24.0 % error increase in pedestrian speed when compared to simply applying the Boston multi-agent policy.

The differences in speed prediction are lower when comparing Boston to Las Vegas. For example, considering the Las Vegas validation dataset, applying the Boston multi-agent policy resulted in a difference of 7.4 % in pedestrian speed when compared to the Las Vegas multi-agent policy, whereas applying the Pittsburgh and the Singapore multi-agent policies led to differences of 25.7 % and 35.0 %, respectively, in pedestrian speed. This shows that the differences in behavior in both Pittsburgh and Singapore are more meaningful when compared to Boston and Las Vegas. Similar conclusions can be observed when applying the multi-agent policies to the Boston validation dataset. For instance, the vehicle speed presented differences of 13.5 %, 26.7 %, and 29.9 % when the multi-agent policies of Las Vegas, Pittsburgh, and

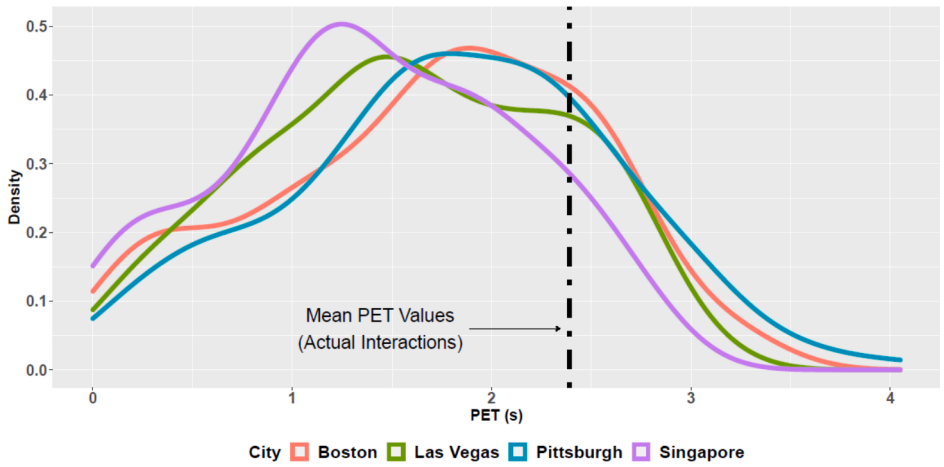


Fig. 7. PET distributions for the Simulated Trajectories in the Boston validation dataset using different multi-agent policies.

Singapore, respectively, were applied to the Boston validation dataset when compared to solely applying the Boston multi-agent policy.

Fig. 7 demonstrates the PET distributions of the simulated trajectories after applying the multi-agent policies. The Boston validation dataset is used as an example, and lower PET values are associated with more severe interactions. The figure shows that, when applying the Singapore multi-agent policies to the Boston validation dataset, the interactions tend to have greater severity. This is different from applying multi-agent policies from both Las Vegas and Pittsburgh. In these cases, the PET distributions approach the PET distribution when the Boston multi-agent policy is applied. These results indicate that considering agents of different environments in another dataset might lead to increased risk levels, especially for circumstances of very different driving situations (e.g., comparing Singapore to Boston). Therefore, caution should be exercised when transferring multi-agent policies of different environments as road user behavior exercises a significant impact on the severity of the interactions. Similar conclusions were found by applying the multi-agent policies in other validation datasets.

6. Conclusions

This paper used MA-AIRL to compare simulated vehicle–pedestrian interactions in four different environments, namely Boston, Las Vegas, Pittsburgh, and Singapore. This framework models drivers and pedestrians in a Markov Game, where both road users are able to take logical decisions whilst considering the equilibrium in the road users' intentions. The model estimated reward functions, which generated insights related to road user behavioral aspects in different environments. Furthermore, a simulation platform was developed with the optimal policies obtained from the MACK DRL algorithm. This resulted in simulated driver and pedestrian trajectories, which were then compared to the actual road user trajectories to evaluate (1) the losses and the explained variances, (2) the differences in speed and position, (3) the prediction of the evasive action mechanisms, (4) the transferability of agents across different environments.

Results have shown that the MA-AIRL obtained better results in some datasets when compared to others. For example, the explained variance and the losses in the Boston and Las Vegas models were considerably better than the other models. Moreover, these models were able to predict the evasive action mechanisms with good accuracy (e.g., 84.1 % and 71.5 % for predicting speed and orientation changes, respectively, for drivers in the Boston model, and 76.6 % and 70.6 % for predicting speed and orientation changes, respectively, for pedestrians in the same model). However, this good prediction accuracy was lower for the other models, especially Pittsburgh and Singapore. Therefore, the MA-AIRL was not able to accurately learn the complex nuances of different driving behavior when compared to the Boston and Las Vegas models. Finally, the transferability analysis showed that switching agents across different environments might lead to increased severity in the interactions.

This study demonstrates that directly applying multi-agent policies without accounting for local behavioral characteristics can result in different risk levels. Therefore, AV policies for collision avoidance mechanisms, which are expected to properly consider the evasive actions of both road users in conflict situations, should be developed with site-specific parameters, such as road users' local preferences (e.g., relative speeds, sudden changes in orientations, distances that can be tolerated), given that road users react to conflicts differently. This work has shown that the MA-AIRL algorithm is able to capture these complex interactions between various road uses as it considers the equilibrium between their intentions. This framework yields reward functions and optimal policies, which can be further incorporated into AV motion planning benchmarks. However, the optimal strategies to avoid crashes might be different depending on the environment.

This study has some limitations, which can be accounted for in future works. For example, the Markov Game framework applied in this study consists of six state variables and two action variables, and additional variables to represent road user behavior (e.g., demographics, violation behavior) can be incorporated if known. Also, this study considered conflicts identified with the PET conflict indicator, and other conflict indicators can be investigated in the future. Furthermore, additional frameworks that model different agents in various environments (e.g., transfer learning, imitation learning) can be implemented. This study used datasets from four locations in the US and Asia, and other locations, such as environments with more intense mixed traffic and less organization, should be considered. Finally, the multi-agent policies developed in this study can be compared with standard collision avoidance systems (Milanés et al., 2012; Shalev-Shwartz et al., 2018).

CRediT authorship contribution statement

Gabriel Lanzaro: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft. **Tarek Sayed:** Conceptualization, Investigation, Supervision, 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.

Data availability

Data will be made available on request.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *Proceedings, Twenty-First International Conference on Machine Learning, ICML, 2004*, 1–8.

Alsaleh, R., & Sayed, T. (2022). Do road users play Nash Equilibrium? A comparison between Nash and Logistic stochastic Equilibriums for multiagent modeling of road user interactions in shared spaces. *Expert Systems with Applications, 205*, Article 117710. https://doi.org/10.1016/j.eswa.2022.117710

Alsaleh, R., & Sayed, T. (2021). Markov-game modeling of cyclist-pedestrian interactions in shared spaces: A multi-agent adversarial inverse reinforcement learning approach. *Transportation Research Part C, 128*, Article 103191. https://doi.org/10.1016/j.trc.2021.103191

Alsaleh, R., & Sayed, T. (2020). Modeling pedestrian-cyclist interactions in shared space using inverse reinforcement learning. *Transportation Research Part F: Traffic Psychology and Behaviour, 70*, 37–57.

Barbosa, H., Cunto, F., Bezerra, B., Nodari, C., & Jacques, M. A. (2014). Safety performance models for urban intersections in Brazil. *Accident Analysis & Prevention, 70*, 258–266. https://doi.org/10.1016/j.aap.2014.04.008

Caesar, H., Kabzan, J., Tan, K.S., Fong, W.K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., Omari, S., 2021. NuPlan: A closed-loop ML-based planning benchmark for autonomous vehicles.

Chai, H., Zhang, Z., Hu, H., Dai, L., & Bian, Z. (2023). Trajectory-based conflict investigations involving two-wheelers and cars at non-signalized intersections with computer vision. *Expert Systems with Applications, 230*, Article 120590. https://doi.org/10.1016/j.eswa.2023.120590

Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., Hays, J., 2019. Argoverse: 3D tracking and forecasting with rich maps. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2019-June, 8740–8749. doi:10.1109/CVPR.2019.00895.

Chao, Q., Deng, Z., & Jin, X. (2015). Vehicle-pedestrian interaction for mixed traffic simulation: Vehicle-pedestrian interaction for mixed traffic simulation. *Comp. Anim. Virtual Worlds, 26*(3–4), 405–412. https://doi.org/10.1002/cav.1654

Chen, P., Zeng, W., & Yu, G. (2019). Assessing right-turning vehicle-pedestrian conflicts at intersections using an integrated microscopic simulation model. *Accident Analysis & Prevention, 129*, 211–224. https://doi.org/10.1016/j.aap.2019.05.018

Duarte, F., & Ratti, C. (2018). The Impact of autonomous vehicles on cities: A review. *Journal of Urban Technology, 25*(4), 3–18. https://doi.org/10.1080/10630732.2018.1493883

Essa, M., & Sayed, T. (2018). Traffic conflict models to evaluate the safety of signalized intersections at the cycle level. *Transportation Research Part C: Emerging Technologies, 89*, 289–302.

Farah, H., & Azevedo, C. L. (2017). Safety analysis of passing maneuvers using extreme value theory. *IATSS Research, 41*(1), 12–21. https://doi.org/10.1016/j.iatssr.2016.07.001

Feng, M., Wang, X., Lee, J., Abdel-Aty, M., & Mao, S. (2020). Transferability of safety performance functions and hotspot identification for freeways of the United States and China. *Accident Analysis & Prevention, 139*, Article 105493. https://doi.org/10.1016/j.aap.2020.105493

Finn, C., Levine, S., Abbeel, P., 2016. Guided cost learning: Deep inverse optimal control via policy optimization. 33rd International Conference on Machine Learning, ICML 2016 1, 95–107.

Formosa, N., Quddus, M., Ison, S., Abdel-Aty, M., & Yuan, J. (2020). Predicting real-time traffic conflicts using deep learning. *Accident Analysis & Prevention, 136*, Article 105429. https://doi.org/10.1016/j.aap.2019.105429

Fu, C., & Sayed, T. (2022). A multivariate method for evaluating safety from conflict extremes in real time. *Analytic Methods in Accident Research, 36*, Article 100244. https://doi.org/10.1016/j.amar.2022.100244

Fu, J., Luo, K., Levine, S., 2017. Learning robust rewards with adversarial inverse reinforcement learning. arXiv preprint arXiv:1710.11248 1–15.

Georgila, K., Nelson, C., Traum, D., 2014. Single-agent vs. multi-agent techniques for concurrent reinforcement learning of negotiation dialogue policies, in: 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. pp. 500–510. doi:10.3115/v1/p14-1047.

Golchoubian, M., Ghafurian, M., Dautenhahn, K., & Azad, N. L. (2023). Pedestrian trajectory prediction in pedestrian-vehicle mixed environments: A review. *IEEE Trans. Intell. Transport. Syst., 24*(11), 11544–11567. https://doi.org/10.1109/TITS.2023.3291196

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *Communications of the ACM, 63*(11), 139–144. https://doi.org/10.1145/3422622

Guo, Y., Essa, M., Sayed, T., Haque, M. M., & Washington, S. (2019). A comparison between simulated and field-measured conflicts for safety assessment of signalized intersections in Australia. *Transportation Research Part C: Emerging Technologies, 101*, 96–110.

Jiang, X., Wang, W., Bengler, K., & Guo, W. (2015). Analyses of pedestrian behavior on mid-block unsignalized crosswalk comparing Chinese and German cases, 168781401561046 *Advances in Mechanical Engineering, 7*(11). https://doi.org/10.1177/1687814015610468.

Kamel, A., Sayed, T., & Fu, C. (2022). Real-time safety analysis using autonomous vehicle data: A Bayesian hierarchical extreme value model. *Transportmetrica B*. https://doi.org/10.1080/21680566.2022.2135634

Kassim, A., Ismail, K., & Hassan, Y. (2014). Automated measuring of cyclist - motor vehicle post encroachment time at signalized intersections. *Canadian Journal of Civil Engineering, 41*(7), 605–614. https://doi.org/10.1139/cjce-2013-0565

La Torre, F., Domenichini, L., Branzi, V., Meocci, M., Paliotto, A., & Tanzi, N. (2022). Transferability of the highway safety manual freeway model to EU countries. *Accident Analysis & Prevention, 178*, Article 106852. https://doi.org/10.1016/j.aap.2022.106852

Lanzaro, G., Sayed, T., & Alsaleh, R. (2022a). Modeling motorcyclist-pedestrian near misses: A multiagent adversarial inverse reinforcement learning approach. *J. Comput. Civ. Eng., 36*(6), 04022038. https://doi.org/10.1061/(ASCE)CP.1943-5487.0001053

Lanzaro, G., Sayed, T., & Alsaleh, R. (2022b). Can motorcyclist behavior in traffic conflicts be modeled? A deep reinforcement learning approach for motorcycle-pedestrian interactions. *Transportmetrica B, 10*(1), 396–420. https://doi.org/10.1080/21680566.2021.2004954

Levine, S., Koltun, V. 2012. Continuous inverse optimal control with locally optimal examples, in: Proceedings of the 29th International Conference on Machine Learning.

Littman, M.L. 1994. Markov games as a framework for multi-agent reinforcement learning, Machine Learning Proceedings 1994. Morgan Kaufmann Publishers, Inc. doi:10.1016/b978-1-55860-335-6.50027-1.

Lu, L., Ren, G., Wang, W., Chan, C.-Y., & Wang, J. (2016). A cellular automaton simulation model for pedestrian and vehicle interaction behaviors at unsignalized mid-block crosswalks. *Accident Analysis & Prevention, 95*, 425–437. https://doi.org/10.1016/j.aap.2016.04.014

McIlroy, R. C., Nam, V. H., Bunyasi, B. W., Jikyong, U., Kokwaro, G. O., Wu, J., Hoque, M. S., Plant, K. L., Preston, J. M., & Stanton, N. A. (2020). Exploring the relationships between pedestrian behaviours and traffic safety attitudes in six countries. *Transportation Research Part F: Traffic Psychology and Behaviour, 68*, 257–271. https://doi.org/10.1016/j.trf.2019.11.006

Milanés, V., Pérez, J., Godoy, J., & Onieva, E. (2012). A fuzzy aid rear-end collision warning/avoidance system. *Expert Systems with Applications, 39*(10), 9097–9107. https://doi.org/10.1016/j.eswa.2012.02.054

Mnih, V., Badia, A.P., Mirza, L., Graves, A., Harley, T., Lillicrap, T.P., Silver, D., Kavukcuoglu, K., 2016. Asynchronous methods for deep reinforcement learning, in: 33rd International Conference on Machine Learning, ICML 2016. pp. 2850–2869.

Nasernejad, P., Sayed, T., & Alsaleh, R. (2022). Multiagent modeling of pedestrian-vehicle conflicts using adversarial inverse reinforcement learning. *Transportmetrica A: Transport Science*. https://doi.org/10.1080/23249935.2022.2061081

Nasernejad, P., Sayed, T., & Alsaleh, R. (2021). Modeling pedestrian behavior in pedestrian-vehicle near misses: A continuous Gaussian process inverse reinforcement learning (GP-IRL) approach. *Accident Analysis and Prevention, 161*(May), Article 106355. https://doi.org/10.1016/j.aap.2021.106355

Ng, A.Y., Russel, S., 2000. Algorithms for Inverse Reinforcement Learning, in: International Conference on Machine Learning.

Nordfjærn, T., Şimşekoğlu, Ö., & Rundmo, T. (2014). Culture related to road traffic safety: A comparison of eight countries using two conceptualizations of culture. *Accident Analysis and Prevention, 62*, 319–328. https://doi.org/10.1016/j.aap.2013.10.018

Nordfjærn, T., & Zavareh, M. F. (2016). Individualism, collectivism and pedestrian safety: A comparative study of young adults from Iran and Pakistan. *Safety Science, 87*, 8–17. https://doi.org/10.1016/j.ssci.2016.03.005

Pakgohar, A., Tabrizi, R. S., Khalili, M., & Esmaeili, A. (2011). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: A data mining approach. *Procedia Computer Science, 3*, 764–769. https://doi.org/10.1016/j.procs.2010.12.126

Parada, R., Aguilar, A., Alonso-Zarate, J., & Vazquez-Gallego, F. 2021. Machine Learning-based Trajectory Prediction for VRU Collision Avoidance in V2X Environments, in: 2021 IEEE Global Communications Conference (GLOBECOM). Presented at the GLOBECOM 2021 - 2021 IEEE Global Communications Conference, IEEE, Madrid, Spain, pp. 1–6. doi:10.1109/GLOBECOM46510.2021.9685520.

Penmetsa, P., Sheinidashtegol, P., Musaev, A., Adanu, E. K., & Hudnall, M. (2021). Effects of the autonomous vehicle crashes on public perception of the technology. *IATSS Research, 45*(4), 485–492. https://doi.org/10.1016/j.iatssr.2021.04.003

Rossato, L., Silva, Luis A. L., Assunção, Joaquim, 2020. A Markovian model for the Game of Truco, in: SBC – Proceedings of SBGames 2020.

Saunier, N., & Sayed, T., 2006. A feature-based tracking algorithm for vehicles in intersections, in: Third Canadian Conference on Computer and Robot Vision, CRV 2006.

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry, 36*(8), 1627–1639.

Sayed, T., Abdelwahab, W., & Navin, F. (1995). Identifying accident-prone locations using fuzzy pattern recognition. *Journal of Transportation Engineering, 121*(4), 352–358. https://doi.org/10.1061/(ASCE)0733-947X(1995)121:4(352)

Sayed, T., & Zein, S. (1999). Traffic conflict standards for intersections. *Transportation Planning and Technology, 22*(4), 309–323.

Shalev-Shwartz, S., Shammah, S., & Shashua, A., 2018. On a Formal Model of Safe and Scalable Self-driving Cars.

Shou, Z., Chen, X., Fu, Y., & Di, X. (2022). Multi-agent reinforcement learning for Markov routing games: A new modeling paradigm for dynamic traffic assignment. *Transportation Research Part C: Emerging Technologies, 137*, Article 103560. https://doi.org/10.1016/j.trc.2022.103560

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature, 550*(7676), 354–359. https://doi.org/10.1038/nature24270

Song, J., Ren, H., Ermon, S., & Sadigh, D. (2018). Multi-agent generative adversarial imitation learning. *Advances in Neural Information Processing Systems*, 7461–7472.

Sutton, R.S., & Barto, A.G., 2018. Reinforcement Learning: An Introduction, IEEE Transactions on Neural Networks. MIT Press.

Tageldin, A., & Sayed, T. (2019). Models to evaluate the severity of pedestrian-vehicle conflicts in five cities. *Transportmetrica A: Transport Science, 15*(2), 354–375.

Tageldin, A., Sayed, T., & Shaaban, K. (2017). Comparison of time-proximity and evasive action conflict measures case studies from five cities. *Transportation Research Record, 2661*, 19–29. https://doi.org/10.3141/2661-03

Talebpour, A., Mahmassani, H. S., & Elfar, A. (2017). Investigating the effects of reserved lanes for autonomous vehicles on congestion and travel time reliability. *Transportation Research Record, 2622*(1), 1–12. https://doi.org/10.3141/2622-01

Tinella, L., Koppel, S., Lopez, A., Caffò, A. O., & Bosco, A. (2022). Associations between personality and driving behavior are mediated by mind-wandering tendency: A cross-national comparison of Australian and Italian drivers. *Transportation Research Part F: Traffic Psychology and Behaviour, 89*, 265–275. https://doi.org/10.1016/j.trf.2022.06.019

Tolksdorf, L., Tejada, A., Van De Wouw, N., & Birkner, C. (2023). Risk in Stochastic and Robust Model Predictive Path-Following Control for Vehicular Motion Planning. In *in: 2023 IEEE Intelligent Vehicles Symposium (IV). Presented at the 2023 IEEE Intelligent Vehicles Symposium (IV)* (pp. 1–8). Anchorage, AK, USA: IEEE. https://doi.org/10.1109/IV55152.2023.10186708.

Waizman, G., Shoval, S., & Benenson, I. (2015). Micro-simulation model for assessing the risk of vehicle-pedestrian road accidents. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, 19*(1), 63–77. https://doi.org/10.1080/15472450.2013.856721

Wang, C., Xu, C., Xia, J., Qian, Z., & Lu, L. (2018). A combined use of microscopic traffic simulation and extreme value methods for traffic safety evaluation. *Transportation Research Part C: Emerging Technologies, 90*, 281–291. https://doi.org/10.1016/j.trc.2018.03.011

Wu, G., Tan, G., Deng, J., & Jiang, D. (2021). Distributed reinforcement learning algorithm of operator service slice competition prediction based on zero-sum markov game. *Neurocomputing, 439*, 212–222. https://doi.org/10.1016/j.neucom.2021.01.061

Wu, Y., Mansimov, E., Liao, S., Grosse, R., & Ba, J. (2017). Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation. *Advances in Neural Information Processing Systems*, 5280–5289.

You, C., Lu, J., Filev, D., & Tsiotras, P. (2019). Advanced planning for autonomous vehicles using reinforcement learning and deep inverse reinforcement learning. *Robotics and Autonomous Systems, 114*, 1–18. https://doi.org/10.1016/j.robot.2019.01.003

Yu, L., Song, J., & Ermon, S., 2019. Multi-agent adversarial inverse reinforcement learning, in: International Conference on Machine Learning (Pp. 7194-7201).

Zeng, W., Chen, P., Yu, G., & Wang, Y. (2017). Specification and calibration of a microscopic model for pedestrian dynamic simulation at signalized intersections: A hybrid approach. *Transportation Research Part C: Emerging Technologies, 80*, 37–70. https://doi.org/10.1016/j.trc.2017.04.009

Zhang, K., Yang, Z., Başar, T., 2019. Multi-Agent reinforcement learning: A selective overview of theories and algorithms 1–73.

Zhang, Z., & Fu, D. (2022). Modeling pedestrian–vehicle mixed-flow in a complex evacuation scenario. *Physica A: Statistical Mechanics and its Applications, 599*, Article 127468. https://doi.org/10.1016/j.physa.2022.127468

Zheng, L., & Sayed, T. (2020). A novel approach for real time crash prediction at signalized intersections. *Transportation Research Part C: Emerging Technologies, 117*, Article 102683. https://doi.org/10.1016/j.trc.2020.102683

Zheng, L., Sayed, T., & Mannering, F. (2021). Modeling traffic conflicts for use in road safety analysis: A review of analytic methods and future directions. *Analytic Methods in Accident Research, 29*, Article 100142. https://doi.org/10.1016/j.amar.2020.100142

Zhu, J., & Tasic, I. (2021). Safety analysis of freeway on-ramp merging with the presence of autonomous vehicles. *Accident Analysis and Prevention, 152*(May 2020), Article 105966. https://doi.org/10.1016/j.aap.2020.105966

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K., 2008. Maximum entropy inverse reinforcement learning., in: 23rd AAAI Conference on Artificial Intelligence. Chicago, IL, USA, pp. 1433–1438.