VeMo: Enabling Transparent Vehicular Mobility Modeling at Individual Levels with Full Penetration[*]

Yu Yang
Rutgers University
yy388@cs.rutgers.edu

Xiaoyang Xie
Rutgers University
xx88@cs.rutgers.edu

Zhihan Fang
Rutgers University
zhihan.fang@cs.rutgers.edu

Fan Zhang
Shenzhen Institute of Advanced Technology
zhangfan@siat.ac.cn

Yang Wang
University of Science and Technology of China
Angyan@ustc.edu.cn

Desheng Zhang
Rutgers University
desheng.zhang@cs.rutgers.edu

ABSTRACT

Understanding and predicting real-time vehicle mobility patterns on highways are essential to address traffic congestion and respond to the emergency. However, almost all existing works (e.g., based on cellphones, onboard devices, or traffic cameras) suffer from high costs, low penetration rates, or only aggregate results. To address these drawbacks, we utilize Electric Toll Collection systems (ETC) as a large-scale sensor network and design a system called VeMo to transparently model and predict vehicle mobility at the individual level with a full penetration rate. Our novelty is how we address uncertainty issues (i.e., unknown routes and speeds) due to sparse implicit ETC data based on a key data-driven insight, i.e., individual driving behaviors are strongly correlated with crowds of drivers under certain spatiotemporal contexts and can be predicted by combining both personal habits and context information. More importantly, we evaluate VeMo with (i) a large-scale ETC system with tracking devices at 773 highway entrances and exits capturing more than 2 million vehicles every day; (ii) a fleet consisting of 114 thousand vehicles with GPS data as ground truth. We compared VeMo with state-of-the-art benchmark mobility models, and the experimental results show that VeMo outperforms them by average 10% in terms of accuracy.

KEYWORDS

Vehicular Mobility Modeling, Static Sensors, Toll Systems, Destination Prediction, Route Prediction, Speed Prediction

ACM Reference Format:
Yu Yang, Xiaoyang Xie, Zhihan Fang, Fan Zhang, Yang Wang, and Desheng Zhang. 2019. VeMo: Enabling Transparent Vehicular Mobility Modeling at Individual Levels with Full Penetration. In *The 25th Annual International Conference on Mobile Computing and Networking (MobiCom '19), October 21–25, 2019, Los Cabos, Mexico.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3300061.3300130

1 INTRODUCTION

Understanding and modeling individual vehicular mobility on highways have various applications, e.g., congestion prediction [23], route planning [6] and ramp metering [37]. However, modeling and predicting individual vehicle locations in fine spatial-temporal granularity are extremely challenging due to a large number of vehicles and limited infrastructures on highways compared to cities [1][35].

The existing approaches for vehicle location prediction can be basically categorized into two groups: (i) mobile infrastructure based solutions such as cellphones (e.g., Online Map Services [18]) and onboard devices (e.g., OBD devices [7]), and (ii) static infrastructure based solutions: traffic cameras [45], loop sensors [39], and RFID [52]. For mobile infrastructure based solutions, they typically have privacy issues since they require real-time GPS locations of vehicles [55]; for static infrastructure based solutions, they typically introduce low spatial coverage or high costs for a complete highway system coverage [43]. Further, both of them may suffer low penetration rates, e.g., some commuters do not use navigation apps when traveling some familiar routes [38]; traffic cameras are not pervasive on highways in some countries [16].

[*] Prof. D. Zhang and Y. Wang are the joint corresponding authors.

MobiCom '19, October 21–25, 2019, Los Cabos, Mexico

ACM ISBN 978-1-4503-6169-9/19/10.
https://doi.org/10.1145/3300061.3300130

In this paper, to address these drawbacks, we utilize a highway Electric Toll Collection (ETC) system as a sensor network for vehicular mobility modeling and prediction. Compared to the existing approaches, our ETC based solution has the following features: (i) it requires no additional infrastructure since it relies on data already gathered in real time over highway networks for toll collections; (ii) it poses no additional privacy threats because it does not collect vehicle-specific GPS data; (iii) it does not suffer from low penetration rates since all vehicles have to be charged by an ETC system when using highway systems. Even some highways are installed with induction loops, they cannot achieve individual level modeling compared to the ETC system.

However, since an ETC system is deployed for toll collections instead of mobility modeling, we have the following new challenges. (i) An ETC system only logs when and where a vehicle enters and leaves a highway system for billing purposes and it leads to extremely sparse location records for each vehicle, i.e., only two data points per trip, which makes predicting destinations without intermediate locations be challenging. Without any historical routes or speeds logged, it is difficult to train a model. (ii) In a complicated highway network, given an entrance and exit, there are many potential routes as shown by our later analyses, and ETC data do not log any information regarding which route was taken during a particular origin and destination pair. Based on our data, we found that the shortest routes are not the first choices for many vehicles due to congestion. (iii) Traffic speeds vary by different spatiotemporal contexts, and ETC data do not directly log speeds. Straightforward solutions (e.g., assuming real-time speeds vary near the speed limit) usually do not perform well because of various driving behaviors under different contexts.

To address these challenges, in this paper, we perform a systemic investigation of a large-scale ETC system along with its data, and we found a key data-driven insight: even with complicated highway networks and real-time context, individual travel behaviors are strongly correlated with crowds under certain spatiotemporal contexts and can be predicted by combining both personal habits and context information. Built upon this insight, we design a model called VeMo to model and predict individual vehicular mobility patterns based on sparse observations on real-time origins as well as historical origins and destinations only. In particular, the key contributions of this paper are as follows.

- To our knowledge, we conduct the first systematic investigation of real-time vehicular mobility modeling and prediction based on large-scale ETC and GPS data. Our investigation is based on real-time and historical ETC data from 7.8 million vehicles and GPS data from 114 thousand vehicles. This large-scale vehicle sensing study enables us to find mobility insights that are not possible to obtain with small-scale systems and data. By working with our collaborators, we released some processed sample data for the benefit of the research community[1].
- We analyze both ETC and GPS data and provide some in-depth discussions on vehicular mobility patterns on highways. Based on the insights from our analyses, we design a mobility prediction system called VeMo with three key components to predict destinations, routes, and speeds for individual vehicles based on both historical and real-time ETC data. Technically, we extract unobserved routes and speeds through a joint optimization model. By studying various mobility features at both the individual level and crowd level, we fuse them based on a Mondrian Forests model to address the uncertainty issue in the mobility prediction.
- More importantly, we implement and evaluate the VeMo in Guangdong Province, China with (i) an ETC system covering 1,439 highway entrances and exits, and it captures around 2 million vehicles per day; (ii) a vehicle fleet and its GPS data including 114 thousand vehicles for evaluation only, where 20% of vehicles have the trajectories on highways.
- We evaluate VeMo through a two-month set of ETC and GPS data by showing both intermediate results (e.g., predicting destinations, routes, and speeds) and end-to-end results (e.g., predicting real-time locations). We study the performance sensitivity of our system to different spatial-temporal contexts. Compared with state-of-the-art solutions, VeMo provides a 10% performance gain on average in terms of prediction accuracy.

2 MOTIVATION

2.1 Use cases

VeMo aims to predict the real-time locations of individual vehicles, which enables various applications that cannot be achieved by previous solutions. As collaboration with the highway administrators, we gives two exemplary applications that matter a lot to the highway management.

- **Highway anomaly detection:** One important task for highway administrators is to detect the highway anomaly at the first time, such as traffic accidents. However, it is quiet expensive to arrange regular road check manually or cannot detect anomalies in time. Through predicting the real-time location of a vehicle, we can know when the vehicle is expected to leave the highway in the regular situation. Conversely, we could

[1] https://www.cs.rutgers.edu/ dz220/

know there may be an anomaly event if a number of vehicles do not leave the highway as expected.

- **Highway risk assessment:** Improving driving safety on highways is always an important topic for the highway administration companies. Noticeably, there are more than 6 million crashes on highways in the United States during 2015, including more than 30 thousand fatalities and 1 million injuries [24]. By transparently predicting the locations of individual vehicles, highway administration companies can understand the number of affected vehicles if there were an accident on certain road segments, and provide some contingency plans accordingly. Another safety related application is to localize a vehicle of interest (e.g., a vehicle with dangerous cargo or suspects) for public safety after it enters the highway.

Uniqueness of ETC based systems: To implement those applications, previous works either require extra installed infrastructures or suffer from low penetration rates of vehicles. For example, mobile phone based solutions can only know the locations of a number of vehicles, which cannot provides the accurate number of vehicles in a certain location. Induction loop based solutions cannot identify the uniqueness of a vehicle. Traffic cameras are potentially used to detect individual vehicles but limited by the laws in many countries such as U.S. Moreover, in developing countries, satellite images or mobile infrastructure is not well penetrated and it is really hard to predict the real-time locations. The ETC based toll system is universal and exist almost everywhere even in developing countries. Therefore, ETC based systems utilize widely deployed infrastructure (i.e., ETC), which can transparently obtain information from vehicles (i.e., when charging toll) with extremely low marginal cost. Moreover, the full penetration rate on highways can also make up for the weakness of mobile phone based solutions.

2.2 Challenges

It is not trivial to predict the real-time locations of vehicles because of the uncertainties caused by various traffic conditions and driving behaviors. To show these challenges, we study one-month data (both ETC transactions and trajectories of sample vehicles) in the Guangdong province of China and identify several challenges regarding three key factors including destinations, routes and speeds. The detailed data description is presented in Section 3 and Section 5.

(i) Destination uncertainty: To predict the real-time locations of vehicles, it is important to understand the destinations and routes. However, it is not trivial to predict the routes and destinations. To characterize the inherent predictability across vehicles, we present the destination entropy of each vehicle in Fig 1. The figure reveals two peaks as the entropy equals 0 and 1, which indicates the next location of a vehicle could be found on average in any $2^0 = 1$ and $2^1 = 2$ locations, respectively. Especially, we find most vehicles travel on highways only once in one month when the entropy=0; vehicles are more like to commute between two locations when the entropy=1. Many works [51] [11] have been done to predict the destinations of vehicles whose entropy is greater than or equal to one since those vehicles generally have regular commute patterns or extensive historical data. However, it is not clear how to predict the destinations of vehicles with only a few historical transactions. We refer this problem as a *destination sparsity* problem.

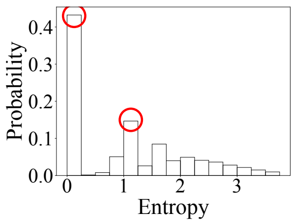

Fig 1: Destination entropy

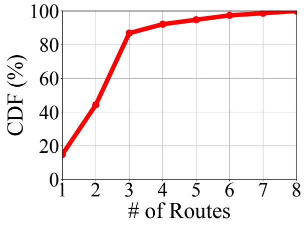

Fig 2: Number of routes

(ii) Unobserved routes and speeds: Previous studies have been done to model the route choices and driving speeds [6] [54]. Through studying the historical routes and speeds in the trip recorded by GPS-based devices, some sophisticated models are proposed to predict vehicular mobility in the near future. However, in our setting, one of the key characteristics of the ETC system is that it can only obtain very sparse information (i.e., the time and location when entering and exiting highways). This leads to the problem that we cannot obtain detailed routes and speeds to learn the route choice model and the driving speed model, which is not solved in the previous work.

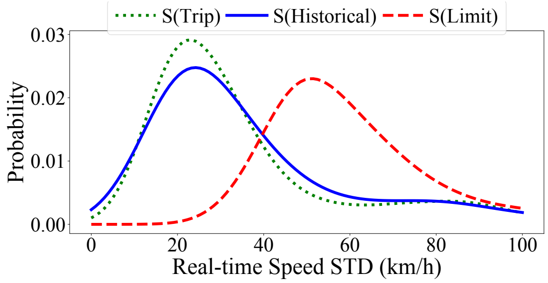

Fig 3: Speed STD

Moreover, routes and speeds also vary depending on user behaviors and contexts. For a given origin and destination, people can choose different routes if the road network is not trivial (i.e., only one route from the origin to the destination). Fig 2 illustrates the number of routes between the origin-destination pairs. We found that only 17% of station pairs have only one route based on GPS trajectories obtained from 114 thousand vehicles. It is impractical to assume only shortest routes are used by vehicles. (Note that these trajectories are only used in the motivation and evaluation rather than the model design.) As for speeds, people empirically expect that the driving speeds of vehicles are around certain speeds (e.g., speed limit or average speed) with less variance. However, in our study, we found the real-time speed is more complicated than the empirical intuition. To illustrate the characteristics of real-time speed, we study the real-time speed standard deviation (STD) across vehicles by replacing the mean value in the standard formula of standard deviation with the speed limit($S(Limit)$), the historical average speed($S(Historical)$), the current trip average speed($S(Trip)$), respectively. Fig 3 demonstrates both $S(Historical)$ and $S(Trip)$ have a Gaussian-like distribution with the mean STD near 20 km/h. It leads to a 330-meter offset in one-minute driving if only the average speed is utilized to obtain the real-time location. It also revels the fact that it is difficult for people to drive at the speed limit (e.g., can only drive at 60 km/h compared to the speed limit of 120 km/h) because of the heavy traffic.

2.3 Summary

The ETC based system provides an unprecedented opportunity to transparently model and predict vehicular mobility with a full penetration rate, which enables various potential applications such as highway safety management and adaptive dynamic toll strategies. However, due to the unique characteristic of only observing vehicles at entrances and exits, there are several challenges to be solved including destination sparsity problem and unobserved routes and speeds.

3 ETC SYSTEM AND DATA DESCRIPTION

We first introduce some notations to facilitate our discussion, and then give a brief description of an ETC system based on our infrastructure access in Guangdong and finally provide some data-driven insights.

Notations: Given ETC data on the vehicle's trip levels,

- An **edge** e is a highway segment between two adjacent toll stations, i.e., the finest spatial unit for ETC data-based modeling.
- A **route** r is a set of adjacent edges, which connect the origin toll station and the destination toll station of a particular trip.
- A **K-edge trip** is a trip of a particular vehicle with K edges in its route between the origin and the destination. Specifically, a **single-edge trip** has only one edge in the route.

Based on the above terms, our problem definition is "Given a vehicle entering a highway network from a toll station as an origin S_o at time T_o, predict its real-time locations on highways at any given time T_r until it exits the highway.

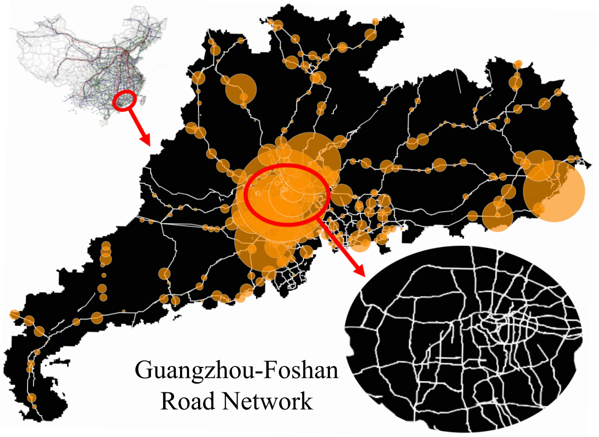

Fig 4: ETC Systems in Guangdong Province

Infrastructure Overview: Fig 4 shows the road structure and the locations of toll stations in the Guangdong province, which has 69 highways and 773 ETC toll stations with 1,439 highway entrances and exits covering an area of $179,800km^2$. The circles represent toll stations and the larger the icon, the heavier the daily traffic volume. It shows the traffic mainly concentrates on the central area and the road structure in that area is also complex as shown in the Guangzhou-Foshan Road Network. Each toll station detects all vehicles when they enter the highway system, and then logs the records as transactions after they leave the highway system. The toll station identifies a vehicle by ETC RFID devices (for regular charging) or cameras (for the purpose of detecting escaping charges).

As shown in Table 1, each generated transaction contains information including entering and exit station, entering and exiting time, vehicle id, vehicle type (i.e., car, bus, truck), axis count and weight. Such a transaction was generated when a vehicle enters and exits the highway network with both ETC cards or cash. On average, there are more than 4 million transactions generated every day from 2 million vehicles.

Statistic description: Fig 5 plots the average traffic volume in 24 hours of a day. It shows there are two peak hours (i.e., 10 am and 6 pm), which potentially make prediction

Table 1: ETC Transaction Description

Field	Value
Entering/Exit Toll Station	Humen Station
Entering/Exit Time	2016-07-01 13:00:01
Vehicle Id	F37SS1D4GU
Vehicle Type	Car/Bus/Truck
Axis Count	2
Weight	1500kg
Number of Daily Transactions: 4 millions	
Number of Daily Vehicles: 2 millions	

challenging due to uncertainty (e.g., route choice, traffic jam, etc) introduced by high traffic volume. Fig 6 depicts the daily transaction volume of all the toll stations, where 25 % of the stations contribute 75 % of the transactions. It suggests the major number of vehicles enter the highway from a limited number of stations, indicating prediction related to unpopular stations may suffer from lack of historical and real-time data.

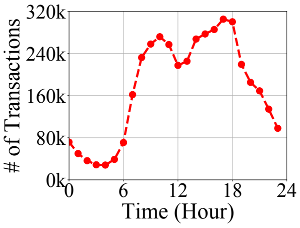

Fig 5: Volume over time

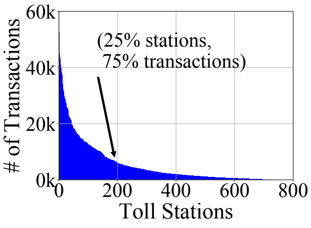

Fig 6: Volume over stations

4 VEMO DESIGN

4.1 Framework

In this section, we first depict the overview framework of VeMo, which is then followed by feature extraction of three components including (i) destination prediction, (ii) route inference, and (iii) speed estimation. Specifically, in the route and speed inference, we utilize a joint optimization model to learn the historical routes and speeds with only transaction data, to obtain necessary training data. These features are fed into a learner to learn predictors for different tasks.

Fig 7 shows the framework of our system, which consists of two parts: offline learning and online prediction. In the offline learning, all the data come from three data sources including the road map, historical ETC transactions and context data. In the feature extraction, we divide all the features into three categories, which are individual features, crowd features and context features. The feature summary is presented in Table 2 (next page). By fitting these features into the learner, we train three predictors for destinations, routes and speeds. By combining these predictor together, we predict the real-time locations of vehicles. In the next three subsections, we introduce three predictors for destinations, routes, and speeds from a feature perspective respectively, and then unify them together with a prediction model based on Mondrian Forest.

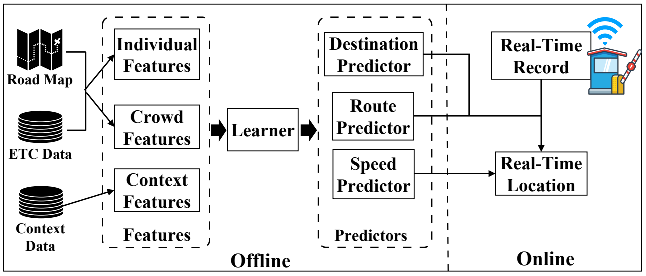

Fig 7: Framework

4.2 Destination Predictor

Destination prediction has been intensively studied in the past few years [17, 57]. The existing approaches for the vehicle destination prediction mainly rely on transition probabilities between different locations through learning historical trajectories using various Markov chain based models [12, 28]. One of the key prerequisites is that there should be enough historical data of individuals to learn the transition probabilities. However, in our context, most vehicles only have limited historical data (as we discussed in Section 2), which makes it hard to directly apply the Markov chain based models. To address this issue, we explore more individual features, crowd features and context features.

Individual Features: Since individual destinations essentially are based on personal habits, we utilize a set of individual features.

- **Historical Destinations:** As shown in Fig 1, the mobility patterns of most individuals in terms of destinations are relatively stable. Therefore, historical destinations may largely represent their future destinations.
- **Time Factor:** Considering the commute pattern in Fig 1 when the entropy is equal to 1, by introducing the entering time factor, the uncertainty of destinations is reduced. We use half-hour time windows to split one day into 48 time slots.
- **Vehicle Type:** It has three values: cars, buses, and trucks. Intuitively, the trucks most probably go to areas with high cargo demand (e.g., industry parks) and buses often go to areas with a dense population (e.g., commercial districts or transportation hubs). Fig 8

Table 2: Mobility Modeling Features

	Individual Features	Crowd Features	Context Features
Destination Predictor (Section 4.2)	Historical Destinations, Time of Day, Vehicle Type	Crowd Destination Distributions	Day of Week Weekday/weekend
Route Predictor (Section 4.4)	Historical Routes (Section 4.3), Driving Experience, Time of Day	Crowd Route Distributions	Day of Week Traffic Speed
Speed Predictor (Section 4.5)	Historical Driving Speed (Section 4.3), Time of Day, Vehicle Type	Crowd Speed Distributions	Weekday/weekend, Weather

shows the proportion of different vehicle types in different types of areas. We select three exemplary areas and calculate the proportional of different types of vehicle whose destinations are in the area. We found only a few trucks go to the commercial areas; cars and buses contribute major volume in the commercial and transportation hub areas, respectively.

Crowd Features: The individual vehicle's historical data can be very sparse (as we suggested in Section 2). we try to use the crowd destinations to provide complementary information. Fig 9 shows the possible destinations from the same origins by half of all the vehicles. We found almost 50% of vehicles go to at most 10 destinations. It indicates lots of vehicles from the same origins share the similar destinations, which can be used to infer the destination of a vehicle without any historical destination data.

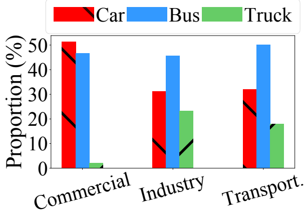

Fig 8: Dest. variance

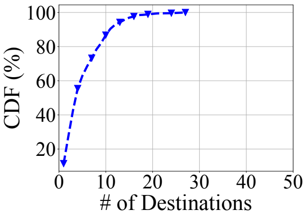

Fig 9: Crowd dest.

Context Features: We further consider other context features, i.e., the day of the week, weekday/weekend, holidays, that may have impacts on the destination choices. We choose the 10 most popular destinations for each origin and compare the rank of these destinations in a regular day with that in other days with different contexts using the measurement of Normalized Discounted Cumulative Gain *NDCG* [25]. The lower the *NDCG*, the lower similarity the destination choices. Fig 10 shows that the measurement between weekdays, weekend, and holiday. The holiday has very different destination choices compared to other days. In the early morning and the late afternoon of weekends, the *NDCG* is also lower than that of weekdays. It suggests these factors have impacts on people's choice of the destinations.

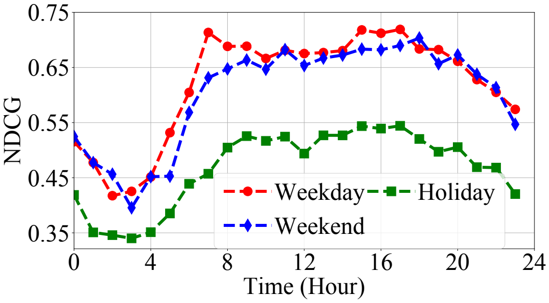

Fig 10: Context impacts

4.3 Historical Route and Speed Learning

As we discuss in Section 2, the reason that previous works are not feasible in our setting is that the historical routes and speeds of individual vehicles cannot be observed by the ETC system. In order to learn the mobility of individual vehicles, we propose a joint learning approach to obtain the historical routes and speeds of vehicles simultaneously, which are utilized as training data to model the route choices and real-time speeds in Section 4.4 and Section 4.5.

Several studies [15] [54] have been done to investigate the relationship between the travel routes and real-time speeds, which found the route of vehicles can be inferred with only speeds information. This finding indicates the strong correlation between the routes and speeds, which inspires our idea to learn the routes and speeds simultaneously.

To achieve this, we first present a few preliminaries.

- Time: we divide a day of 24 hours into K time slots(t) (i.e., each time slot is equal to 10 minutes).
- Location: we split the highway road networks into M equal length road segments(s) (i.e., 1 km).
- Speed: instead of treating the speed as a continuous variable, we discretize it into H discrete integer speed(v) by the smallest unit of 1 km/h (e.g., if the speed limit is 120 km/h, then we can have 121 different speed values ranging from 0 to 120km/h).

In this way, the states of vehicles in each trip on highways can be presented as a sequence of states <t, s, v> between the origin and the destination. As an example of the trip

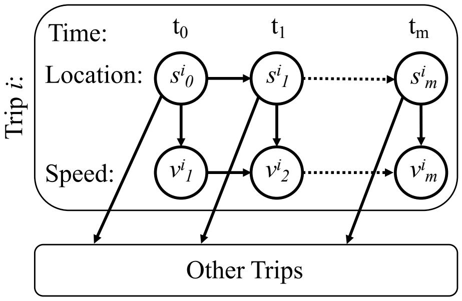

Fig 11: Route and speed Correlation

i in Fig 11, the vehicle enters the highway from the road segment s_0 at the time t_0 and exits the highway from the road segment s_m at the time t_m. It is worth mentioning that, in other trips, vehicles can be at the same location as the same time as the trip i. Then our objective is to infer the most likely state sequence of each trip. The solution is motivated by the key observation that at the same time multiple vehicles are traveling on the same road segments and their real-time speeds can be considered as samples of the speed distribution. The following insights reveal the characteristics of the distribution.

- **Speeds distribution on the road segment**: By analyzing the sample GPS trajectories, we observe that speeds of vehicles on the same road segment follow a normal distribution, which is also validated in other contexts [20].
- **Speed STD distribution**: Moreover, as shown in Fig 3, we also observe strong normality of the speed.

Since both insights show the normality, to quantify them, we utilize Kolmogorov-Smirnov test to test the normality. Specifically, the states of different trips within the same time and location are grouped as samples to test the normality of speed on the road segments. For the speed STD distribution insight, it is measured as suggested in Section 2. Then all the STDs are considered as samples to test the normality.

Given the normality test of both the speed distribution in each road segment and speed STD distribution of all the vehicles, our problem can be transformed into an optimization problem to find the best state sequence combination for the maximization of the number of the acceptance of normality tests. Suppose we have N trips with J vehicles, we formulate the problem as following:

$$\underset{sc}{\text{maximize}} \quad \sum_{i}^{N} 1_A(Rnorm(sc)) + \sum_{j}^{J} 1_A(Snorm(sc))$$

where sc is the combination of the state sequences of different trips, $Rnorm$ is a test function to check the normality of the speed distributions, $Snorm$ is a test function to check the normality of the speed STD distribution. 1_A is an indicator function of the test acceptance.

A straightforward approach to solve the optimization problem is to search all the possible state sequence combinations. For each trip, the possible state sequence is $K \times M \times H$. Then the total search space is $O(N^{K \times M \times H})$, which is time consuming to search. To reduce the search space, we introduce several simple but effective heuristics to guide the search.

- State sequences constrained by routes: Shown in Fig 2, there is a limited number of routes between origins and destinations, which naturally reduces the search space of possible location sequences.
- Spatial smoothness: Constrained by the structure of the road network and the speed limit, the next location of the vehicle can be the reachable road segments under the speed limit. (e.g., suppose the speed limit is 120km/h, the next location in 5 minutes can only be the road segments within a range of 5 minutes×120 km/h = 10 km.)

Given these heuristics, we perform a standard search algorithm (e.g., DFS) to find the best combination of the state sequence. Then the historical routes can be obtained by concatenating the locations in each trip and speeds can be directly obtained from the state sequence.

4.4 Route Predictor

Similar to the destination prediction, we study the features from three perspectives: individual features, crowd features and context features.

Individual Features: We utilize the following features for the route prediction at the individual level.

- **Historical Routes:** Based on a previous study, people are more reluctant to change their regular routes if they have more experience with these routes [6], which indicates historical routes are most likely to be their future routes given the same origin and destination.
- **Driving Experience:** Empirically, experienced people are good at finding the best routes [6]. We quantify the experience by two factors: (i) the frequency of driving on highways, which can be obtained from historical ETC transactions; (ii) the saved travel time compared with the average travel time, which can also be computed from historical ETC data.
- **Time Factor:** Empirically, people generally have their own estimations about the route traffic at a different time, e.g., taking a detour during the rush hour to avoid the traffic. It affects their future route choices.

Crowd Features: For those people who have no or only limited historical data, we incorporate the route choices of crowds to infer their route choice. Specifically, we use the probability of historical crowds' routes between particular origin/destination at the certain time.

Context Features: People's route choices are affected by the real-time context [5], i.e., the day of the week and real-time traffic speed, which can be estimated with ETC transactions in the recent past.

4.5 Speed Predictor

In this subsection, we introduce different features that are correlated to the real-time speed. The key idea is to learn the relation between individual driving speed and other features (e.g., crowd speed) in order to predict the real-time speed given all these features.

4.5.1 ***Features:*** We introduce our features on the individual, crowd, and context level.

Individual Features: Since the driving speed is essentially based on people's behaviors, we define a set of individual vehicle's features.

- **Historical Driving Speed:** As shown in Fig 3, the driving speed is relatively stable for a particular person. We use their average speeds of historical trips to reflect their general driving speed.
- **Vehicle Type:** This feature reflects the vehicle's type (i.e., cars, buses, trucks). Intuitively, the driving speed of cars should be higher than trucks and buses. Fig 12 also validates this intuition.
- **Time Factor:** Fig 12 shows that the driving speed varies at the different time of a day, which is mainly due to the different traffic conditions.

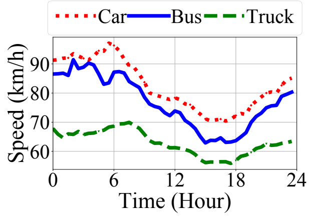

Fig 12: Speed variance

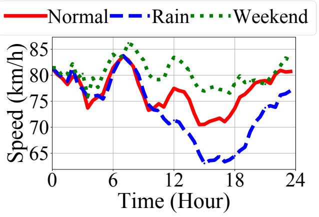

Fig 13: Context impacts

Crowd Features: People may behave differently under different traffic conditions. Instead of studying the detailed behavior patterns of individuals, which may have many factors to discuss, we directly investigate the correlation between the individual speed and crowd speed. Fig 14 shows the Pearson correlation of the individual speed and crowd traffic speed. More than 80% of vehicles have at least 0.89 correlation coefficient with the crowd traffic speed. Motivated by the strong correlation, the crowd traffic speed is an important feature to estimate the individual driving speeds on specific edges. Therefore, we extract the features of the vehicle speed samples, which are incorporated to estimate the crowd traffic speed. Instead of using the average crowd traffic speed (which may cause an estimation bias), we consider the statistic values of the crowd traffic speed distribution, including minimum, lower fourth, median, upper fourth and maximum of the samples. We reply on the crowd features to learn how the driver would react under different situations, in order to predict the real-time speed in the future.

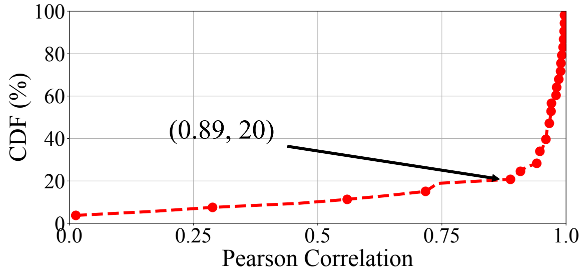

Fig 14: Speed Correlation

Context Features: Besides the vehicle-related features, we also consider other factors that may have impacts on the driving speed, including weather and weekday/weekend. As shown in Fig 13, the speed is decreased by 10% at most in the rainy day and increased by 5% on weekend. This is reasonable because people tend to drive slower when raining and fewer people use highways to work on the weekend, which makes the highways less congested.

4.6 Learning with Mondrian Forest

Mondrian forests [27] is an online random forest model using Mondrian processes to construct ensembles of decision trees. Compared to the offline or online random forest [27], it provides the ability to process online data and online updates faster and more accurately. Compared with other algorithms, the Mondrian forests model has the following advantages:

- It is more robust to heterogeneous features. In our data input, we have both numerical variables (i.e., speed) and categorical values (i.e., vehicle type, weather, weekday/weekend). These variables can be input into the model directly without conversion or normalization.
- It provides self-check on the importance of the features during the training stage. For example, such as the weather condition and holidays, these variables would only have high importance under certain conditions with a low frequency.
- Compared to other neural based model (e.g., deep neural network), the results are more explainable because of the internally used decision tree [14].

For different tasks (i.e., destination prediction, route inference, speed estimation), we fit all the extracted features into Mondrian forests and learn three predictors to work collaboratively on the real-time location prediction, which is illustrated in Section 4.7. Even we choose Mondrian forests, our system is flexible to many machine learning methods. The more important aspect is the analysis process and find the effective features.

4.7 Put them all together

In the previous sections, we have conducted an analysis of the three key tasks: destination prediction, route inference and speed estimation. Based on multiple extracted features, we learn three predictors *d-predictor, r-predictor, s-predictor* for each of the tasks, perceptively. The procedure of real-time location prediction is described in Algorithm 1.

Algorithm 1: Real-time Location Prediction

```
Input  : d-predictor: the destination predictor,
         r-predictor: the route predictor,
         s-predictor: the speed predictor,
         $entrance$: the entering toll station,
         $interval$: the updating time interval
         $t_0$: the entering time.
Output : real-time locations
destination ← d-predictor given entrance
route ← r-predictor given destination
$distance = 0$
while $distance < route.length$ do
    speed ← s-predictor at $t_i$
    $distance$ += $speed \times interval$
    location ← match distance to route
end
```

5 EVALUATION

In this section, we introduce our data-driven evaluation in terms of methodology and results.

5.1 Evaluation Methodology

Ground Truth: To obtain the ground truth of real-time vehicle locations, we introduce another real word dataset with detailed GPS trajectories in Guangdong, which provide the real-time locations of 114 thousand vehicles including 75% cars, 13% buses and 12% trucks. These vehicles upload their real-time locations in every 10 to 30 seconds. The detailed data format is presented in Table 3. Fig 15 shows the trajectories visualization on the main roads in Guangdong. It shows our dataset covers most of the main roads, which can be utilized to evaluate the state-level mobility. Two highlighted areas are the two largest cities, i.e., Guangzhou and Shenzhen, which are densest areas in terms of vehicles. For each vehicle, we first apply a map matching algorithm [34] to map trajectories onto the road network. Then only the trajectories on highways are remained to obtain entering toll stations, exit toll stations, routes and real-time locations, which cover 20% of the vehicles in our dataset. Since the training and testing are conducted on different datasets, we do not need to split the datasets for cross-validation.

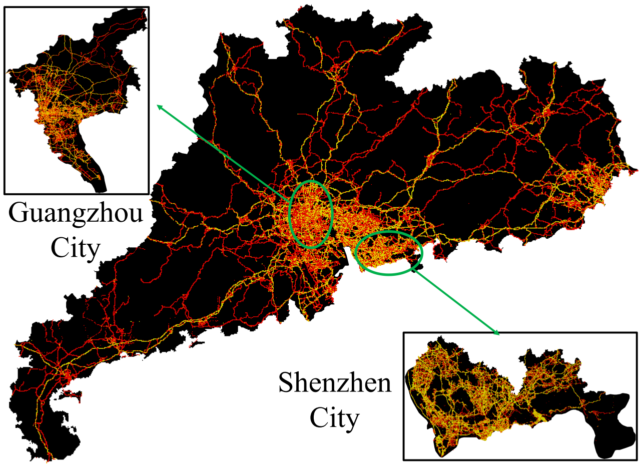

Fig 15: Ground truth visualization

Table 3: Ground Truth Format

Field	Value	Field	Value
Id	P0SF51B4GU	Type	Car/Bus/Truck
Longitude	113.402904	Latitude	23.167894
Time	2016-06-01 00:00:34		
75% cars, 13% buses, 12% trucks		#Vehicle: 114k	

Evaluation Metrics: For each component, we define the evaluation metrics as follows:

- Destination and Route prediction:

$$accuracy = \frac{\#prediction_{correct}}{\#prediction_{all}} \times 100\% \qquad (1)$$

where $\#prediction_{correct}$ is the number of corrected prediction and $\#prediction_{all}$ is the total number.

- Speed Prediction:

$$accuracy = 1 - \frac{|speed_{predict} - speed_{actual}|}{speed_{actual}} \qquad (2)$$

where $speed_{predict}$ is the predicted speed and $speed_{actual}$ is the ground truth.

- Real-Time Location Prediction: we quantify the location accuracy by measuring the percentage of predicted locations within the accuracy threshold (i.e., 100 meters) of the ground truth considering the GPS errors every 15 seconds (i.e., the average uploading time interval of data in ground truth) [19]. The accuracy formula is defined as

$$accuracy = \frac{\#prediction_{correct}}{\#prediction_{all}} \times 100\% \qquad (3)$$

.

Baselines for Intermediate Results: For the three individual prediction components, i.e., predictions for destinations, routes, and speeds, since we utilize a unified algorithm for all of them, we evaluate them from the perspective of a learning model by comparing it with the other learning models. The selected learning models are presented as follows, and each of them is representative of a group of methods with the similar bases:

- **Empirical Estimation (Emp)**: The baseline represents the prediction based on the naive empirical knowledge. For the destination and route prediction, we consider the most frequently visited destinations and routes. For speed prediction, we utilize their historical average speed.
- **Bayesian Network (Bayes)** [14]: Bayesian network is a typical graph-based algorithm, which is representative for the probability based models.
- **Neural Network (Neural)** [14]: Neural network represents the models that focus on learning the linear or non-linear combination between features and targets.

Baselines for End-to-End Results: For the overall performance of the real-time locations, we choose the baselines based on two principles: (i) static infrastructure based methods; (ii) mobile sensor based methods.

- **STrack**: This baseline represents a wide range of static infrastructure based methods, e.g., cameras [60]. Considering traffic cameras are set to detect motoring offenses without open location information, we implement STrack by assuming a given percentage of edges (defined in Section 3) have been installed with cameras that can track vehicles. In the middle of each edge, we assume one traffic camera is installed that can recognize vehicle plates. The real-time locations of vehicles are obtained as being observed by the cameras. For the location estimation of vehicles between cameras, we assume they are uniformly distributed on the roads between cameras. Fig 16 shows the edge length in the highway road network. Different percentages are also evaluated to show the performance.

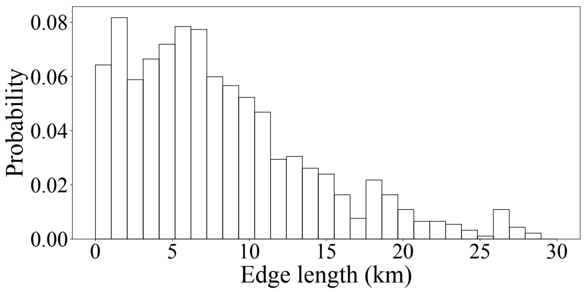

Fig 16: Edge length

- **CTrack** [41]: This baseline aims to track individual vehicles based on cellular networks by periodical communications between onboard cellphones and cell towers. Based on the locations of communicated cell towers, it infers the locations of the cellphones (thus vehicles). The cell tower locations we use are located in Shenzhen City (shown in Fig 17), where the ETC system is also widely spread with 79 toll stations. We implement CTrack by assuming each vehicle has an onboard cellphone to interact with cell towers and follow the trajectory mapping algorithm in [41].

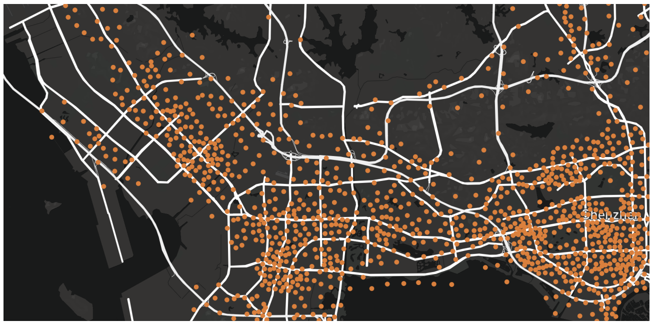

Fig 17: Cell Tower Locations in Shenzhen

Impacts of Factors: We evaluate several factors to show the impacts on the performance of VeMo,

- **Weather**: Weather condition is a factor that affects the driving behavior such as driving speed. We evaluate the accuracy in both regular day and extreme weather day (e.g., heavy rain).
- **Accuracy threshold**: Given different accuracy threshold to declare the accuracy, the performance may be varied. We choose several threshold values to show the accuracy changes.
- **Time factors:** The performance may vary at different time. We evaluate it in weekday, weekend and holiday.

- **Spatial factors:** Shown in the previous sections, different areas have different densities of toll stations and different volume of traffic. We evaluate VeMo at different areas in Guangdong, i.e., both the downtown areas and suburb areas.
- **Vehicle Types:** Different types of vehicles may have different challenges of prediction. We evaluate this factor by applying VeMo on different types of vehicles, i.e., car, bus, truck.

In the next subsection, we first compare the three individual components with the baselines (i.e., Naive Empirical, Bayesian, Neural). Then we comparing the overall performance of the real-time location prediction with *STrack* and *CTrack* followed by the impacts of the factors.

5.2 Evaluation Results

5.2.1 Efficiency. We implement VeMo on a server with Intel Xeon E5-1660 3.00GHz CPU and 32GB RAM in 16 threads. After loading all the data, the training process takes 450 seconds. The speed prediction is 500 times per thread every second on average, which can satisfy the real-time need of 4 million daily transactions.

5.2.2 Real-Time Edge-Cloud Design. Since most of the applications built on our system require real-time response, it is necessary to have real-time cloud components. Even it is feasible to conduct prediction in a powerful server, however, it is challenging to update the model in the cloud in real-time. Our solution is to combine both the cloud (i.e., center servers) and the edges (i.e., computer systems in the toll stations).
Cloud: All the data is stored in the cloud system for security issues. As the new data collected in the edges, the data is transmitted to the cloud through Ethernet. All the trained models are also stored in the cloud to distribute to the edges.
Edge: Given the truth that a vehicle only appears in a few toll stations, we could pre-distribute the trained individual models to top frequent edges according to historical records. Considering the online updating feature of our model, we update the model directly in the edge devices. Then the model itself is transmitted back the cloud to distribute to other station. Generally, a vehicle leaving the toll station would not get back to the highways immediately. There there is enough time to transmit the model to the cloud.

5.2.3 Comparison to baselines. We evaluate both individual predictors and overall location predictor. For each of the individual component, we evaluate it by comparing it to the three baselines, respectively. Then three predictors work collaboratively to predict the locations of vehicles.
(i) Destination prediction: Fig 18 plots the result of the destination prediction. It shows VeMo has better performance than other three learning models with an average performance gain of 11%. The Bayesian network baseline performs better than the neural network, which means the probability relationship is better to model the destination prediction problems. Moreover, the naive empirical baseline achieves 60% accuracy during the day time, which suggests the destination choices are relatively stable on highways.

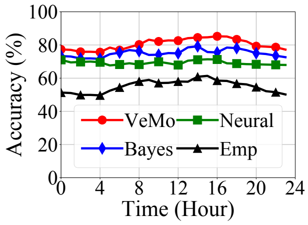

Fig 18: Destination pred.

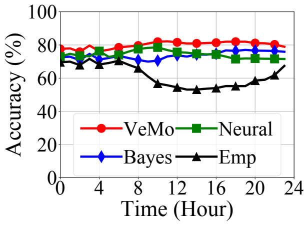

Fig 19: Route pred.

(ii) Route prediction: Fig 19 presents the result of the route prediction. Compared to the other three baselines, VeMo achieves an average performance gain of 6%. It suggests the performance does not vary much in terms of different learning models. The naive empirical baseline has similar performance in the early morning but poor performance during the daytime, which means the route choices are flexible when there is heavy traffic.
(iii) Speed prediction: Fig 20 shows the result of average speed prediction. VeMo has an average performance gain of 17%. During the day time, the accuracy is higher, because the heavy traffic constrains the speed variation. The naive empirical baseline shows poorer performance during the daytime because the empirical knowledge cannot obtain the real-time traffic information. Moreover, the neural network baseline is better than the Bayesian-based baseline, which suggests the advantage of linear combination based method on the speed prediction tasks.

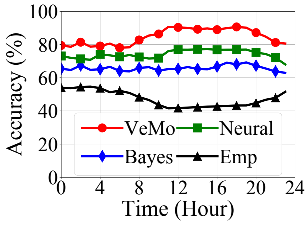

Fig 20: Speed pred.

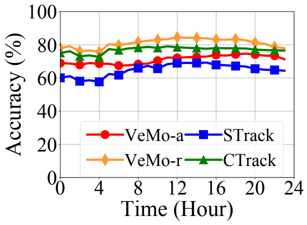

Fig 21: Location pred.

(iv) Location prediction performance: After the individual predictors' evaluation, we combine them together to evaluate the real-time locations of vehicles. Since the route has dominating impacts on the locations of vehicles, to show more sophisticated evaluations, we test the accuracy of both (i) the vehicles (*VeMo-a*) and (ii) those vehicles with correctly predicted routes (*VeMo-r*). Then we compare them with *STrack* and *CTrack*. Fig 21 plots the evaluation results. Considering the vehicles with correctly predicted routes, VeMo (shown as *VeMo-r*) has the average accuracy about 82%. The reason that VeMo has similar accuracy as *CTrack* is that the baseline experiment is conducted inner city, which has a dense cell tower distribution. Even including all the vehicles (shown as *VeMo-a*), VeMo achieves average accuracy of 70%, which is still at the same level of *STrack*, which means VeMo can be an alternative solution of *STrack* without introducing extra infrastructures.

We also evaluate the impacts of coverage percentage of *STrack*, and show the result in Fig 22. After the coverage percentage increases to 50%, *STrack* achieves better performance. Since it is expensive to provide such high infrastructure coverage, VeMo outperforms *STrack* in terms of feasibility.

5.2.4 Impacts of factors. Five factors are evaluated including accuracy threshold, weather, time factors, spatial factors and vehicle types. The metrics are the same as the equation 3.

(i) Impacts of Accuracy Threshold: We choose accuracy threshold including 25, 50, 100, 150 meters to show how the accuracy changes in Fig. 23. The lower the line, the better the accuracy. We found higher thresholds lead to higher accuracy. 100-meter and 150-meter thresholds have closed accuracy while 25-meter and 100-meter thresholds have obvious lower accuracy.

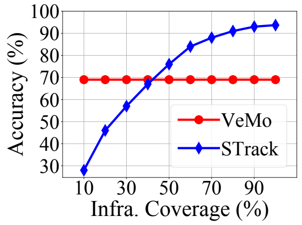

Fig 22: % of infra.

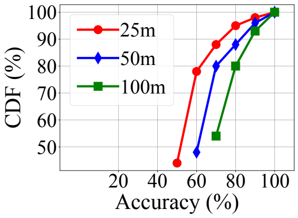

Fig 23: Accuracy threshold

(ii) Impacts of Weather: We select one day with heavy rain and compare the result with that of a regular day. We surprisingly found the rain even increase the prediction accuracy. Since people tend to drive slowly in the heavy rain, the individual speed is reduced and there is a smaller range of speed variance on the way, which benefits the prediction accuracy.

(iii) Impacts of Time Factors: Fig 25 shows the performance of VeMo in weekday, weekend and holiday. The accuracy in weekday and weekend is similar. Moreover, the performance in the holiday is different than other days, especially during the morning. This is because the destination choices are less predictable on the holidays when people generally do not follow regular mobility patterns.

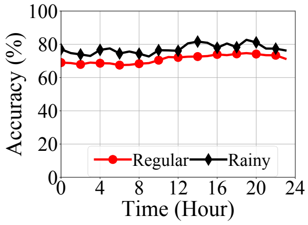

Fig 24: Weather Impact

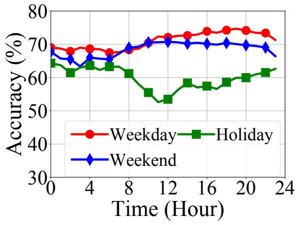

Fig 25: Time impact

(iv) Impacts of Spatial Factors: We investigate the performance of VeMo in both downtown areas and suburb areas and show the result in Fig 26. In the early morning, two areas have similar accuracy. During the daytime starting at 8 am, the performance in the downtown areas decreases. This is because the road structure is more complex in that areas, which makes the route prediction less accurate.

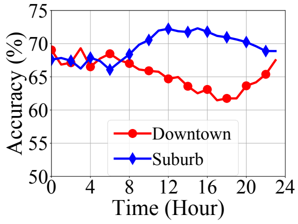

Fig 26: Spatial impact

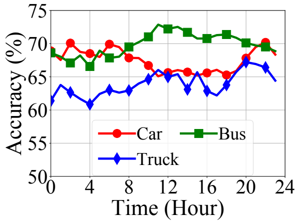

Fig 27: Type impact

(v) Impacts of Vehicle Types: Fig 27 shows the performance of different types of vehicles. Trucks have the lowest accuracy because they generally have longer travel distances and irregular mobility patterns (e.g., one truck may travel between different areas for cargo services as long as there are demands of cargo transportation). Buses have higher accuracy because they have most regular mobility patterns compared to trucks and cars. Cars' accuracy decreases during the daytime because they generally travel inner cities, which is impacted by both traffic conditions and road structures.

6 DISCUSSIONS

Lessons Learned: Based on our results in Guandong, we learned a few valuable lessons.

- vehicle's mobility pattern in terms of destinations can be identified as three major groups, single-time travel vehicles, commuting vehicles and multi-destination vehicles;
- the overall distributions of both speed STDs cross vehicles and speeds on the road segment follow strong

normality, which can be considered as constrains to infer the routes and speeds simultaneously;

- individual highway speeds vary based on driving behaviors, but is highly correlated with generic traffic speeds. The overall derivation of individual speeds follows a Gaussian-like distribution;
- Both individual and crowd level features (e.g., historical destinations, routes, vehicle types, driving experience, etc) are helpful to predict vehicle locations along with contexts (e.g., time of day, day of week, weather).

Why ETC Data Only? In this work, we explore the possibility of using ETC data alone to predict real-time vehicle locations by solving some uncertainty issues, e.g., unknown routes. This is because ETC systems can provide a full penetration rate transparently based on data already collected. Moreover, the ETC based toll system is universal and exist almost everywhere even in developing countries where satellite images or mobile infrastructure is not well penetrated. If combined with other datasets even with small scale, e.g., GPS data from highway service vehicles or traffic camera data, we may be able to further improve our accuracy significantly. But due to space limitation, we focus on our core contribution on ETC data.

Data Collection and Privacy Protection: In this work, the ETC data we utilized are collected by an ETC company, which is a part of Guangdong Highway Administration Agency; the GPS data we utilized are collected by an insurance company under drivers' agreement, which is a part of usage-based insurances for discounts. In the ETC service agreement and highway usage agreement, people are notified that their data will be used to analyze traffic patterns and improve traffic condition. Instead of tracking individual vehicles, our project is to understand and improve traffic patterns, which potentially benefits all ETC users.

Real-world Applications: Our project is part of highway improvement initiative Guangdong Highway Administration Agency. One key application of our vehicle location prediction is to address the traffic congestion on Guangdong highway during peak hours. Based on our results, we can estimate the number of vehicles on each edge of the highways by predicting the real-time locations of all vehicles, which can be utilized to design applications such as ramp meters [37] and adaptive toll strategies [48].

Rest Area Stop: Since the ETC data only give the station to station travel duration, it may contain time a person spent at rest areas, which cannot be directly obtained from the ETC transaction data. But based on our dataset, we found that 76% of transactions have a duration less than 60 minutes, during which a person is unlikely to go to rest areas unless it is a part of a longer trip starting outside Guangdong Highway System. Unfortunately, we cannot validate this assumption based on ETC data alone. However, with GPS data, we found that for the trips shorter than 60 minutes, only 8% of vehicles went to rest areas. It indicates rest area stops may not have significant impacts on our results.

Limitations and Open Problems: We discuss some limitations and open problems related to our system.

- Each highway system has its unique geographic and demographic features, so data-driven insights and our evaluation results we have in Guangdong may not apply to other highways with very different features. However, we believe the techniques we develop to predict destinations, infer routes, and estimate speeds are generic and can be applied to other highway systems if their data are available.
- Our system work in a controlled environment, i.e., a highway system with both entering and existing records. Therefore the same technique may not be applied to local streets without toll booths to track every vehicle enter or leave a street. In this case, additional data, e.g., partial GPS, can be combined with our solution for prediction. However, we believe our solution can be generalized to stationary sensors that can capture the vehicle's passing (such as cameras, cellular tower, etc.).
- Even ETC systems only capture the vehicle twice, it still has the privacy issue of exposing locations. However, compared with GPS based solutions, it is a better/low-cost privacy-reserved approach since ETC data have already been collected as a mandatory process for billing; whereas other approaches need new devices or dedicate data collection process with potential continuous location collection.
- Our solution can help to detect the abnormal events when there are a certain number of vehicles (potentially spatial correlated vehicles, e.g., their real-time predicted locations are the same road segments) do not leave the highway after their travel duration on highways. But since in this work, we focused on the fundamental location prediction and did not try to explicitly handle the anomalies, which would be a good direction for our future work.
- Our solution replies on the historical data of vehicles to learn their driving behaviors. In the ETC system, we found there is only 9% of the new vehicles without any historical data after 10-day data accumulation, which is very small number of vehicles. For those without historical data, we can only infer their behaviors according to majority behaviors (i.e., crowd features). It is still an interesting open problem.

7 RELATED WORK

The most related work to this paper is a system called SharedEdge [53] where ETC data are utilized to infer the generic traffic speed on each highway edge. However, VeMo is different in both the objective and methods. In particular, VeMo focuses on location prediction, whereas SharedEdge focuses on speed prediction. Even though VeMo also has a speed estimation but it focuses on individual speeds, whereas SharedEdge focuses on generic speed. Moreover, there a good body of literature for vehicle mobility modeling and predictions based on various sensing infrastructures [41][63][33]. Shown in the Table 4, we divide them into two major parts: GPS based approaches and None-GPS based approaches, where highlight our position with the full penetration.

Table 4: Vehicular Mobility Survey

	Aggregate	Individual	
Mobile	[59] [26] [22] [56]	[58] [64] [40] [2] [50] [31]	
Static	[33] [60] [53] [36]	Partial Penetration	Full Penetration
		[42] [41] [8]	Our work

Static Infrastructure: Static infrastructures, e.g., traffic cameras [60], cell towers [41], WiFi access point [42], are widely used for vehicle mobility modeling. Some communication related works are also studies based on the static infrastructure [4] [29] [3] [13]. However, one disadvantage of these approaches is either the lower coverage of infrastructures [44] or the low penetration of the apps that are used to interact with the infrastructure. Without the installation of extra infrastructures for a full coverage, it is difficult to model and predict the mobility of all the vehicles. Most of these approaches require continuous movement detection, which is not always satisfied in the real world [30]. Compared with existing work, our approach makes use of the existing infrastructures to predict vehicle mobility without extra cost. All the vehicles entering the highways are detected, which does not require the installation of interaction apps. The requirement of only single real-time observations, e.g., entrance to a highway, largely increases the feasibility of our approach in the real world. Some approaches such as cell phone network may have potential to infer traffic condition with low cost. But normally the cellphone data are not available for highway administrators. They can only use the data collected by themselves. Further, the cell phone network cannot be narrowed down to vehicular mobility since the driver and passage cannot be distinguished from the cell phone data only, which may introduce extra bias.

Mobile Infrastructure: Mobile infrastructures, i.e., smartphones and onboard devices, are extensively studied to understand both individual and groups of vehicles. [42] [64] [21] use smartphones to track vehicles in real time. The inference on mobility is studied in details by smartphone data [40]. [2] estimates the urban traffic using vehicular fleets with onboard devices. [49] implements regular vehicle tracking through commercial vehicles with onboard devices. Other works such as crowdsoucing information collection and energy issues can also benefit from the mobile infrastructures [10] [61] [32]. However, these approaches are either limited by low penetration rates of apps [38] or focus on the aggregated level [2], and typically rise privacy issues of exposing vehicle GPS data [64].

Vehicular mobility on the highways is also studied in the transportation community (i.e., the destination and speed prediction [9] [47] [46] [62]). However, previous works mainly focused on the aggregated traffic characteristics, such as origin-destination matrix or traffic speed on the road segments. Different from these works, our system aims at the mobility model of individual vehicles, which requires microscope analysis of the vehicle mobility pattern. In addition, our result can be extended to the aggregated traffic characteristics by aggregating the individual vehicles. Moreover, we utilize extra dataset as ground truth, which avoid the drawbacks of cross-validation in the previous works.

Summary: Based on our discussion, most of the existing approaches are limited by extra deployment cost, low penetration rates, or requirement for privacy-prying GPS locations. In contrast, our approach makes use of the existing infrastructures with a full penetration rate to track individuals with only sparse location information, which makes our work significantly different from the existing approaches.

8 CONCLUSION

In this paper, we focus on vehicle location prediction on large-scale highway systems with sparse ETC data. In particular, we motivate and design a novel system called VeMo with three key technical components for the destination prediction, route inference, and speed estimation. More importantly, we implement and evaluate VeMo based on the large-scale data in the Guangdong highway network in China, utilizing an large-scale ETC system with 773 stations and a large-scale vehicle fleet with GPS data as ground truth. We advance state-of-the-art vehicle mobility modeling approaches by some key lessons we learned. We envision our results may benefit various applications including highway anomaly detection and risk assessment that we have been working with our partner.

ACKNOWLEDGMENTS

The authors would like to thank anonymous reviewers for their valuable comments and suggestions. This work is partially supported by NSF 1849238 and Rutgers Global Center.

REFERENCES

[1] 511nj. 2017. 511NJ: Get Connected and go! http://www.511nj.org/cameras.aspx. (2017).

[2] Javed Aslam, Sejoon Lim, Xinghao Pan, and Daniela Rus. 2012. City-scale traffic estimation from a roving sensor network. In *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. ACM, 141–154.

[3] Fan Bai, Keyvan Rezaei Moghadam, and Bhaskar Krishnamachari. 2015. A tale of two cities-Characterizing social community structures of fleet vehicles for modeling V2V information dissemination.. In *SECON*. 506–514.

[4] Aruna Balasubramanian, Ratul Mahajan, Arun Venkataramani, Brian Neil Levine, and John Zahorjan. 2008. Interactive wifi connectivity for moving vehicles. *ACM SIGCOMM Computer Communication Review* 38, 4 (2008), 427–438.

[5] Eran Ben-Elia, Roberta Di Pace, Gennaro N Bifulco, and Yoram Shiftan. 2013. The impact of travel information's accuracy on route-choice. *Transportation Research Part C: Emerging Technologies* 26 (2013), 146–159.

[6] Eran Ben-Elia and Yoram Shiftan. 2010. Which road do I take? A learning-based model of route-choice behavior with real-time information. *Transportation Research Part A: Policy and Practice* 44, 4 (2010), 249–264.

[7] Nicholas Capurso, Eric Elsken, Donnell Payne, and Liran Ma. 2014. Poster: A robust vehicular accident detection system using inexpensive portable devices. In *MobiSys*.

[8] Gayathri Chandrasekaran, Tam Vu, Alexander Varshavsky, Marco Gruteser, Richard P Martin, Jie Yang, and Yingying Chen. 2011. Tracking vehicular speed variations by warping mobile phone signal strengths. In *Pervasive Computing and Communications (PerCom), 2011 IEEE International Conference on*. IEEE, 213–221.

[9] Gang-Len Chang and Jifeng Wu. 1994. Recursive estimation of time-varying origin-destination flows from traffic counts in freeway corridors. *Transportation Research Part B: Methodological* 28, 2 (1994), 141–160.

[10] Xi Chen, Xiaopei Wu, Xiang-Yang Li, Yuan He, and Yunhao Liu. 2014. Privacy-preserving high-quality map generation with participatory sensing. In *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2310–2318.

[11] Rinku Dewri, Prasad Annadata, Wisam Eltarjaman, and Ramakrishna Thurimella. 2013. Inferring trip destinations from driving habits data. In *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*. ACM, 267–272.

[12] Trinh Minh Tri Do and Daniel Gatica-Perez. 2012. Contextual conditional models for smartphone-based human mobility prediction. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. ACM, 163–172.

[13] Zhihan Fang, Fan Zhang, Ling Yin, and Desheng Zhang. 2018. Multi-Cell: Urban Population Modeling Based on Multiple Cellphone Networks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 106.

[14] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. 2001. *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

[15] Xianyi Gao, Bernhard Firner, Shridatt Sugrim, Victor Kaiser-Pendergrast, Yulong Yang, and Janne Lindqvist. 2014. Elastic pathing: Your speed is enough to track you. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 975–986.

[16] GHSA. 2018. Speed Cameras on Highways. https://www.ghsa.org/state-laws/issues/speed-and-red-light-cameras. (2018).

[17] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. 2008. Understanding individual human mobility patterns. *nature* 453, 7196 (2008), 779.

[18] Google. 2017. Google Map. https://www.google.com/maps. (2017).

[19] Mohinder S Grewal, Lawrence R Weill, and Angus P Andrews. 2007. *Global positioning systems, inertial navigation, and integration*. John Wiley & Sons.

[20] Bruce Hellinga, Pedram Izadpanah, Hiroyuki Takada, and Liping Fu. 2008. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. *Transportation Research Part C: Emerging Technologies* 16, 6 (2008), 768–782.

[21] Bo-Jhang Ho, Paul Martin, Prashanth Swaminathan, and Mani Srivastava. 2015. From pressure to path: Barometer-based vehicle tracking. In *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM, 65–74.

[22] Minh X Hoang, Yu Zheng, and Ambuj K Singh. 2016. FCCF: forecasting citywide crowd flows based on big data. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 6.

[23] Dijiang Huang, Swaroop Shere, and Soyoung Ahn. 2010. Dynamic highway congestion detection and prediction based on shock waves. In *Proceedings of the seventh ACM international workshop on VehiculAr InterNETworking*. 11–20.

[24] Insurance Information Institute. 2017. Highway safety. https://www.iii.org/fact-statistic/facts-statistics-highway-safety. (2017).

[25] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[26] Amin Vahedian Khezerlou, Xun Zhou, Lufan Li, Zubair Shafiq, Alex X Liu, and Fan Zhang. 2017. A traffic flow approach to early detection of gathering events: Comprehensive results. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 6 (2017), 74.

[27] Balaji Lakshminarayanan, Daniel M Roy, and Yee Whye Teh. 2014. Mondrian forests: Efficient online random forests. In *Advances in neural information processing systems*. 3140–3148.

[28] Mu Li, Amr Ahmed, and Alexander J Smola. 2015. Inferring movement trajectories from GPS snippets. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 325–334.

[29] Yuanjie Li, Chunyi Peng, Zengwen Yuan, Jiayao Li, Haotian Deng, and Tao Wang. 2016. Mobileinsight: Extracting and analyzing cellular network information on smartphones. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. ACM, 202–215.

[30] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. 2015. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 33.

[31] Luyang Liu, Hongyu Li, Jian Liu, Cagdas Karatas, Yan Wang, Marco Gruteser, Yingying Chen, and Richard P Martin. 2017. BigRoad: Scaling Road Data Acquisition for Dependable Self-Driving. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 371–384.

[32] Suhas Mathur, Tong Jin, Nikhil Kasturirangan, Janani Chandrasekaran, Wenzhi Xue, Marco Gruteser, and Wade Trappe. 2010. Parknet: drive-by sensing of road-side parking statistics. In *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, 123–136.

[33] Chuishi Meng, Xiuwen Yi, Lu Su, Jing Gao, and Yu Zheng. 2017. City-wide Traffic Volume Inference with Loop Detector Data and Taxi Trajectories. (2017).

[34] Paul Newson and John Krumm. 2009. Hidden Markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*. ACM, 336–343.

[35] NYC Department of Transportation. 2017. New York City Camera. http://dotsignals.org. (2017).

[36] Zhou Qin, Zhihan Fang, Yunhuai Liu, Chang Tan, Wei Chang, and Desheng Zhang. 2018. EXIMIUS: A Measurement Framework for Explicit and Implicit Urban Traffic Sensing. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems*. ACM, 1–14.

[37] Thorsten Schmidt-Dumont and Jan H van Vuuren. 2015. Decentralised reinforcement learning for ramp metering and variable speed limits on highways. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS* 14, 8 (2015), 1.

[38] Statista. 2017. Number of internet users who used the internet for route planning,to access maps or road maps (e.g. Google Maps) in Germany from 2013 to 2016, by frequency (in millions). https://www.statista.com/statistics/432169/online-route-planning-and-map-usage-eg-google-maps-germany. (2017).

[39] Saber Taghvaeeyan and Rajesh Rajamani. 2014. Portable roadside sensors for vehicle counting, classification, and speed measurement. *IEEE Transactions on Intelligent Transportation Systems* 15, 1 (2014), 73–83.

[40] Arvind Thiagarajan, James Biagioni, Tomas Gerlich, and Jakob Eriksson. 2010. Cooperative transit tracking using smart-phones. In *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*. ACM, 85–98.

[41] Arvind Thiagarajan, Lenin Ravindranath, Hari Balakrishnan, Samuel Madden, and Lewis Girod. 2011. Accurate, low-energy trajectory mapping for mobile devices. (2011).

[42] Arvind Thiagarajan, Lenin Ravindranath, Katrina LaCurts, Samuel Madden, Hari Balakrishnan, Sivan Toledo, and Jakob Eriksson. 2009. VTrack: accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM conference on embedded networked sensor systems*. ACM, 85–98.

[43] US DOT. 2017. https://www.itscosts.its.dot.gov/ITS/benecost.nsf/DisplayRUCByUnitCostElementUnadjusted?ReadForm&UnitCostElement=CCTV+Video+Camera&Subsystem=Roadside+Detection+(RS-D). (2017).

[44] US DOT. 2017. CV Pilots Take Advantage of Multiple Communication Media. https://www.its.dot.gov/pilots/cvp_media.htm. (2017).

[45] Yiwen Wan, Yan Huang, and Bill Buckles. 2014. Camera calibration and vehicle tracking: Highway traffic video analytics. *Transportation Research Part C: Emerging Technologies* 44 (2014), 202–213.

[46] Xiang Wang, Kang An, Liang Tang, and Xiaohong Chen. 2015. Short term prediction of freeway exiting volume based on SVM and KNN. *International Journal of Transportation Science and Technology* 4, 3 (2015), 337–352.

[47] Jiancheng Weng, Rongliang Yuan, Ru Wang, and Chang Wang. 2014. Freeway travel speed calculation model based on ETC transaction data. *Computational intelligence and neuroscience* 2014 (2014), 48.

[48] Wired. 2017. Dynamic tolls in Virginia. https://www.wired.com/story/virginia-i66-toll-road/. (2017).

[49] Xiaoyang Xie, Yu Yang, Zhihan Fang, Guang Wang, Fan Zhang, Fan Zhang, Yunhuai Liu, and Desheng Zhang. 2018. coSense: Collaborative Urban-Scale Vehicle Sensing Based on Heterogeneous Fleets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 196.

[50] Xiaoyang Xie, Fan Zhang, and Desheng Zhang. 2018. PrivateHunt: Multi-Source Data-Driven Dispatching in For-Hire Vehicle Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 45.

[51] Mengwen Xu, Dong Wang, and Jian Li. 2016. DESTPRE: a data-driven approach to destination prediction for taxi rides. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 729–739.

[52] Lei Yang, Yekui Chen, Xiang-Yang Li, Chaowei Xiao, Mo Li, and Yunhao Liu. 2014. Tagoram: Real-time tracking of mobile RFID tags to high precision using COTS devices. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 237–248.

[53] Yu Yang, Fan Zhang, and Desheng Zhang. 2018. SharedEdge: GPS-Free Fine-Grained Travel Time Estimation in State-Level Highway Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 48.

[54] Jiadi Yu, Hongzi Zhu, Haofu Han, Yingying Jennifer Chen, Jie Yang, Yanmin Zhu, Zhongyang Chen, Guangtao Xue, and Minglu Li. 2016. Senspeed: Sensing driving conditions to estimate vehicle speed in urban environments. *IEEE Transactions on Mobile Computing* 15, 1 (2016), 202–216.

[55] Hui Zang and Jean Bolot. 2011. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*. ACM, 145–156.

[56] Desheng Zhang, Tian He, Shan Lin, Sirajum Munir, and John A Stankovic. 2017. Taxi-passenger-demand modeling based on big data from a roving sensor network. *IEEE Transactions on Big Data* 3, 3 (2017), 362–374.

[57] Desheng Zhang, Jun Huang, Ye Li, Fan Zhang, Chengzhong Xu, and Tian He. 2014. Exploring human mobility with multi-source data at extremely large metropolitan scales. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 201–212.

[58] Desheng Zhang, Juanjuan Zhao, Fan Zhang, and Tian He. 2015. UrbanCPS: a cyber-physical system based on multi-source big infrastructure data for heterogeneous model integration. In *Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems*. ACM, 238–247.

[59] Junbo Zhang, Yu Zheng, and Dekang Qi. 2017. Deep Spatio-Temporal Residual Networks for Citywide Crowd Flows Prediction.. In *AAAI*. 1655–1661.

[60] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. 2017. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 3687–3696.

[61] Tan Zhang, Ning Leng, and Suman Banerjee. 2014. A vehicle-based measurement framework for enhancing whitespace spectrum databases. In *Proceedings of the 20th annual international conference on Mobile computing and networking*. ACM, 17–28.

[62] Jiandong Zhao, Yuan Gao, Yujie Guo, and Zhiming Bai. 2018. Travel time prediction of expressway based on multi-dimensional data and the particle swarm optimization–autoregressive moving average with exogenous input model. *Advances in Mechanical Engineering* 10, 2 (2018), 1687814018760932.

[63] Mingmin Zhao, Tao Ye, Ruipeng Gao, Fan Ye, Yizhou Wang, and Guojie Luo. 2015. Vetrack: Real time vehicle tracking in uninstrumented indoor environments. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*. ACM, 99–112.

[64] Yiran Zhao, Shen Li, Shaohan Hu, Lu Su, Shuochao Yao, Huajie Shao, Hongwei Wang, and Tarek Abdelzaher. 2017. Greendrive: A smartphone-based intelligent speed adaptation system with real-time traffic signal prediction. In *Cyber-Physical Systems (ICCPS), 2017 ACM/IEEE 8th International Conference on*. IEEE, 229–238.