A Dual-Flow Attentive Network with Feature Crossing for Chained Trip Purpose Inference

SUBMISSION DATE / POSTED DATE

08-03-2022 / 01-11-2022

CITATION

Lyu, Suxing; Han, Tianyang; Li, Peiran; Luo, Xingyu; Kusakabe, Takahiko (2022): A Dual-Flow Attentive Network with Feature Crossing for Chained Trip Purpose Inference. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.19322279.v3

DOI

10.36227/techrxiv.19322279.v3

A Dual-Flow Attentive Network with Feature Crossing for Chained Trip Purpose Inference

Suxing Lyu, Tianyang Han, Peiran Li, Xingyu Luo, and Takahiko Kusakabe*

***Abstract*—Trip purpose is essential information supporting tasks in intelligent transportation systems, such as travel behaviour comprehension, location-based service, and urban planning. The observation of trip purpose is a necessary aspect of travel surveys. However, owing to the sampling volume, survey budget, and survey frequency, relying solely on travel surveys in the era of big data is a difficult task. There has long been a demand for an accurate, generalizable, and robust inference method for trip purposes. Although existing studies contributed significant efforts to improve the trip purpose inference, the potential of leveraging the trip chain is insufficient. The spatial correlations and chaining patterns hidden in travelled zones are worthy of further exploration. The unequal importance within trip chains has not been clearly represented. Additionally, complex activity-zone mutual interdependence has not been considered in previous models. Herein, we propose a framework-Dual-Flow Attentive Network with Feature Crossing (DACross), specifically for inferring the chained trip purpose. We form trip chains innovatively that treat trip activities and travelled geographic zones as two chains with mutual interactions. We propose DACross, which consists of two parallel attentive branches and a co-attentive feature crossing module, for fully learning the intra- and inter-chain dependencies. We conducted extensive experiments on four large-scale real-world datasets to evaluate not only the performance of DACross but also the generalizability of the proposed framework among different cities and scenarios. Notably, the Experimental results prove the overall superiority of the proposed DACross.**

***Index Terms*—Travel behaviour, trip purpose inference, deep learning, intelligent transportation systems.**

I. Introduction

Trip purpose plays a critical role in human mobility, which has been recognised as a crucial behavioural pattern [1]. In this perspective, trip purpose can contribute to urban planning and mobility information systems by facilitating travel behaviour analysis. From the perspective of urban planners and policymaking, travel behaviour is tightly correlated with the urban structure [2] and socio-demographic attributes of travellers [3], [4]. Fine-grained information and modelling in travel behaviour studies are expected to better estimate and forecast travel demand, which is essential for addressing urban planning-related issues [5], such as site selection and infrastructure function evaluation. On the other hand, modern mobility information systems, by determining the motivation that enables people to travel, can improve transport mode detection [6], origin-destination (OD) extraction [7], and destination prediction [8], [9], and past studies have proved that only knowing the when and where is not sufficient.

Although the trip purpose represents important information and has considerable potential for use, the long-standing issue has been that the trip purpose is challenging to observe. Since the 1980s [10], behavioural surveys (particularly the travel diary surveys) have been the most reliable way to acquire trip purposes. As an essential item asked in the questionnaires of diary travel surveys, trip purposes in daily travel are determined through inquiry and subsequently recorded. Moreover, by leveraging mobile devices, researchers conducted a variant of travel diary surveys, termed the Prompted Recall (PR) surveys [6], [11], [12], which segment trips from mobile records and ask respondents to answer their trip purposes online to improve the survey quality [13]. Currently, PR surveys provide more realistic records and continuously contribute to travel surveys; nevertheless, the difficulties of conducting broad and continuous long-term observation using these specially designed surveys are yet to be solved. As early as 2001, researchers began a proof-of-concept study examining the feasibility of deriving trip purpose from movement data [14]. This study shows that for accurate trip-purpose inference, mobile movement data can be an efficient replacement for traditional travel surveys, which can respond passively to the demand for broad and long-term observation. Hence, the research on methods to infer trip purpose accurately has long been motivated.

In past research, considerable efforts have been invested to develop an accurate means for trip purpose inference. For instance, importing precise and abundant extra information was approved as meaningful in this context. In this case, precise geographical contexts, such as the commercial point of interest (POI) distributions and geo-located social media records, were considered [15]–[17]. Moreover, acquiring travellers' long-term travel histories and detailed personal attributes has significantly improved the task [18]. These data augmentation methods cannot always be satised in real-world scenarios, as they could be cost-intensive for and pose privacy-sensitive issues to the target. Therefore, instead of importing complicated data pre-requirement, we focused on investigating the modelling framework in this study.

There have been two types of modelling frameworks, the trip-level and trip chain-level (Fig. 1). The trip-level methods are constructed on the assumption that each trip is independent. The independent assumption ignores the contextual

Suxing Lyu is with Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan

Tianyang Han is with Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8565, Japan

Peiran Li is with Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan

Xingyu Luo is with the Department of System Biology, George Mason University, Fairfax, VA, 22032, USA.

Takahiko Kusakabe is with Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8568, Japan

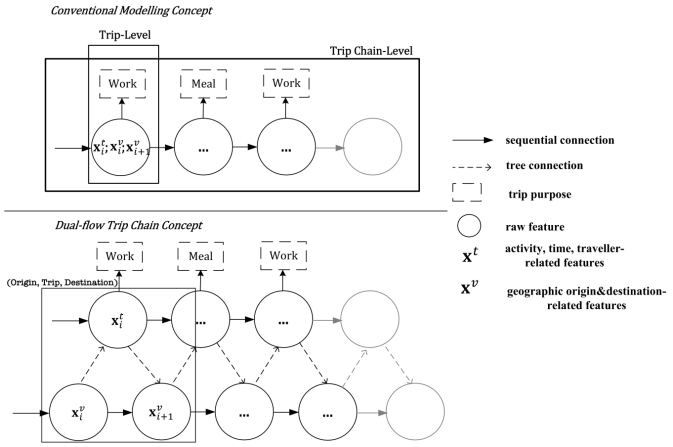


Fig. 1. Sketch of the three modelling concepts for trip purpose inference.

patterns in the chain. Hereof, there could be counterintuitive orders of trip purposes, for example, $\mathrm{Home} \rightarrow \mathrm{Work} \rightarrow \mathrm{Home} \rightarrow \mathrm{Work}$. In recent years, the trip chain-level methods have received more attention; however, compared with the trip-level methods, these methods have rarely been studied as only a few studies have been reported. The trip chain-level methods attach importance to the chaining behaviour, which is an approved significant benefit to our task [19].

Regardless of the trip- or trip chain-level methods, previous studies have rarely considered separating the behavioural contexts (trips) and geographical contexts (zones)[1]. Considering the most essential definition, the trip purpose inference can be established on the three-element tuple $(\mathrm{Origin}, \mathrm{Trip}, \mathrm{Destination})$. Most of the previous studies simply concatenated all of three together, nevertheless, we consider this approach will result in an insufficient usage of the inherent feature patterns. In particular, concatenating three elements together will flatten the hierarchical structure that a trip naturally represents a behavioural connection between two functional zones. Moreover, in this manner, the chaining patterns in zones will be changed to be subsidiary to trips, where rich information of both spatial and temporal views in chained zones [20] will be lost.

To overcome the aforementioned issues and achieve our targets, we propose a new modelling concept implemented on deep learning technologies, which are seen to be effective and efficient for extracting high-level features [21]–[24]. The proposed modelling concept is based on the chained tuple relations, termed the dual-flow architecture, in which we processed the chained zones and trips simultaneously. We consequently introduce a framework-**D**ual-Flow **A**ttentive Network with Feature **Cross**ing (DACross), for chained trip purpose inference. The DACross primarily takes account into 1) augmenting the usage of zone information by modelling zones' geographical adjacencies in graphs and maintaining the zones' self-organised chain in order of daily travel, and 2) importing the unequal importance within chains (the intradependencies) and the mutual interactions between trips and zones (the interdependencies) by the attentive methods. The DACross was designed to not rely on traveller information for achieving relatively accurate trip-purpose inference, which is convenient to apply and adaptable to real-world application scenarios. Moreover, we decided to build our model on existing travel surveys, which are the largest and most complete datasets so far, following the diary-like format. Although compared to the emergent passively collected movement data, the volume of travel surveys is relatively small, model evaluations on these small data are crucial to bridge the gap to big data [25] up. Finally, we organise our contributions as follows:

- We initiated an attempt to better exploit the information of zones and to provide a new perspective, i.e., dual-flow architecture, of general modelling for trip purpose inference, which was solely based on the essential three-element tuples.
- We proposed a new framework, DACross, for chained trip-purpose inference. DACross consisted of two equally important branches with a co-attentive feature crossing module, which leveraged interdependencies within the trip chain.
- We conducted extensive experiments using four large-scale real-world datasets. Furthermore, we evaluated the practicability of DACross through comprehensive analysis. The results showed that the proposed DACross outperformed various comparative studies.

The paper is organised as followings. In Section II, we report a comprehensive literature review with further explanation of our research motivations. The necessary definitions and terms are introduced in Section III. The detailed framework construction is described in Section IV. In Section V and VI, we introduce the details of the experiments and summarise the results on quantitative and qualitative analysis. Finally, we conclude the paper in Section VII.

II. Trip Purpose Inference Review

The trip purpose inference has been a long-standing topic. Nowadays, there has been considerable evolved progress on this topic. We primarily described the related works by the evolutionary tendency of methods and modelling concepts. During the early stage, deterministic rule-based methods [11], [14], [26]–[28] were initially developed. Such rule-based methods heavily relied on the design of complicated rules to determine the trip purposes with the search and query functions of the Geographic Information System (GIS). These rule-based methods usually can only correspond to a few categories of trip purposes. Subsequently, in some special cases, given the long-time travel trajectory histories or travel surveys as prior references, the Bayesian methods [17], [29], [30] were proposed for modelling trip purposes from a probabilistic view. These methods were developed for specific transport modes, and thus, were difficult to generalise to whole-day travel.

In the past decade, machine learning-based methods have [6], [12], [15], [16], [18], [31], [32] became mainstream, as data conditions have improved correspondingly. These methods were established on the trip-level modelling, and attention was paid for importing the extra information to improve the

[1]Note that a trip refers to a unit of movement activity made while moving from the origin to the destination. Correspondingly, a zone denotes the choice of the geographic area or boundary in the real world.

performance. Owing to the elaborated data pre-requirement and heavy feature engineering, the models' practicability is constrained. Moreover, the trip chain-level formulation has become popular in recent years. The Hidden Markov model (HMM) and variants [8], [21], [22], [33]–[35] are widely applied for chained trip purpose inference. [19] validated the significant benefits considering the trip sequence on trip purpose inference. However, current studies of the trip chains-level modelling are yet to be sufficient. For instance, the unsupervised learning for specific transport modes [22], [33] still kept the difficulty to be applied on the whole day travel as relatively a few numbers of trip purposes can be inferred. The poor performance of inferring long and complex trip chain [21] was validated. The reliable and generalised trip chain-level inference is temporarily absent.

Most recently, a few studies have focused on implementing deep learning technologies for inferring the trip purpose. The burgeoning development of deep neural networks leads to a flexible way for extracting high-level features in a data-driven manner. Thus, it can be easily extended to fit various scenarios. For instance, the studies [21], [22] of the trip chain-level formulation consist of parts of deep neural networks to ensure abundant high-level features as the inputs to the following HMM. [24] proposed a framework to leverage Check-in and POI semantics, which were applied to the taxi trip purpose inference. The potential merits of using deep neural networks prompted our study, focusing on the trip chain-level modelling by the proposed DACross framework. Moreover, we argue that the unequal importance of chained trip purposes has not been explicitly considered in previous studies; nevertheless, the unequal importance has long been explored and discussed. For instance, in terms of trip generation [36], there are priorities for different trip purposes, and several compulsory purposes primarily decide the composition of a trip chain [37]. Furthermore, discretionary purposes are distributed among the travellers' schedules and the time budget between compulsory purposes. This concept was also sustained under the destination choice [38], which can be correlated to the travelled zone order. Therefore, we propose the DACross, which reflects the unequal importance of attentive modelling intra- and inter-dependencies in deep learning. The DACross is constructed to leverage the inherent chaining patterns within the trip chain and eventually provide a whole-day travel inference under the naive data condition, which is critical to practicability in real-world scenarios. Based on our previous study [23], we maintain the mining of graph-constructed zone information and dive into the dual-flow architecture.

III. Preliminaries

We first define the graph for geographic adjacency and the method of chaining the trip. Utilizing this strategy, the formulation of the chained trip-purpose inference task can be obtained. Herein, we use the terms sequence and chain to express the same meaning.

Definition 1 **Geographic Adjacency Graph:** Given a certain research area (such as a city), we partitioned the area into numerous small zones by using administrative boundaries, street blocks, and mesh grids. We defined these small zones as a set of nodes $\mathcal{V}$. Regarding the adjacent relationships among the zones, we defined the pairs of adjacencies as a set of edges $\mathcal{E}$. Consequently, we present the geographic adjacency graph as $G(\mathcal{V}, \mathcal{E})$. For each node v in the graph, a vector $\mathbf{x}^v \in \mathbb{R}^{d_v}$ denotes its attributes, such as numbers of POIs and categories of land use. Our geographic adjacency graph has no weights on edges and no limitations on their directions. Thus, given an edge e_{ij} between v_i and v_j, e_{ij} only denotes the adjacent relationship, where $e_{ij} \equiv e_{ji}$.

Definition 2 **Trip Chain:** Within a certain period and given a specific traveller p, a trip chain refers to the scheduling of micro-level activities of the traveller in time and space. We defined two sequences of observations, a sequence of trips $\mathcal{T}^{(p)} = \{t_1^{(p)}, ..., t_N^{(p)}\}$ and a sequence of travelled zones $\mathcal{V}^{(p)} = \{v_1^{(p)}, ..., v_M^{(p)}\} | \mathcal{V}^{(p)} \subset \mathcal{V}$, which compose a trip chain $\mathcal{C}^{(p)} = (\mathcal{T}^{(p)}, \mathcal{V}^{(p)})$. Specifically, as each single trip $t_i^{(p)}$ bridges the origin $v_i^{(p)}$ and the destination $v_{i+1}^{(p)}$, we have $M = N + 1$. When a traveller finishes a trip at the same zone, $v_i^{(p)} \equiv v_{i+1}^{(p)}$ is established. Similar to the node vector, the vector $\mathbf{x}^t \in \mathbb{R}^{d_t}$ denotes the trip's attributes, such as travel speed, distance, duration, and travel mode.

Problem: ***Chained Trip Purpose Inference*****.** Given a geographic adjacency graph $G = \{\mathcal{V}, \mathcal{E}\}$ and trip chain $\mathcal{C}^{(p)} = (\mathcal{T}^{(p)}, \mathcal{V}^{(p)})$, the goal of our task is to train a deep neural network $f(\cdot)$ for inferring the trip chaining purposes, $\{y_1^{(p)}, ..., y_N^{(p)}\} = f(G, \mathcal{C}^{(p)})$.

IV. Methodology

A. Framework Overview

An overview of the DACross framework is shown in Figure 2. In accordance with the dual-flow concept, we split a trip chain into batches of trips and zones and treat them equally. In this manner, we processed the travelled zones delicately and enriched their feature representations. Through feature crossing, we further enhanced two branches with the full usage of inter-trip chain dependencies to improve the robustness of the inference. We describe four modules designed for specific objectives as follows:

Input Feature Embedding: We started with coarse-level feature embedding before entering the principal modules. Initially, raw feature vectors of trips and zones may have different dimensions and belong to different latent feature spaces, which are troublesome for later processing. Hence, they are organised to have the same dimensions in this module. Moreover, inspired by [23], [24], [39], we employed a 1-layer Graph Neural Network (GNN), specifically, a Graph Attention Network (GAT) [40], to capture zones with adjacent relevance to their neighbours. The GAT can emphasise most useful information on neighbours through the attention mechanism. Moreover, the objective of this module is to roughly aggregate adjacent spatial information into the travelled zones.

Intra-Sequence Encoders: We instantiated the shared sequential encoder of Transformer [41] twice (marked as the 1st encoder and 2nd encoder in the Fig. 2). As mentioned before, multiple trip purposes in a trip chain are considered to have different importance and priority [36], [37]. This should

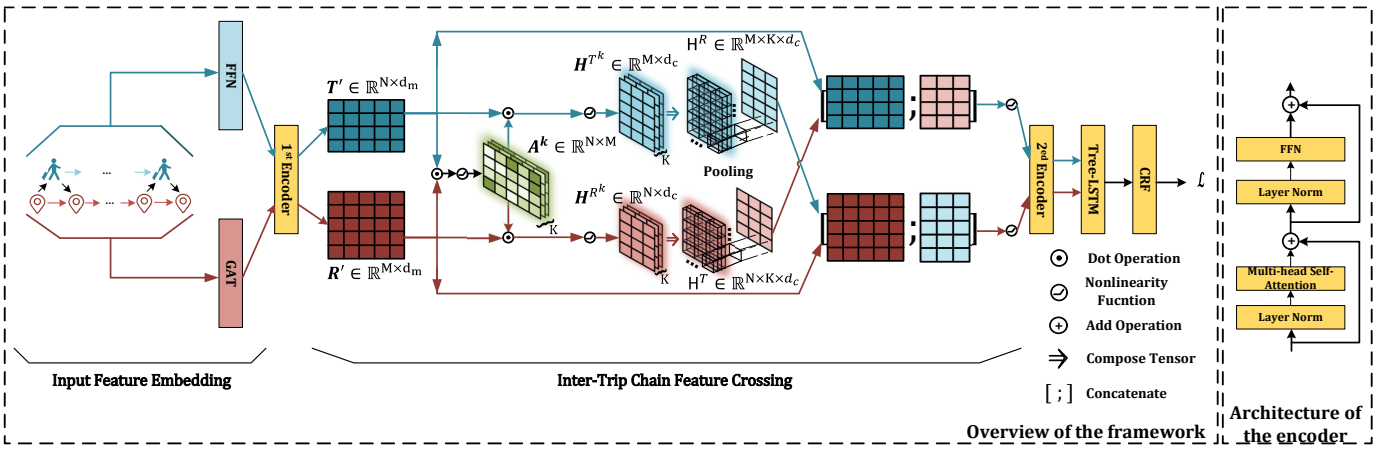


Fig. 2. The overview framework of the proposed DACross. The blue colour scheme denotes the procedures of trips, and the red colour scheme denotes the zones. The yellow colour scheme denotes the shared or common procedures. We implemented an encoder that differed from the locations of the Layer Norm from [41], in which it was proved to have a better training efficiency [48].

also be true for the regional sequence. Certain zones are not crucial for the chain, but they occur when the traveller just drops in. These encoders are responsible for capturing the intra-relevance and intra-attractiveness. The first encoder is designed for lightly encoding feature matrices from the input embeddings in which these matrices only contain the information inside themselves. Therefore, we did not consider any interactions between trips and zones within the first encoder. After enhancing the inter-chain interactive information, the second encoder is designed for the same targets as the first encoder but devotes itself to capture the changed information.

Inter-Trip Chain Feature Crossing: This module is inspired by the co-attentive structures, which have been utilised widely for Natural Language Processing (NLP) tasks, such as Q&A [42], [43] and machine comprehension [44]. These tasks, besides NLP [45], generally contain multiple sources as inputs, and their objectives ultimately lead to a synthetic output that judges, evaluates, or answers a binary label or a similarity score. As inherent correlations may exist across multiple inputs, the co-attention mechanism can markup mutual effective information to improve the performance. In our scenario, we argue that within the same trip chain, the occurrence of a certain trip purpose may be indirectly related to travelled zones. Referring to [46], we propose a parametric co-attentive feature crossing module. This module practises our modelling perspective and concurrently enhances the information of inherent correlations through addition.

Tree-Structure Aggregation: Following our previous work [23], this module seeks to organise and aggregate trips and zones for outputs. The natural tree structure that bridges the origin and the destination of the trip is suitable for utilizing Tree-LSTM [47].

B. Dual-Flow Attentive Network with Feature Crossing

1) Input Feature Embedding: We first introduced the processing using GAT to aggregate geographic adjacent information. The GAT linearly transforms node feature vectors and pairs the target node and neighbours to calculate attention scores. The 1-layer GAT with multi-head attention mechanism can be formulated as:

$$\alpha_{ij}^{k} = \frac{\exp(\sigma(\mathbf{a}^{\mathbf{k}}[\boldsymbol{W}_3^{I^k}\mathbf{x}_i^v; \boldsymbol{W}_3^{I^k}\mathbf{x}_j^v]))}{\sum_{l\in\mathcal{N}_i}\exp(\sigma(\mathbf{a}^{\mathbf{k}}[\boldsymbol{W}_3^{I^k}\mathbf{x}_i^v; \boldsymbol{W}_3^{I^k}\mathbf{x}_l^v]))} \tag{1}$$

Consequently, in our case, for each zone $v^{(p)} \in \mathcal{V}^{(p)}$, the aggregated zone vector is as follows:

$$\mathbf{h}^v = \sigma(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}(v^{(p)})}\alpha_{ij}^{k}\boldsymbol{W}_3^{I^k}\mathbf{x}^v) \tag{2}$$

, where $\mathcal{N}(\cdot)$ denotes the 1^{st} order sampling function. From the perspective of a graph, the index i denotes the target node (zone), and the index j denotes the 1^{st} order neighbours (including the target node i) to the target node. In our case, we set the parameter matrices $\mathbf{a}^{\mathbf{k}}$ and $\boldsymbol{W}_3^{I^k}$ with dimensions of $\mathbb{R}^{2d_m}$ and $\mathbb{R}^{d_v\times d_m}$, respectively. For convenience, we only applied average operation in Equation 2, regardless of the optional concatenation operation reported in [40].

For each trip $t^{(p)} \in \mathcal{T}^{(p)}$, we fed it into a feed-forward network (FFN) as:

$$\mathbf{h}^t = \sigma(\boldsymbol{W}_1^I\mathbf{x}^t + \mathbf{b}_1^I)\boldsymbol{W}_2^I + \mathbf{b}_2^I \tag{3}$$

, where $\boldsymbol{W}_1^I, \boldsymbol{W}_2^I \in \mathbb{R}^{d_t\times 4d_m}, \mathbb{R}^{4d_m\times d_m}$ are trainable matrices for the input feature extraction of trips, and $\boldsymbol{b}_1^I, \boldsymbol{b}_2^I \in \mathbb{R}^{4d_m}, \mathbb{R}^{d_m}$ denote biases. The simple 2-layer FFN can project raw trip features to the same space with the outputs from GAT.

Consequently, the vectors $\mathbf{h}_i^t, \mathbf{h}_i^v \in \mathbb{R}^{d_m}$ are prepared. Throughout this article, σ denotes nonlinearity functions (e.g. ReLU, ELU, and LeakyReLU). The index $k = 1, ..., K$ denotes the k-th head when multi-head setting is available, and $[;]$ denotes the concatenation operation.

2) Intra-Sequence Encoder without Interdependence: In the implementation, specifically, we employed the encoder architecture from Transformer, which fully leveraged the intra-sequence self-attention mechanism. Similar to GAT, the self-attention mechanism can be beneficial from a multi-head setting. Following the notations in the origin, given an input matrix $\boldsymbol{X} \in \mathbb{R}^{L\times d_m}$, the single sequential encoder layer is calculated as:

$$\begin{bmatrix} \boldsymbol{Q}^k \\ \boldsymbol{K}^k \\ \boldsymbol{V}^k \end{bmatrix} = \boldsymbol{X} \begin{bmatrix} \boldsymbol{W}^{Q^k} \\ \boldsymbol{W}^{K^k} \\ \boldsymbol{W}^{V^k} \end{bmatrix} \tag{4}$$

$$\mathrm{Att}^k(\boldsymbol{Q}^k, \boldsymbol{K}^k, \boldsymbol{V}^k) = \mathrm{softmax}(\frac{\boldsymbol{Q}^k {\boldsymbol{K}^k}^\top}{\sqrt{d_k}})\boldsymbol{V}^k \tag{5}$$

$$\mathrm{MultiHead}(\boldsymbol{X}) = [\mathrm{Att}^k; |k = 1, ..., K]\boldsymbol{W}^O \tag{6}$$

$$\boldsymbol{X}' = \sigma(\mathrm{MultiHead}(\boldsymbol{X})\boldsymbol{W}_1 + \mathbf{b}_1)\boldsymbol{W}_2 + \mathbf{b}_2 \tag{7}$$

, where $\boldsymbol{W}^{Q^k}, \boldsymbol{W}^{K^k}, \boldsymbol{W}^{V^k} \in \mathbb{R}^{d_m \times d_k}$ are linear transformation weights for the k-th head with $d_k = \frac{d_m}{K_1}$. $\boldsymbol{W}^O \in \mathbb{R}^{d_m}$ is the weight of organizing outputs from multiple heads. We set the weights of Equation 7 as $\boldsymbol{W}_1 \in \mathbb{R}^{d_m \times 2d_m}$ and $\boldsymbol{W}_1 \in \mathbb{R}^{2d_m \times d_m}$. Finally, we summarised Equations 4-7 into a single functional layer. A complete sequential encoder consecutively stacks multiple encoder layers. We set the number of layers as a controllable hyper-parameter L_1. Correspondingly, for notation simplification, we represent the stacked encoder as:

$$\begin{aligned} \boldsymbol{T}' &= \mathrm{Encoder}^1(\boldsymbol{T}) \\ \boldsymbol{R}' &= \mathrm{Encoder}^1(\boldsymbol{R}) \end{aligned} \tag{8}$$

, where the trip matrix $\boldsymbol{T} \in \mathbb{R}^{N \times d_m}$ and zones matrix $\boldsymbol{R} \in \mathbb{R}^{M \times d_m}$ are composed of $\mathbf{h}_i^t, i = 1, ..., N$ and $\mathbf{h}_j^v, j = 1, ..., M$, respectively.

3) Inter-Trip Chain Feature Crossing: The module is inspired by the co-attention mechanisms, specifically, the parametric co-attentive computation [42]. Instead of directly calculating the dot-product of two matrices, we acquired the affinity feature map $\boldsymbol{A}$ using a learnable weight matrix $\boldsymbol{W}^A \in \mathbb{R}^{d_m \times d_m}$. Following the multi-head settings to capture various aspects, the k-th head's affinity feature map $\boldsymbol{A}^k \in \mathbb{R}^{N \times M}$ is defined as

$$\boldsymbol{A}^k = \tanh(\boldsymbol{T}' \boldsymbol{W}^{A^k} {\boldsymbol{R}'}^\top) \tag{9}$$

Consequently, the $\boldsymbol{A}^k$ is used to indicate mutual effects. We process these mutual attention weights as:

$$\boldsymbol{H}^{T^k} = \sigma({\boldsymbol{A}^k}^\top (\boldsymbol{T}' \boldsymbol{W}^{T^k})) \tag{10}$$

$$\boldsymbol{H}^{R^k} = \sigma(\boldsymbol{A}^k (\boldsymbol{R}' \boldsymbol{W}^{R^k})) \tag{11}$$

At the k-th head, the trip-to-zone features are represented as $\boldsymbol{H}^{T^k} \in \mathbb{R}^{M \times d_c}$, and correspondingly the zone-to-trip features are represented as $\boldsymbol{H}^{R^k} \in \mathbb{R}^{N \times d_c}$. $\boldsymbol{W}^{T^k}, \boldsymbol{W}^{R^k} \in \mathbb{R}^{d_m \times d_c}$ are the weights for linear transformations. In this manner, we have acquired mutual correlated information already. In a multi-head setting, there are several aspects drawing trip-to-zone or opposite correlations. For simplicity, we desire to aggregate them together. Thus, we composed co-attentive features of all heads into two 3-dimensional tensors $\mathsf{H}^T \in \mathbb{R}^{M \times K_c \times d_c}$ and $\mathsf{H}^R \in \mathbb{R}^{N \times K_c \times d_c}$, respectively.

$$\boldsymbol{T}'' = \sigma([\boldsymbol{T}'; \mathrm{Pooling}(\mathsf{H}^R)]\boldsymbol{W}^{TO} + \mathbf{b}^{TO}) \tag{12}$$

$$\boldsymbol{R}'' = \sigma([\boldsymbol{R}'; \mathrm{Pooling}(\mathsf{H}^T)]\boldsymbol{W}^{RO} + \mathbf{b}^{RO}) \tag{13}$$

The pooling operations are applied on the 3rd dimension to aggregate various aspects. Subsequently, through a residual connection, we concatenated aggregated features back to the original inputs. Two 1-layer FFNs are set for fusing features. Hereof, we prepare $\boldsymbol{T}''$ and $\boldsymbol{R}''$, which contain each other's information that enhances the interactions. $\boldsymbol{W}^{TO}, \boldsymbol{W}^{RO} \in \mathbb{R}^{(d_m + d_c) \times d_m}$ and $\mathbf{b}^{TO}, \mathbf{b}^{RO} \in \mathbb{R}^{(d_m + d_c)}$ are trainable weights and biases. For computational efficiency, we set $d_c = \frac{d_m \times scale}{K_c}$. $scale$, and the number of heads K_c are set as hyper-parameters.

4) Inference Composition by Intra-Sequence Encoder: After the inter-trip chain feature crossing, intra-chain dependencies should be changed frequently. Moreover, we instantiated the other shared $\mathrm{Encoder}^2(\cdot)$ using Equation 8 to complete intra-sequence inference composition. Similar to the aforementioned step, we also introduced a hyper-parameter L_2 to control the number of stacked layers.

5) Tree-Structure Aggregation: By splitting matrices $\boldsymbol{T}'''$ and $\boldsymbol{R}'''$ back into vectors, we could acquire the processed trip vectors $\{\mathbf{h}_1^{t'}, ..., \mathbf{h}_N^{t'}\} \in \mathbb{R}^{d_m}$ and zone vectors $\{\mathbf{h}_1^{v'}, ..., \mathbf{h}_M^{v'}\} \in \mathbb{R}^{d_m}$. For each trip vector $\mathbf{h}_i^{t'}$ and its zones $\{\mathbf{h}_i^{v'}, \mathbf{h}_{i+1}^{v'}\}$, we aggregated the output vector $\mathbf{h}_i^O \in \mathbb{R}^{d_r}$ as:

$$\mathbf{h}_i^R = \mathbf{h}_i^{v'} + \mathbf{h}_{i+1}^{v'} \tag{14}$$

$$\mathbf{h}_i^{OI} = \sigma\left(\boldsymbol{W}^{OI}\mathbf{h}_i^{t'} + \mathbf{U}^{OI}\mathbf{h}_i^R + \mathbf{b}^{OI}\right) \tag{15}$$

$$\begin{aligned} \mathbf{h}_i^{OF} &= \sigma\left(\boldsymbol{W}^{OF}\mathbf{h}_i^{t'} + \boldsymbol{U}^{OF}\mathbf{h}_i^{v'} + \mathbf{b}^{OF}\right) \\ \mathbf{h}_{i+1}^{OF} &= \sigma\left(\boldsymbol{W}^{OF}\mathbf{h}_i^{t'} + \boldsymbol{U}^{OF}\mathbf{h}_{i+1}^{v'} + \mathbf{b}^{OF}\right) \end{aligned} \tag{16}$$

$$\mathbf{h}_i^{OO} = \sigma\left(\boldsymbol{W}^{OO}\mathbf{h}_i^{t'} + \boldsymbol{U}^{OO}\mathbf{h}_i^R + \mathbf{b}^{OO}\right) \tag{17}$$

$$\mathbf{h}_i^{OU} = \tanh\left(\boldsymbol{W}^{OU}\mathbf{h}_i^{t'} + \boldsymbol{U}^{OU}\mathbf{h}_i^R + \mathbf{b}^{OU}\right) \tag{18}$$

$$\mathbf{h}_i^{OC} = \mathbf{h}_i^{OI} \odot \mathbf{h}_i^{OU} + \sum_{j \in [i, i+1]} \mathbf{h}_j^{OF} \odot \mathbf{c}_j \tag{19}$$

$$\mathbf{h}_i^O = \mathbf{h}_i^{OO} \odot \tanh(\mathbf{h}_i^{OC}) \tag{20}$$

, where $\boldsymbol{W}^*, \boldsymbol{U}^*, \mathbf{b}^*$ denote the parameterised matrices of the Tree-LSTM unit, and they have the same dimensions as d_{agg}. Once the aggregations are processed, they are projected by a linear transformation, which confirms that the final outputs have the same dimension as the number of classes.

C. Training

In our scenario, the trips are set as chained. Consequently, during the training process, we aimed to keep this setting consistently. The training loss should be estimated regarding the rationality of the whole chain. Hence, the Conditional Random Field (CRF) [49], [50] was added in the end to maintain the chaining pattern throughout. Instead of acquiring distinct accumulated Cross Entropy loss values, we constrained the chaining pattern via CRF-Loss as:

Dataset	Geo-Adjacency Graph			Trip Chain				Zone			
	#nodes	#edges	avg of degrees	#trips	#chains	length	avg of length	avg of area (km^2)	median of area (km^2)	avg of #POIs	median of #POIs
Tokyo 23	3,192	9,772	3.06	78,449	30,152	[1,22]	2.60	0.20	0.17	15.47	8
Downtown Yokohama	1,780	5,348	3.00	31,531	12,386	[1,23]	2.55	0.27	0.16	10.64	5
Twin Cities	8,600	27,018	3.14	186,099	43,829	[1,38]	4.24	45.91	2.37	4.92	2
Chicago Metropolitan Area	6,669	21,705	3.25	57,277	17,193	[1,30]	3.33	3.12	0.51	1.50	1

TABLE I
STATISTICAL INFORMATION OF THE EXPERIMENTED DATASETS.

$$\mathrm{S}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} (\boldsymbol{A}_{i-1,i} + \boldsymbol{P}_{i,y_i}) \tag{21}$$

, where $\mathrm{S}(\cdot)$ is a score function, $\mathbf{x}$ denotes the set of inputs of the DACross, and $\mathbf{y}$ corresponds to the true consecutive path of the labels, $\{y_1^{(p)}, ..., y_N^{(p)}\}$. Herein, $\boldsymbol{A}$ is a weight matrix of the CRF, named the transition matrix. $\boldsymbol{P}$ is the emission matrix, which is obtained from the linearly projected outputs at the last step. Consequently, the probability of the true path among all possible paths is as follows:

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp(\mathrm{S}(\mathbf{x}, \mathbf{y}))}{\sum\limits_{\mathbf{y}' \in \mathcal{Y}} \exp(\mathrm{S}(\mathbf{x}, \mathbf{y}'))} \tag{22}$$

Particularly, the goal of optimization is to minimise the CRF-Loss function.

$$\mathcal{L} = -\ln p(\mathbf{y}|\mathbf{x}) \tag{23}$$

V. EMPIRICAL EXPERIMENTS

In this section, we describe the details of the experimental settings, such as the datasets utilised in this study, baselines, and hyper-parameter settings.

A. Datasets

We used a total of four large datasets for our experiments. We collected these datasets from three actual large-scale travel surveys. Two of the datasets were used for training, validation, and testing, whereas the other two datasets were used only for testing. We randomly chose 20% of the trip chains to form the test set. A total of 10% of the remaining trip chains were randomly selected again to form the validation set. The remaining trip chains formed our training set. The datasets used only for testing were set as the control groups, which were completely used to evaluate the model performance in terms of generalizability. Statistical information and details about these datasets are presented in Table I.

The first two datasets, **Tokyo 23** (23 special wards of Tokyo) and **Downtown Yokohama**, are a part of the dataset of the Greater Tokyo Area, which is the most populated and largest economic metropolitan area in the world. We constructed these datasets using the 2018 Person Trip (PT) Survey[2] in Japan with approximately one million residents as respondents. **Twin Cities** contains trips between two large cities, Minneapolis, the most populated city in the state, and its neighbour to the east, Saint Paul, the state capital. We formulated the dataset using the Travel Behavior Inventory household (TBI)[3] survey from October 2018 through September 2019. **Chicago Metropolitan Area** is a major urban area in the midwestern United States; it is among the forty largest urban areas in the world. We prepared this dataset using the My Daily Travel Survey (MDTS)[4] considering the period of 2018-2019. Furthermore, to ensure fairness for all datasets, we collected POIs from OpenStreetMap as the raw features for nodes in the geographic adjacency graph. Notably, we could observe the vast average number of zone areas in the Twin Cities. This was owing to the large non-residential zones between the two cities and the broad lake area zones; otherwise, the dense urban zones have a median of 2.37 km^2 only.

B. Baselines

To fairly evaluate the effectiveness of DACross, we compare comprehensive baselines as follows.

Conventional machine learning methods: Conventional machine learning methods have mostly been implemented in previous research. These methods include

- **Random forest (RF):** This is the most common model used in trip-based prediction considering no chaining patterns [6], [15].
- **Bayesian neural network (BNN):** This is a variation of the neural network, which contains a prior probabilistic distribution on the weights. [16] utilised it to address the inherent uncertainty of trips.
- **Support vector machine (SVM):** This is the second most common comparative method for trip-level inference [16], [32]. We implemented the Radial Basis Function (RBF) kernel in this study.
- **XGBoost:** This was utilised for cycle's trip purpose inference problem [32].
- **Hidden Markov model (HMM):** This is a probabilistic model. [8], [21] applied it for chaining purpose inference.

Modified deep learning methods: Modified deep learning methods are adjusted according to the specified tasks. Two universal changes were made during their modifications: 1) we retained the input feature embedding instead of their original word embeddings, as our tasks are not meant for NLP, and 2) we replaced the output layers of the original models with the tree aggregation module in which we maintained our modelling concept, the dual-flow architecture.

- **Multilayer perceptron (MLP):** This is the basic structural element of 2-layer fully connected neural networks, which have the same meaning as the Artificial neural network (ANN) in the previous study [12].
- **Bi-LSTM+CRF [49] (dual-flow):** This was originally proposed for the sequential tagging of NLP. As our task

[2]https://www.mlit.go.jp/toshi/tosiko/toshi_tosiko_tk_000031.html
[3]https://msptravelstudy.org/mspweb/pages/home?locale=en
[4]https://www.cmap.illinois.gov/data/transportation/travel-survey

was similar to sequential tagging, we implemented it as the baseline, which did not involve feature crossing.

- **Bi-LSTM+CRF (trip chain-level):** This is a modified variant of the Bi-LSTM+CRF (dual-flow). All features are concatenated together to a single branch Bi-LSTM.
- **ESIM [46]:** This framework was proposed for the Natural Language Inference (NLI) task. As we borrowed ideas and insights from ESIM, we decided to use it for our task. It represents the architecture that implements dual-flow processing and co-attention enhancement. We trained it using the CRF loss.

C. Metrics

We evaluated all methods by means of the F1 scores and Micro- and Macro-views to investigate the unitary and category-specific performance. The Micro-F1 score is the harmonic mean of the values of precision and recall [51]. Compared to the general accuracy metric, the Micro-F1 score takes accounts into the balance of both precision and recall and into the imbalance class distribution, which is realistic to the trip purpose distribution. In the same manner, the Micro-F1 score was evaluated in a unitary way regardless of diversities among classes. We could evaluate the robust overall performance by the Micro-F1 score. By using the universal definitions of precision and recall, the Micro-F1 score can be defined as follows:

$$\mathrm{F}_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{24}$$

Assuming a total of C classes, acquiring the average of the sum of each class's Micro-F1 score yields the Macro-F1 score as:

$$\mathrm{Macro} - \mathrm{F}_1 = \frac{1}{C} \sum_{c \in C} \mathrm{F}_1^c \tag{25}$$

The Macro-F1 score (averaged Micro-F1 scores [52]) was used to evaluate the balance between classes. The Macro-F1 score is an important indicator reflecting the fine-grain performance, which evaluates whether methods can optimally handle all types of trip purposes.

VI. Results & Discussion

In this study, we propose a framework for trip purpose inference, which does not rely on traveller information but is robust, practicable, and accurate. In the following sub-sections, the experimental results of the proposed model are presented to verify if it achieves the pre-defined expectations.

A. Performance Analysis

The overall performance of the proposed DACross model was evaluated against the baseline models, and the results are summarised in Table II. We firstly trained and then tested the model performance on the Tokyo 23 and the Twin Cities datasets, and subsequently, tested the model again on the Downtown Yokohama and the Chicago Metropolitan Area datasets. The average values of the performance metrics after ten runs are reported. As the RF is the most implemented model in literature, we set it as the standard, and calculated the ratios of improvement relative to the RF. From the results, the following features were observed:

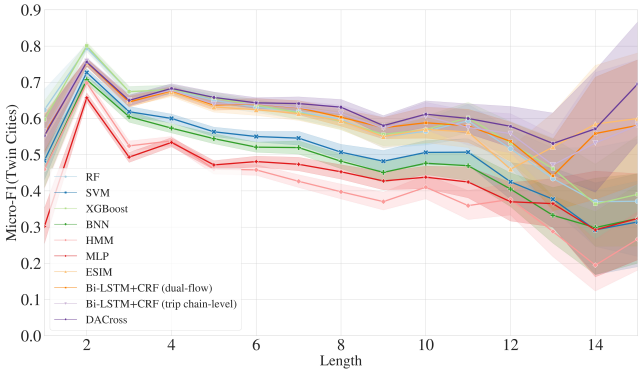


Fig. 3. Overall accuracy changes with the trip chain length. The number of samples decreases as the length increases. To ensure the validity of the plot, trip chain lengths with an inadequately small number of samples were removed. At least 20 samples were used for each trip chain length.

- The proposed DACross model outperformed all the models, except in the Macro-F1 score of the Twin Cities dataset. The relatively weak result may have resulted from the overfitting problems of the RF, which constantly underperformed on two generalization tests. We observed the highest accuracy losses of RF on the generalization test of the Twin Cites dataset compared to the Chicago Metropolitan Area dataset. Setting the results of RF as standard, the highest relative improvements of $+20.81\%$ and $+36.51\%$ were observed for the Chicago Metropolitan Area dataset. Thus, the superiority of the proposed DACross model was proved. In particular, the DACross model performed optimally on trip chains in unknown cities.
- The performance of SVM, HMM, and BNN is apparently worse than that of the other models. We consider that they are less suitable in our predefined scenario, which requires that the mining of trip chain is fully based on trips and zones. In previous studies, [16], [21], [32] implemented these three methods with elaborate feature selection or also with abundant respondent attributes. Moreover, to retain uncertainties during learning, BNN can be captious to its inputs. For instance, [16] collected more precise POIs from Google Maps and social media as the features of zones. However, the HMM shows more acceptable accuracy losses of generalization tests than those of BNN and SVM. Trip chaining patterns are helpful to prevent generalization losses.
- The performance of the BiLTSM+CRF (dual-flow) and ESIM models are similar and suboptimal compared to the top-tier models. The dual-flow variant of the BiLSTM+CRF weakly outperforms the conventional trip chain-level variant. As we modified them to adapt to our modelling concepts, we consider that the equally important dual-flow architecture is beneficial. Moreover,

Method	JP				US			
	Micro-F1		Macro-F1		Micro-F1		Macro-F1	
	Tokyo 23	Downtown Yokohama	Tokyo 23	Downtown Yokohama	Twin Cities	Chicago Metropolitan Area	Twin Cities	Chicago Metropolitan Area
RF	76.15	71.32	52.86	46.82	65.42	54.12	**54.05**	32.21
BNN	72.15	67.96	43.03	42.01	54.82	54.57	34.13	27.93
SVM	73.31	66.33	47.70	42.27	57.18	51.52	39.96	32.21
XGBoost	77.73	71.87	56.48	49.12	64.46	55.18	53.95	34.52
HMM	73.88	69.47	47.09	43.33	52.52	56.73	34.89	35.32
MLP	69.28	63.72	40.01	35.23	49.62	45.85	34.64	24.28
Bi-LSTM+CRF (dual-flow)	79.44	75.96	56.02	52.82	64.69	65.16	48.40	43.31
Bi-LSTM+CRF (trip chain-level)	78.62	75.90	54.89	52.80	65.03	64.77	48.51	42.47
ESIM	76.68	75.66	55.58	52.65	63.92	64.98	48.40	43.46
DACross	**80.09**	**77.55**	**58.86**	**55.79**	**65.84**	**65.38**	50.27	**43.97**
Relative %Improvement (RF)	+5.17	+8.74	+11.69	+19.16	+0.64	+20.81	-6.99	+36.51
Relative %Improvement (HMM)	+8.41	+11.63	+24.99	+28.76	+25.36	+15.25	+44.08	+24.49

TABLE II
PERFORMANCE EVALUATION.

deep learning-based models usually perform better than conventional machine learning models on generalisation tests, except the naive MLP. The MLP performed the worst model, suggesting that the absence of an appropriate model structure does not lead to performance gains.

B. Length-related & Purpose-specific Accuracy

We conducted a detailed analysis on the basis of the test results of Twin Cities, as the Twin Cities holds longer trip chains than the others. As trip chains could increase in complexity with increasing length, we evaluated the accuracy related to the chain length. Fig 3 shows that the accuracy changes with length. In the shorter length interval $[1, 5)$, all models perform similarly, as within this interval the trip chains are not severely complicated. When the length range increases to $[5, 9)$, MLP, SVM, BNN, and HMM start to fall behind the other models. When the length reaches 12 or more, our modified deep learning model starts to outperform conventional machine learning models. Simultaneously, the proposed DACross performs better in complicated trip chains. This phenomenon is reasonable as the longer the trip chain, the richer will be the interactions that the DACross can capture. The proposed DACross is effective for the long-chain inference, which was difficult to obtain in previous studies [21].

In addition to the aforementioned observations, we visualised purpose-specific accuracy matrices, as shown in Fig 4. The results are summarised as follows:

- The trip purposes, Work and Home, constantly show high accuracies, as they have strong and regular behavioural patterns within trip chains. Moreover, they are more distinguishable than the other purposes.
- Observably, certain sets of trip purposes can easily confuse the model prediction. This may result from the close proximity of semantics in the feature space. For instance, the sets of [Business, Work] and [Recreation, Shopping(, Meal)]. This also demonstrates that our proposed model can learn natural representations of purposes.
- The Education can often be confused with Work, as they almost have the same behavioural pattern. However, when compared to our previous results [23], we achieved a huge improvement (+21%). As we imported the intra- and inter- chain dependencies, the model could identify these two dependencies by referring to subsequent trips.

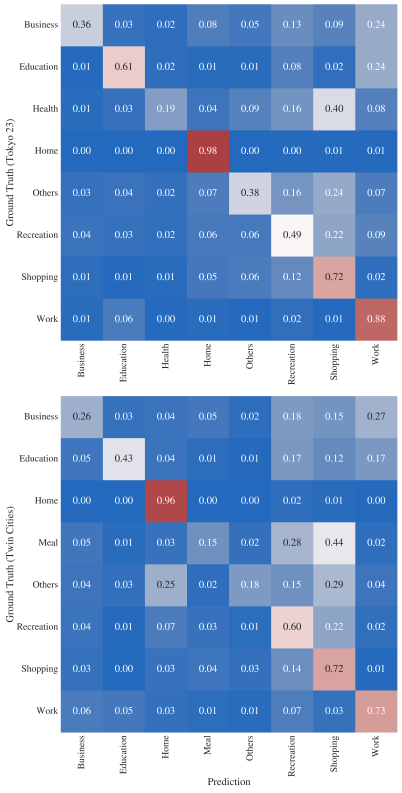


Fig. 4. Sub-class accuracy matrices of Tokyo 23 and Twin Cities, normalised by rows.

- The problem of recognizing Health and Others persists. These two are less distinguishable than the others and are difficult to classify optimally using the current model design and inputs.

C. Ablation Studies

To verify the functionalities of modules, we conducted ablation studies to analyse the effects of the proposed DACross framework as follows:

- DACross-*c*: All encoders were dropped, and only the co-attentive feature crossing module was left to evaluate whether capturing intra-sequence dependencies would be beneficial. The evident decrease is affected by the loss of capturing intra-sequence correlations, which appear to hold influence.
- DACross-*e*: The second encoder Encoder^2 and co-attentive feature crossing module were dropped in which

only the first encoder was applied for capturing the intra-sequence dependencies. The performance was seriously impaired by the loss of feature crossing and enhancement. Moreover, this sub-network architecture was the same as the Transformer Encoder+CRF in our scenario. We verified that only intra-chain dependencies were not sufficiently strong to support better performance.

- DACross-e^2: The second encoder $\mathrm{Encoder}^2$ was completely dropped to evaluate whether compositing intra-sequence dependencies again after feature crossing would be beneficial. We noticed that the second encoder demonstrated significant impact, which proved the existence of a dramatic change in features after feature crossing.
- DACross-e^1: The first encoder $\mathrm{Encoder}^1$ was dropped to evaluate whether enhancing intra-sequence dependencies before feature crossing would be beneficial. The results show that the first encoder could slightly decrease the model accuracies but not as significantly as the second encoder $\mathrm{Encoder}^2$.
- DACross-emb: The input feature embedding module was dropped. We observed the most serious accuracy loss in this case. Moreover, the processing of raw inputs was much impactful.
- DACross-t: The Tree-LSTM was dropped and replaced by concatenating all three elements together. Slight accuracy decrease was observed of which we consider the prior modules were sufficiently effective.

Variant	Micro-F1		Macro-F1	
	Tokyo 23	Twin Cities	Tokyo 23	Twin Cities
DACross	80.09	65.84	58.86	50.27
-c	79.07	64.08	57.00	48.51
-e	79.49	62.79	56.93	47.77
-e2	79.55	63.33	57.46	47.98
-e1	79.98	64.19	58.68	48.63
-emb	**71.41**	**55.08**	**43.29**	**38.21**
-t	80.00	64.46	58.48	48.37

TABLE III
RESULTS OF THE ABLATION STUDIES.

D. Hyper-Parameter Sensitivity Analysis

Considering the relatively wide range of selections of hyper-parameters, we focused on analyzing the combinatorial hyper-parameter trends among primary modules. Figure 5 presents experimental results on the Twin Cities dataset. We used subscripts c, 1, and 2 to denote the hyper-parameters of the feature crossing module, the first, and the second encoders, respectively. K_c is the number of heads of the co-attentive feature crossing module. L_1 and L_2 denote the number of layers of the first and second encoders, respectively.

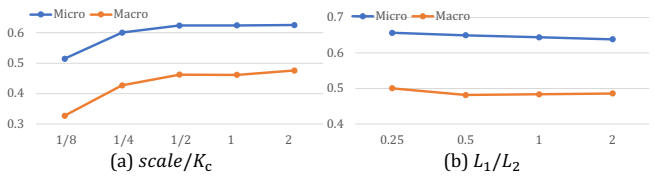


Fig. 5. Parameter sensitivity analysis on the Twin Cites dataset.

1) Ratio of $scale/K_c$: We input the selections $[1, 2, 4]$ of K_c and $[0.5, 1, 2]$ of $scale$, and average results to obtain Figure 5a. Observably, the ratios between 0.5 to 2 differ slightly. Recall that the dimensions of the inner single layer FFN was set to $d_c = \frac{d_m \times scale}{K_c}$. When d_m is over four times larger than d_c, the co-attentive feature crossing module cannot capture sufficient interdependencies.

2) Ratio of L_1/L_2: As two encoders are responsible for processing different information, their diversities are worthy of an investigation. We set L_1 to $[1, 2, 4, 8]$ and set the number of L_2 lower, equal, or higher than L_1. The results in Figure 5b, suggest that empirically $L_1 \leq L_2$ performs better. The first encoder contributes to capturing intra-sequence correlations without crossing. Thus, it might process less information than the second encoder, relatively, in which $L_1 \leq L_2$ is beneficial.

E. Zone Scale Sensitivity Analysis

As the different ways of splitting zones can lead to diverse results, we investigated the influences by evaluating the proposed DACross on four scales of the administrative boundaries in Japan (Fig. 6). The finest scale, cho-me, is the default setting in our experiments.

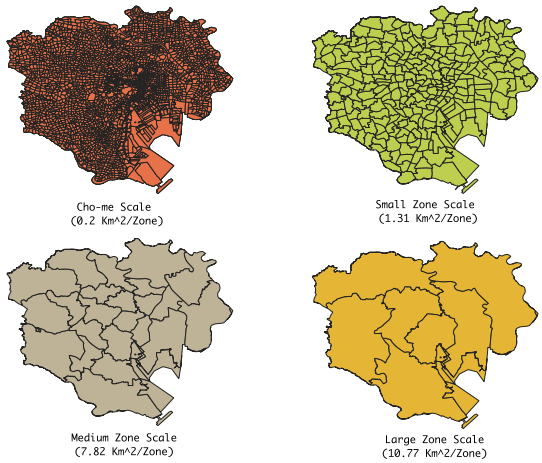


Fig. 6. Four scales of splitting zones in Tokyo 23.

As shown in the Table IV, aggregating zones into the small scale does not affect the loss of performance greatly, but the number of zones declines 12 times than that of the current value. However, the medium and large scales present heavy loss of both of Micro- and Macro-F1 scores. Although the larger zone can lead to more inclusion of POIs, excessive POIs will conversely injure the model's discernibility as more noises are imported. From a practical point of view, we suggest that a zone of 1-2 km^2 is sufficient for the acceptable inference accuracy.

Scale	#zones	avg of #POIs	Micro-F1	Macro-F1
Cho-me	3,192	15.47	80.09	58.86
Small	266	185.94	80.26	58.70
Medium	24	2060.83	78.75	55.31
Large	8	6182.50	74.74	47.10

TABLE IV
RESULTS OF THE ZONE SCALE CHANGES.

F. Feature Sensitivity Analysis

To further explore the factors affecting model performance, in this section, a more complete dataset of POIs is introduced and previously absent traveller information is added back for feature combination analysis in the Tokyo 23 dataset. We conduct extensive experiments to reveal the impacts, which cannot be avoided in the real-world applications.

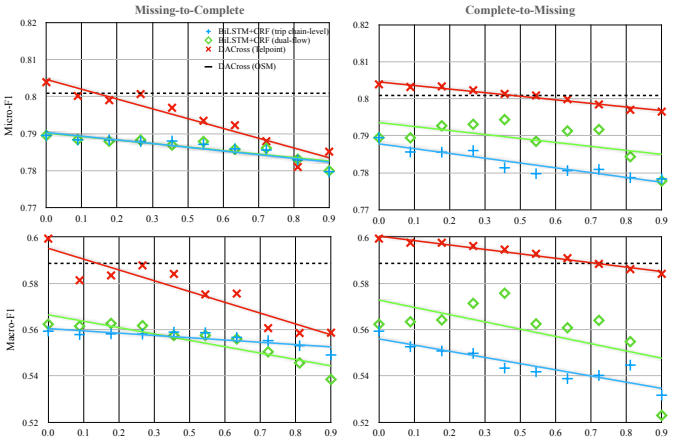


Fig. 7. Missing data test using Telepoint POI dataset in Tokyo 23.

1) Impact of geographic contextual information: We investigated whether the absence of POIs affected the inference performance of the well-trained model, considering that the missing data problem is common in most real-world applications. In this experiment, we simulated the inference scenarios with missing data. A detailed POI dataset, Telepoint Pack DB provided by ZENRIN CO., was newly imported. This dataset collected POIs from approximately 320,000 telephone directories. Each record in the dataset contains detailed information, such as phone numbers, coordinates, addresses, categories, and names. We could roughly recognise that the Telepoint dataset contained complete POI information in the Tokyo 23 dataset. There are 433 second-class categories of POIs. Following the same criterion used to process OpenStreetMap, we acquired distributions of the numbers of various categorical POIs as the node features of our predefined geographic adjacent graph. Subsequently, we repeated random elimination of POIs for simulating the absence of POIs under different conditions from 10% to 90% missing.

We performed two types of comparative experiments on the DACross and variants of BiLSTM+CRF: 1) The missing-to-complete experiment was used to assess the damage to models that were trained on the missing data conditions, and 2) the complete-to-missing experiment was used to evaluate the models robustness to the missing data conditions after a round of training on the full dataset. The Test results are illustrated in Fig 7. When no missing data exists, the Micro- and Macro-F1 scores are 80.39% ($+0.37\%$) and 59.93% ($+1.82\%$), respectively. From the results of the missing-to-complete experiment, we observed that the dual-flow architecture was more susceptible to missing data, probably because more noise was imported from the separated branch of chained zones. Furthermore, we found that the DACross suffered more decline than the model without feature crossing, as the feature crossing further imported noises into the other branch. From the results of the complete-to-missing experiment, we observed much flatter decline of all approaches, but the dual-flow architecture presented a relatively better tolerance to missing data conditions. When the missing ratio approached 70%, we acquired similar results, 79.85% and 58.82%, with OSM. Moreover, we observed that the missing data influenced the values of Macro-F1 scores more than the Micro-F1 scores. The loss of geographic contextual information can actually damage the fine-grain inference performance of the model.

2) Impact of feature combination: We categorise raw features into 4 types, which are activity-related, OD-related, traveller-related, and time-related features. Further details are listed in Appendix. A. We explored potential influence under the combinations between 4-type of features, and the results are presented in this section. The RF and HMM are the most commonly implemented models for the task of inferring the trip purpose. Thus, we selected them as typical standards of trip-level modelling and trip chain-level modelling. The results of the DACross model are listed in Table I, and the relative improvements are shown in Fig 8. We summarise the findings as follows:

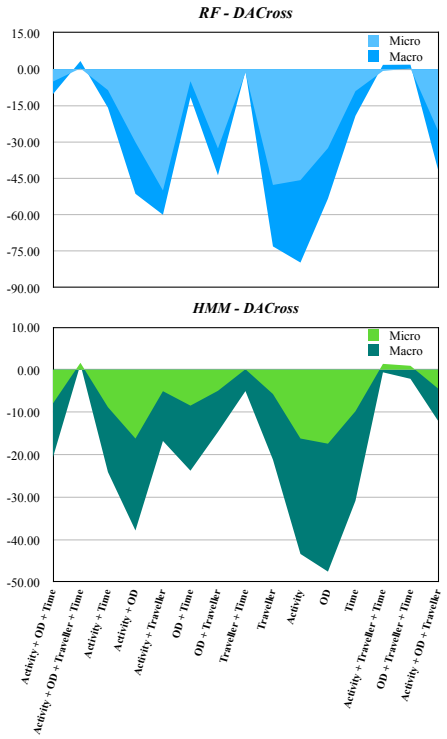


Fig. 8. Percentage relative improvement of F1 scores refer to DACross.

- The combination of $\mathrm{Activity}+\mathrm{OD}+\mathrm{Time}$ is the same as the initial settings, and the values are compared in Table I. After adding back the traveller-related information, no obvious improvements were observed. On one hand, RF and HMM can slightly outperform the proposed DACross model. Traveller-related information is intuitively meaningful for enhancing fine-grain inference. On the other hand, the result also proves that the DACross model

is suitable for our pre-defined expectation in scenarios where traveller information is missing.

- Time-related information is effective for trip purpose inference. The combinations with time-related information consistently perform better than those without time-related information. This finding proves that the assignment of trip purposes is strongly time-related [37]. Furthermore, based on the evidence, the combinations including $\mathrm{Traveller} + \mathrm{Time}$ can lead to the optimal results. In this manner, we assume that temporal patterns exist as the background mechanism affecting trip generation.

G. Qualitative Analysis

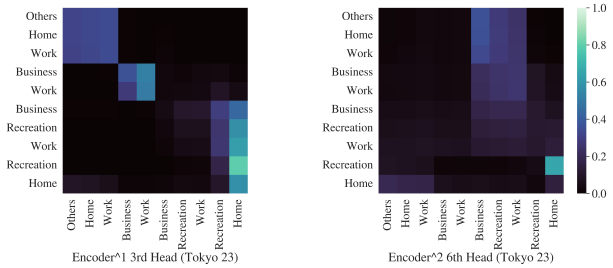


Fig. 9. Visualization of attention weights acquired from the picked trip chain.

We focused on qualitative visualization analysis, and the results are presented in this section. First, we explored whether two encoders can emphasise distinguishable trip purposes. To achieve this, we observed a trip chain that has 11 lengths from the test set in the Tokyo 23 dataset. We visualised two attention weights from $\mathrm{Encoder}^1$ and $\mathrm{Encoder}^2$, as shown in Fig 9. At the third head of $\mathrm{Encoder}^1$, correlations between nearby trips are underlined. Moreover, the trips in the latter half of trip chain show tight relations to the last trip Home. Such phenomenon is also observed from $\mathrm{Encoder}^2$. We argue that the previous hypothesis [36] that the composition of a trip chain is primarily decided by several compulsory trips can be validated in this manner.

By using the CRF loss set-up, we explored the trip purpose transition matrices (Fig 10) for comparing behavioural differences in two cities. In both the Tokyo 23 and Twin Cities datasets, there are great probabilities that $\mathrm{Business}$ and Work activities are triggered after each other. Residents in the Twin Cities are more likely to prefer ongoing recreational activities $\mathrm{Recreation}, \mathrm{Shopping}$ than the residents in Tokyo 23.

H. Discussion

We explored the diversities across the datasets, and the possible reasons that led to differences are listed as follows:

- **Sampling:** The TBI and MDTS were collected from travel dairies through a hybrid way. Most responses were recorded by PR survey that we mentioned in the introduction section. Only partial responses were not conducted online. Thus, the data sampling used here is more realistic compared to that of the PT survey. The PT survey sent survey letters to residents and asked them to record one typical day. From the statistical information in Table I, the lengths of trip chains are obviously different. PR surveys can contribute to realistic sampling. However, they make inference far more difficult. This could be one of the possible reasons that reduced the performance of the proposed model on the Twin Cites and the Chicago Metropolitan Area datasets.
- **Uniformity:** The PT survey only asks respondents an assumed typical day, but the TBI and MDTS processed travel dairies are mostly collected from mobile GPS records within multiple days. There could be trip chains belonging to the same respondents. Our modelling concept does not include the usage of uniformity. In this study, we designed DACross initially from [23], for which we did not consider abundant datasets. Although such disadvantages exist, the DACross can still be an outstanding approach among the baselines.
- **Weekday&Weekend:** The TBI and MDTS also collected trip chains during weekends or holidays. Generally, travel behaviour on weekends is different from that on weekdays [53]. Moreover, we do not exactly discriminate holidays [54] from the TBI survey, and it might account for the differences.
- **POIs:** Observably, the distribution of POIs in the Twin Cities and the Chicago Metropolitan Area are much sparser (nearly 10 times from Fig 11) than in Tokyo 23 and Yokohama. We contemplate that there could be more noise and less useful information about zones. Under extreme situations, given a trip chain, there are no POIs at all in travelled zones. This is the potential reason for the unsuccessful performance of the co-attentive feature crossing module. More studies must be conducted to explore this concept.

VII. Conclusion

We studied the problem of chained trip purpose inference, which is an important task in urban planning and for the analysis of travel behaviour, using information about trips and zones only. In this study, we proposed a new approach to form the trip chain, which consists of an attentive dual-flow or dual-branch architecture. To leverage trip chain interdependencies sufficiently, we proposed a co-attentive feature crossing module with a second intra-trip chain encoder. The performance of the proposed model was evaluated via ablation

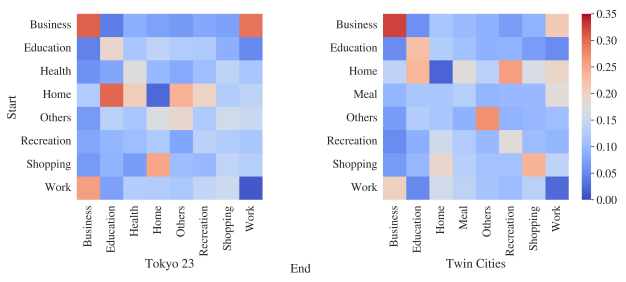


Fig. 10. Visualization of attention weights acquired from the picked trip chain.

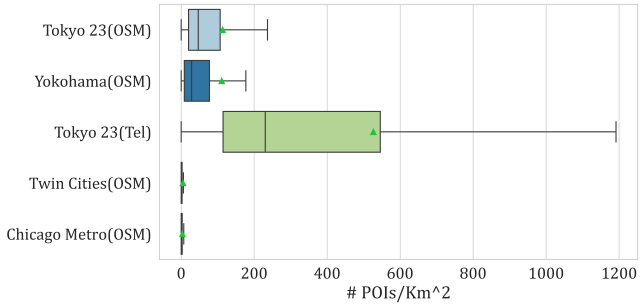


Fig. 11. Densities of POIs.

studies. Extensive experiments were conducted on four large-scale real datasets, and the results were compared with various benchmark models. The results confirmed the performance, generalization ability, and practicability of the proposed framework. The proposed framework achieved all the research objectives of our study. However, there are some questions that are yet to be answered. The proposed framework did not perform as expected on the Twin Cities and the Chicago metropolitan area datasets. The diversity of travel behaviour between different countries is worth exploring. We believe the limitation of the proposed framework could inspire an in-depth investigation into understanding human behaviour in future research.

In summary, we considered two directions of future research. An intuitive idea is to explore methods to save the learned interactive features, specifically the trip-to-zone direction, and pass them back into the geographic adjacent graph. In this way, as the learning process iterates, we can extract the geographic contexts from a traveller's perspective, which is the same as clustering the travellers' statistical information in zones [18]. This approach can benefit not only the model performance but also overcome the serious issue of the drop in performance caused by the missing POIs. In the real-world applications, we cannot know the complete geographic contextual information overtime. As we observed strong temporal correlations among trip purposes, continuous exploration of the attitudes of travellers who value time is meaningful for improving travel behaviour comprehension. Eventually, we plan to conduct a large-scale empirical experiment based on the proposed method in the future.

ACKNOWLEDGMENTS

This study was supported by Joint Research Program No. 1061 at CSIS, UTokyo (Telepoint Pack DB August 2020 by ZENRIN CO.) and JST SPRING, Grant Number JPMJSP2108.

APPENDIX A. DATASET & PREPROCESSING

The four datasets are roughly categorised in terms of the country of origin. The Tokyo 23 and Downtown Yokohama datasets are from Japan and the other datasets are from the US. We present the detailed information in Table V. Notably, the features of the traveller were not included by default but were set for the feature combination test. For all numerical features, we applied standardization, such as pre-processing. All categorical inputs were converted by means of one-hot encoding. In particular, time-related features, such as the departure and arrival time, were transformed to seconds. For deep learning-based models, we further implemented linear projection and add operation on time-related features. Given a trip chain of length N, there should be N trips and $N+1$ travelled zones. Subsequently, setting departure and arrival time as corresponding trigger points for trips and zones respectively, we acquired N departure times and $N+1$ arriving times, which gives rise to two time-related vectors $\mathbf{x}_{td} \in \mathbb{R}^N$ and $\mathbf{x}_{ta} \in \mathbb{R}^{N+1}$. The first arrival time was set to 0:00 AM or 3:00 AM for datasets in JP or US, respectively. We firstly transformed these into a cyclical format and thereafter projected the new products to the same dimension with $\mathbf{h}^t$ and $\mathbf{h}^v$ in Equations 2-3.

$$\hat{\mathbf{h}^t} = \mathbf{h}^t + [\sin(\frac{2\pi \mathbf{x}_{td}}{3600 \times 24}); \cos(\frac{2\pi \mathbf{x}_{td}}{3600 \times 24})]\boldsymbol{W}_{td} \quad (26)$$

$$\hat{\mathbf{h}^v} = \mathbf{h}^v + [\sin(\frac{2\pi \mathbf{x}_{ta}}{3600 \times 24}); \cos(\frac{2\pi \mathbf{x}_{ta}}{3600 \times 24})]\boldsymbol{W}_{ta} \quad (27)$$

$\boldsymbol{W}_{td}, \boldsymbol{W}_{ta} \in \mathbb{R}^{2 \times d_m}$ are the weights for linear projection. Thus, we replaced the outputs of Equations 2-3 with $\hat{\mathbf{h}^t}$ and $\hat{\mathbf{h}^v}$.

Category		Feature	Format	JP	US
OD	$\mathbf{x}^v$	#Public	Integral	√	√
		#Health		√	√
		#Leisure		√	√
		#Catering		√	√
		#Accommodation		√	√
		#Shopping		√	√
		#Tourism		√	√
		#Transport		√	√
Activity	$\mathbf{x}^t$	Travel time (minute)	Float	√	√
		Stay duration (minute)			√
		Speed (mile/minute)			√
		Distance (mile)			√
		Transportation mode	Categorical (4)	√	√
Time		Departure time (second)	Intergral	√	√
		Arriving time (second)		√	√
Traveller		#Household	Intergral	√	√
		Age	Categorical (10)	√	√
		Annual income	Categorical (5)	√	√
		Gender	Binary	√	√
		IF full-time worker		√	√
		IF student		√	√
		IF unemployed		√	√
Purpose	y	Business	Binary	√	√
		Education		√	√
		Health		√	
		Home		√	√
		Meal			√
		Others		√	√
		Recreation		√	√
		Shopping		√	√
		Work		√	√

TABLE V
DATASET DESCRIPTION

APPENDIX B. HYPER-PARAMETER & TRAINING PROGRESS

We implemented DACross and deep learning models using Tensorflow 2.4.0 and Python 3.7. The experiments were conducted on an Ubuntu 20.04 operating system with a single NVIDIA RTX 3090 24-GB graphics processing unit. For the Tokyo 23 and Twin Cities datasets, we conducted hyper-parameter search 80 times using Bayesian optimization each time. Setting all hyper-parameters as searchable renders the search space considerably wide. Hence, we only performed search on critical hyper-parameters. The other hyper-parameters were configured empirically. Details of hyper-parameters are listed in Table VI. An early stopping strategy

Hyper-Parameter	Tokyo 23	Twin Cites	Searchable
#Batch	64	256	
Learning rate	5×10^{-4}	1×10^{-4}	
Dropout rate	0.1	0.1	
Bias	Yes	No	$\surd$
d_m	64	64	
#GAT heads K	4	4	
#Encoder^1 heads K_1	4	8	
#Encoder^2 heads K_2	8	8	
#Encoder^1 layers L_1	1	1	$\surd$
#Encoder^2 layers L_2	1	3	$\surd$
#Cross scale	0.5	0.5	$\surd$
#Cross heads K_c	8	4	$\surd$
#Cross pooling	mean	mean	
d_{agg}	512	512	$\surd$

TABLE VI
HYPER-PARAMETER SETTINGS

was used to stop training progress when the loss from the validation set was larger than the loss from the training set. Moreover, deep learning-based baselines shared the same hyper-parameters from search results. The hyper-parameters of conventional machine learning baselines were selected by the optimum results from complete grid searches, except the BNN. We utilised the same settings of [16] for the BNN.

REFERENCES

[1] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Physics Reports*, vol. 734, pp. 1–74, Mar. 2018.

[2] S. Handy, "Methodologies for exploring the link between urban form and travel behavior," *Transportation Research Part D: Transport and Environment*, vol. 1, no. 2, pp. 151–165, Dec. 1996.

[3] S. Hanson and J. Huff, "Classification issues in the analysis of complex travel behavior," *Transportation*, vol. 13, no. 3, pp. 271–293, Sep. 1986.

[4] X. Lu and E. I. Pas, "Socio-demographics, activity participation and travel behavior," *Transportation Research Part A: Policy and Practice*, vol. 33, no. 1, pp. 1–18, Jan. 1999.

[5] D. E. Boyce and H. C. Williams, *Forecasting Urban Travel: Past, Present and Future*. Edward Elgar Publishing, 2015.

[6] E. F. d. S. Soares, K. Revoredo, F. Baião, C. A. d. M. S. Quintella, and C. A. V. Campos, "A Combined Solution for Real-Time Travel Mode Detection and Trip Purpose Prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4655–4664, Dec. 2019.

[7] H. Hamedmoghadam, H. L. Vu, M. Jalili, M. Saberi, L. Stone, and S. Hoogendoorn, "Automated extraction of origin-destination demand for public transportation from smartcard data with pattern recognition," *Transportation Research Part C: Emerging Technologies*, vol. 129, p. 103210, Aug. 2021.

[8] C. M. Krause and L. Zhang, "Short-term travel behavior prediction with GPS, land use, and point of interest data," *Transportation Research Part B: Methodological*, vol. 123, pp. 349–361, May 2019.

[9] L. Amichi, A. C. Viana, M. Crovella, and A. A. Loureiro, "From movement purpose to perceptive spatial mobility prediction," in *Proceedings of the 29th International Conference on Advances in Geographic Information Systems*, ser. SIGSPATIAL '21. New York, NY, USA: Association for Computing Machinery, Nov. 2021, pp. 500–511.

[10] T. T. Golob and H. Meurs, "Biases in response over time in a seven-day travel diary," *Transportation*, vol. 13, no. 2, pp. 163–181, Jun. 1986.

[11] W. Bohte and K. Maat, "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 3, pp. 285–297, Jun. 2009.

[12] G. Xiao, Z. Juan, and C. Zhang, "Detecting trip purposes from smartphone-based travel surveys with artificial neural networks and particle swarm optimization," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 447–463, Oct. 2016.

[13] E. Murakami and D. P. Wagner, "Can using global positioning system (GPS) improve trip reporting?" *Transportation Research Part C: Emerging Technologies*, vol. 7, no. 2, pp. 149–165, Apr. 1999.

[14] J. Wolf, R. Guensler, and W. Bachman, "Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data," *Transportation Research Record*, vol. 1768, no. 1, pp. 125–134, Jan. 2001.

[15] A. Ermagun, Y. Fan, J. Wolfson, G. Adomavicius, and K. Das, "Real-time trip purpose prediction using online location-based search and discovery services," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 96–112, Apr. 2017.

[16] Y. Cui, C. Meng, Q. He, and J. Gao, "Forecasting current and next trip purpose with social media data and Google Places," *Transportation Research Part C: Emerging Technologies*, vol. 97, pp. 159–174, Dec. 2018.

[17] C. Chen, S. Jiao, S. Zhang, W. Liu, L. Feng, and Y. Wang, "TripImputor: Real-Time Imputing Taxi Trip Purpose Leveraging Multi-Sourced Urban Data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 10, pp. 3292–3304, Oct. 2018.

[18] Q. Gao, J. Molloy, and K. W. Axhausen, "Trip purpose imputation using GPS trajectories with machine learning," *ISPRS International Journal of Geo-Information*, Aug. 2021.

[19] H. Faroqi and M. Mesbah, "Inferring trip purpose by clustering sequences of smart card records," *Transportation Research Part C: Emerging Technologies*, vol. 127, p. 103131, Jun. 2021.

[20] K.-F. Chu, A. Y. S. Lam, and V. O. K. Li, "Deep Multi-Scale Convolutional LSTM Network for Travel Demand and Origin-Destination Predictions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3219–3232, 2020.

[21] S. Jeong, Y. Kang, J. Lee, and K. Sohn, "Variational embedding of a hidden Markov model to generate human activity sequences," *Transportation Research Part C: Emerging Technologies*, vol. 131, p. 103347, Oct. 2021.

[22] B. Mo, Z. Zhao, H. N. Koutsopoulos, and J. Zhao, "Individual Mobility Prediction in Mass Transit Systems Using Smart Card Data: An Interpretable Activity-Based Hidden Markov Approach," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2021.

[23] S. Lyu and T. Kusakabe, "Graph-aware Chained Trip Purpose Inference," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, Sep. 2021, pp. 3691–3697.

[24] C. Liao, C. Chen, S. Guo, Z. Wang, Y. Liu, K. Xu, and D. Zhang, "Wheels Know Why You Travel: Predicting Trip Purpose via a Dual-Attention Graph Embedding Network," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 22:1–22:22, Mar. 2022.

[25] C. Chen, J. Ma, Y. Susilo, Y. Liu, and M. Wang, "The promises of big data and small data for travel behavior (aka human mobility) analysis," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 285–299, Jul. 2016.

[26] L. Shen and P. R. Stopher, "A process for trip purpose imputation from Global Positioning System data," *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 261–267, Nov. 2013.

[27] L. Alexander, S. Jiang, M. Murga, and M. C. González, "Origin–destination trips by purpose and time of day inferred from mobile phone data," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 240–250, Sep. 2015.

[28] A. Alsger, A. Tavassoli, M. Mesbah, L. Ferreira, and M. Hickman, "Public transport trip purpose inference using smart card fare data," *Transportation Research Part C: Emerging Technologies*, vol. 87, pp. 123–137, Feb. 2018.

[29] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data fusion approach," *Transportation Research Part C: Emerging Technologies*, vol. 46, pp. 179–191, Sep. 2014.

[30] P. Wang, G. Liu, Y. Fu, Y. Zhou, and J. Li, "Spotting Trip Purposes from Taxi Trajectories: A General Probabilistic Model," *ACM Transactions on Intelligent Systems and Technology*, vol. 9, no. 3, pp. 29:1–29:26, Dec. 2017.

[31] L. Montini, N. Rieser-Schüssler, A. Horni, and K. W. Axhausen, "Trip Purpose Identification from GPS Tracks," *Transportation Research Record*, vol. 2405, no. 1, pp. 16–23, Jan. 2014.

[32] S. Nair, K. Javkar, J. Wu, and V. Frias-Martinez, "Understanding Cycling Trip Purpose and Route Choice Using GPS Traces and Open Data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–26, Mar. 2019.

[33] G. Han and K. Sohn, "Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model," *Transportation Research Part B: Methodological*, vol. 83, pp. 121–135, Jan. 2016.

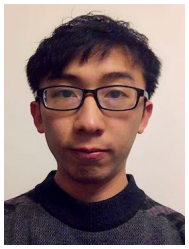
[34] C. Meng, Y. Cui, Q. He, L. Su, and J. Gao, "Travel purpose inference with GPS trajectories, POIs, and geo-tagged social media data," in *2017*

IEEE International Conference on Big Data (Big Data), Dec. 2017, pp. 1319–1324.
[35] ——, "Towards the Inference of Travel Purpose with Heterogeneous Urban Data," *IEEE Transactions on Big Data*, pp. 1–1, 2019.
[36] K. G. Goulias and R. Kitamura, "Recursive Model System for Trip Generationand Trip Chaining," *Transportation Research Record*, no. 1236, pp. 59–66, 1989.
[37] T. F. Golob, "A simultaneous model of household activity participation and trip chain generation," *Transportation Research Part B: Methodological*, vol. 34, no. 5, pp. 355–376, Jun. 2000.
[38] R. Kitamura, "Incorporating trip chaining into analysis of destination choice," *Transportation Research Part B: Methodological*, vol. 18, no. 1, pp. 67–81, Feb. 1984.
[39] Z. Liu, F. Miranda, W. Xiong, J. Yang, Q. Wang, and C. Silva, "Learning Geo-Contextual Embeddings for Commuting Flow Prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, Apr. 2020, pp. 808–816.
[40] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *International Conference on Learning Representations*, Feb. 2018.
[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in {*Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
[42] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical Question-Image Co-Attention for Visual Question Answering," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 289–297.
[43] X. Zhou, L. Li, D. Dong, Y. Liu, Y. Chen, W. X. Zhao, D. Yu, and H. Wu, "Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 1118–1127.
[44] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional Attention Flow for Machine Comprehension," in *International Conference on Learning Representations*, 2017.
[45] Y. Tay, A. T. Luu, and S. C. Hui, "Multi-Pointer Co-Attention Networks for Recommendation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD '18. New York, NY, USA: Association for Computing Machinery, Jul. 2018, pp. 2309–2318.
[46] Q. Chen, X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen, "Enhanced LSTM for Natural Language Inference," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 1657–1668.
[47] K. S. Tai, R. Socher, and C. D. Manning, "Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, Jul. 2015, pp. 1556–1566.
[48] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T.-Y. Liu, "On layer normalization in the transformer architecture," in *ICML 2020*, Jul. 2020.
[49] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF Models for Sequence Tagging," *arXiv:1508.01991 [cs]*, Aug. 2015.
[50] R. Panchendrarajan and A. Amaresan, "Bidirectional LSTM-CRF for Named Entity Recognition," in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Hong Kong: Association for Computational Linguistics, 2018.
[51] Y. Sasaki *et al.*, "The truth of the F-measure," *Teach tutor mater*, vol. 1, no. 5, pp. 1–5, 2007.
[52] J. Opitz and S. Burst, "Macro F1 and Macro F1," Feb. 2021.
[53] J. Zhao, J. Wang, and W. Deng, "Exploring bikesharing travel time and trip chain by gender and day of the week," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 251–264, Sep. 2015.
[54] L. Yang, Q. Shen, and Z. Li, "Comparing travel mode and trip chain choices between holidays and weekdays," *Transportation Research Part A: Policy and Practice*, vol. 91, pp. 273–285, Sep. 2016.

Suxing Lyu is currently a Ph.D. Candidate in the Center for Spatial Information Science (CSIS) and Department of Socio-Cultural Environmental Studies, Graduate School of Frontier Science, the University of Tokyo. His research focus on behaviour analysis, individual mobility, and spatiotemporal data mining.

Tianyang Han is currently a Ph.D. Candidate in Department of Civil Engineering, Graduate School of Engineering, the University of Tokyo. He is with the Transportation Engineering Laboratory (Oguchi Lab.) in Institute of Industrial Science (IIS). His research focus on traffic data analysis, traffic management and control systems and machine learning methods.

Peiran Li is now a Ph.D. candidate of the Center for Spatial Information Science, at The University of Tokyo. He has experience in the field of remote sensing, and currently his work focuses on urban computing and energy issues based on big data and AI tools.



Xingyu Luo is a Ph.D. candidate in the Department of System Biology at George Mason University. He holds a master's degree in Computational Biology and Bioinformatics from the George Mason University and a bachelors degree in Biotechnology in Central China University. His research interests including deep learning modeling, applied machine learning in Bioinformatics.

Takahiko Kusakabe is an Associate Professor in the Center for Spatial Information Science (CSIS), the University of Tokyo. He received the Dr.Eng. degree from Kobe University, Japan in 2010. From 2011 to 2016 he worked at the Tokyo Institute of Technology in Tokyo as an Assistant Professor. His research interest is transportation engineering with emerging technologies. Much of his work has been focusing on the application and implementation of Big data, IoT (Internet of Things), and sensors to traffic and transportation observation and surveys. And incorporating with transportation simulation, applications of these technologies in mobility data platforms, MaaS (Mobility as a Service), and urban mobility design have been explored.