Learning to Discover Causes of Traffic Congestion with Limited Labeled Data

Mudan Wang
Department of Electronic Engineering, Tsinghua University
Beijing, China

Huan Yan*
Department of Electronic Engineering, Tsinghua University
Beijing, China

Hongjie Sui
Department of Electronic Engineering, Tsinghua University
Beijing, China

Fan Zuo
AutoNavi, Alibaba Group
Beijing, China

Yue Liu
AutoNavi, Alibaba Group
Beijing, China

Yong Li
Department of Electronic Engineering, Tsinghua University
Beijing, China

ABSTRACT

Traffic congestion incurs long delay in travel time, which seriously affects our daily travel experiences. Exploring why traffic congestion occurs is significantly important to effectively address the problem of traffic congestion and improve user experience. Traditional approaches to mine the congestion causes depend on human efforts, which is time consuming and cost-intensive. Hence, we aim to discover the known and unknown causes of traffic congestion in a systematic way. However, to achieve it, there are three challenges: 1) traffic congestion is affected by several factors with complex spatio-temporal relations; 2) the amount of congestion data with known causes is small due to the limitation of human label; 3) more unknown congestion causes are unexplored since several factors contribute to traffic congestion. To address above challenges, we design a congestion cause discovery system consisting of two modules: 1) congestion feature extraction, which extracts the important features influencing congestion; and 2) congestion cause discovery, which utilize a deep semi-supervised learning based method to discover the causes of traffic congestion with limited labeled causes. Specifically, it first leverages a few labeled data as prior knowledge to pre-train the model. Then, the deep embedded clustering method is performed to produce the clusters under the supervision of the data reconstruction loss and Kullback-Leibler divergence loss. Extensive experiments show that the performance of our proposed method is superior to the baselines. Additionally, our system is deployed and used in the practical production environment at Amap.

CCS CONCEPTS

• **Information systems** → **Data mining**; *Clustering*.

KEYWORDS

Traffic congestion causes, novel category discovery, transfer clustering, spatio-temporal data mining

ACM Reference Format:
Mudan Wang, Huan Yan*, Hongjie Sui, Fan Zuo, Yue Liu, and Yong Li. 2022. Learning to Discover Causes of Traffic Congestion with Limited Labeled Data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539185

1 INTRODUCTION

Traffic congestion, characterized by slow driving speed rather than free-flow speed, has become a common phenomenon in urban transportation systems. According to statistics in [16], nearly 25% of American adults experienced traffic congestion every day. There are several negative effects caused by traffic congestion such as the long-time delay. For example, people in America spend 4.3 billion extra hours in travel time in 2020 due to traffic congestion [20]. Thus, traffic congestion is a serious problem, which has a great impact on our daily travel experiences.

To tackle the problem of traffic congestion, many map service providers launch their map applications that not only offer the real-time traffic condition information, but also have ability to predict the future traffic condition by traffic prediction models [11, 17, 25, 28]. However, in most cases, people not only want to know where and when the traffic is congested, but also would like to realize why traffic congestion occurs. For example, when a user is experiencing the traffic congestion, he/she is more willing to know the congestion causes. Thus, it is important to discover the causes of congestion. For map service providers, offering the causes of congestion is beneficial to improve the user experience of their map applications. For individual users, they can dynamically adjust their trip routes to alleviate the influence of the traffic congestion.

In order to study the causes of traffic congestion, we collect large-scale traffic data within one week in a city of China. The traffic data includes the traffic conditions, e.g., the average speed and congestion levels, of each road in the road network at each time slot. Meanwhile, we also collect other geo-spatial information like points of interests (POI) in the city, which is helpful for the congestion cause analysis. To comprehensively investigate the traffic congestion, we define a congestion event as the traffic conditions on a

[*] Corresponding author. Email: yanhuan@tsinghua.edu.cn.

KDD '22, August 14–18, 2022, Washington, DC, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539185

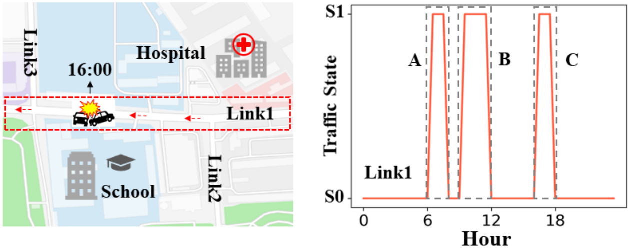


Figure 1: An example of traffic congestion caused by different causes. We use A, B and C to represent different causes, which are attending school, going to hospital and traffic accidents, respectively.

road segment at the maximum consecutive time slots are congested. From our data analysis, we find that traffic congestion events are strongly related to the nearby places and time information, thus their causes are diverse and complicated. For instance, the traffic congestion on road segment *link1* in Figure 1 occurred three times, but their causes were different. The first congestion was caused by going to school, which occurred from 7AM to 8AM. The second happened between 9AM and 10AM due to high traffic near the hospital. The last took place at 4PM because of a severe accident. Since there are a large number of congestion events happening every day in the city, it raises a question of how to automatically and accurately identify these causes.

Traditionally, it relies on considerable experience and expertise knowledge to manually label the congestion causes. It is too time-consuming and costly to find all the causes of traffic congestion events, which cannot be used in the real-time map applications. Thus, how to automatically and accurately identify the causes of traffic congestion is our main focus. To address it, we first manually label a small part of the causes of traffic congestion events from the traffic data. Our goal is to simultaneously discover known and unknown causes from a large amount of unlabeled congestion events with a limited labeled data. However, there are three challenges that need to be addressed.

- **Complicated factors influencing congestion**. Traffic congestion may occur on the same road but at different time periods, as illustrated in Figure 1. This indicates that the congestion has a strong correlation with temporal information. Meanwhile, the causes are not the same, which are related to the nearby POI types. This means that spatial information has a great impact on traffic congestion. Thus, the traffic congestion is affected by several complicated factors with the spatio-temporal correlations.
- **Limited labeled data of known congestion causes**. Because it takes much time and high cost to manually label the congestion causes, it is difficult to know the causes of most congestion events. Although the labeled data of known congestion causes is relatively limited, they can also provide some valuable information to explore which features correlate with traffic congestion. For example, the case that the traffic congestion is caused by going to schools represents the importance of POI information. Thus, how to effectively utilize the limited labeled data of known congestion causes is challenging.
- **More unknown congestion causes**. Although we can empirically list some categories of congestion causes, it is hard to cover all the congestion events because several factors would contribute to traffic congestion to different degrees. For example, the school location and the school hours both play a critical role in traffic congestion on *link1* in Figure 1. To sum up, there are many unknown congestion causes. How to explore the unknown causes from a large amount of unlabeled congestion events is the third challenge.

In order to address them, we design and implement a system that has ability to explore known and novel congestion cause based on limited supervised data. This system contains two modules: 1) congestion feature extraction, which extracts the key features related to the traffic congestion events based on real-world traffic data; and 2) congestion cause discovery, which designs a deep semi-supervised learning based method based on the limited labelled data to discover know and novel causes of traffic congestion. Specifically, we design an encoder-decoder neural network, and pre-train the model based on limited labelled data under the supervision of the reconstruction loss and classification loss. Further, we transfer the knowledge from pre-trained model, and perform clustering on learned feature representations with the reconstruction loss and Kullback-Leibler (KL) divergence loss. Benefit from the transferred knowledge, our clustering method can achieve better performance in accuracy.

We summarized our contributions as follows.

- To the best of our knowledge, we are first to automatically discover unknown causes of traffic congestion with limited labeled data.
- We quantitatively study the representative features of traffic congestion events induced by various causes. These features are extracted from both spatial and temporal aspects, including the related POI information, road type, the start and end time as well as how the congestion events evolve.
- We propose a deep semi-supervised learning based approach on the limited labelled data of known congestion causes. In this approach, we transfer the pre-trained model parameters from the labelled data, and perform the deep embedded clustering algorithm based on learned features representation of traffic congestion events.
- We conduct extensive experiments based on real-world dataset. The results demonstrate that our proposed approach achieves a higher accuracy than the state-of-the-art methods, and can discover unknown causes at the same time.
- We deploy the system in the production environment, which shows its capacity of discovering the congestion causes.

2 OVERVIEW

2.1 Preliminary

Definition 1 (*Traffic condition level*): The traffic condition has four levels: unobstructed, slow, congested and severely congested, which is specified according to the algorithm designed by Amap. Note that there is a unique condition level $c_{t,i}$ on a road segment i at time slot t.

Definition 2 (*Traffic congestion event*): We define a traffic congestion event occurred on road segment i as a sequence $x(i)$ =

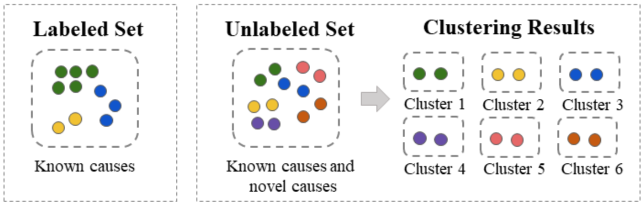


Figure 2: An illustration of our research problem.

$[(t_s, c_{t_s,i}), (t_{s+1}, c_{t_{s+1},i}), ..., (t_e, c_{t_e,i})]$, $t_s < t_e$, which satisfies the following three conditions:

(1) The values of $c_{t_s,i}$ and $c_{t_e,i}$ are congested or severely congested condition.

(2) None of the values of set $[c_{t_e+1,i}, c_{t_e+2,i}, ...c_{t_e+T,i}]$ are congested or severely congested condition. T is a threshold, which is equal to 5 minutes.

(3) $t_e - t_s >= T_d$, in this paper, T_d is equal to 5 minutes.

2.2 Problem Statement

Given a traffic congestion event set $D^l = \{x_i^l, i = 1, \cdots, N\}$ with the labelled causes, a congestion event set $D^u = \{x_i^u, i = 1, \cdots, M\}(M \gg N)$ with the unlabelled causes, as well as the features related to the congestion events in D^l and D^u, we aim to discover the causes of the congestion events in D^u, which may be known or novel. Figure 2 give a clear illustration to our research problem. In Figure 2, we have a labeled set of knowing congestion causes, but in reality, there are a large number of unlabeled events. The unlabeled set contains both known causes and novel causes, and our goal is to separate these events according to their features.

2.3 System Overview

As shown in Figure 3, the overall system consists of two main modules, i.e., feature extraction module and congestion discovery module. To be specific, the *feature extraction* module takes multi-source data including the POIs, road networks, traffic data as the input, and extract the congestion features based on the analysis of the labeled congestion events with known causes. Since the traffic congestion events are influenced by spatial and temporal factors, the feature extraction is divided into two parts: spatial features and temporal features. The *congestion discovery* module uses the limited labeled congestion event data as the prior knowledge to pre-train a classification model, and then transfers it into the clustering task of unlabed data to infer the congestion event causes.

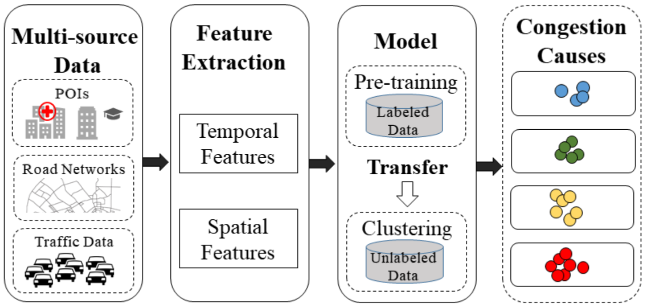


Figure 3: System overview.

Table 1: Summary of Known Traffic Congestion Causes.

Causes	Notation	# of Congestion Events
GH	Going to the Hospital	5395
AS	Attending school	3500
LS	Leaving school	1640
OR	Line up at off-ramps	6722
GS	Going to scenic spots	160
TS	Line up at toll station	62
TA	Traffic accident	787
RS	Road construction	253
TR	Traffic restriction	15

3 CONGESTION EVENT FEATURE EXTRACTION

3.1 Dataset

Firstly, according to our definition of a traffic congestion event mentioned in Section 2, we extract a large number of discrete congestion events from historical traffic condition data. This data comes from AutoNavi department in Alibaba Group and records four traffic conditions, including unobstructed, slow, congested and severely congested. Then, our dataset is divided into two parts: labeled and unlabeled traffic congestion events.

3.1.1 *Labeled Traffic Congestion Events.*

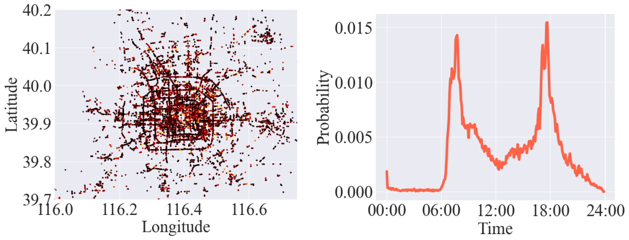
First, we introduce the dataset of traffic congestion causes labeled by domain experts in Table 1. Domain experts from AutoNavi department in Alibaba Group rely on their background knowledge to speculate on the causes of traffic congestion. Due to the high time cost and labor cost, our domain experts only label a few traffic congestion events in Beijing. Specifically, for periodic traffic congestion, the congestion events caused by hospitals are labeled by grade A hospitals only. The congestion caused by scenic spots is labeled by 5A scenic spots only. For other non-periodic traffic congestion, our domain experts also labeled the congestion events caused by traffic accidents, road construction and traffic restriction.

3.1.2 *Unlabeled Traffic Congestion Events.*

In addition to the labeled data, we also randomly sampled 55,638 unlabeled traffic congestion events in Beijing. These events extracted from historical traffic condition data from September 5, 2021 to September 11, 2021, which records road conditions for each link of road every minute. Therefore, the data has a very high temporal resolution and can reflect a lot of congestion properties. We visualize the spatial distribution and temporal distribution of these randomly sampled congestion events in Figure 4. It is observed that these randomly sampled events cover a large part of Beijing. In terms of temporal distribution, most of the congestion occurs during the morning and evening rush hours, which is consistent with what we expect.

3.2 Feature Extraction

Traffic congestion events are affected by complex spatial and temporal factors, which lead to different congestion causes. Based on the labeled traffic congestion event data, we introduce the extracted features from the spatial and temporal aspects.



(a) Spatial distribution (b) Temporal distribution

Figure 4: The overview of unlabeled congestion causes.

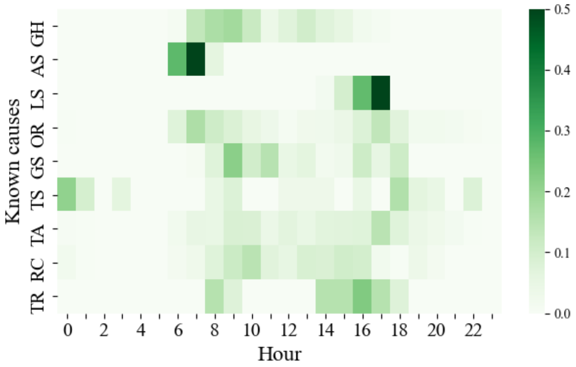


Figure 5: The distribution of start time.

3.2.1 *Temporal Features.*

The start time. The start time is an important feature to describe a traffic congestion event. The distribution of congestion events induced by difference causes in one day is shown in Figure 5. We observe that the start time of the congestion events with different causes is different. For example, the congestion events caused by attending school occur between 6AM and 8AM, while those caused by leaving school happen between 4PM and 6PM.

Three stages of congestion. A traffic congestion event usually experiences three stages, including congestion formation, congestion persistence and congestion dissipation, which is the important information to characterize the traffic congestion. To quantify it, we define the congestion formation as the slow moving stage before congestion, and the congestion dissipation process as the slow moving stage after the end of congestion. Then, we extract the duration of the congestion formation, congestion persistence and congestion dissipation. The normalized formation duration (NFD), the normalized congestion duration (NCD) and the normalized dissipation duration (NDD) of congestion events induced by different causes are illustrated in Figure 6. We find that the congestion events induced by different causes show different patterns of the duration, among which the duration of line up at off-ramps is the longest.

Traffic condition. The stop-and-go trend is a frequently observed phenomenon in traffic congestion.To quantify it, we calculate the proportion of four traffic condition levels in the process of traffic congestion.

Traffic speed. The traffic speed describes how fast the vehicles can drive on a road, which is an important indicator to evaluate how severe the traffic congestion is. The traffic congestion events with different causes have different patterns in the traffic speed before congestion, during congestion and after congestion, so we choose the average speed in these three phases as the key features.

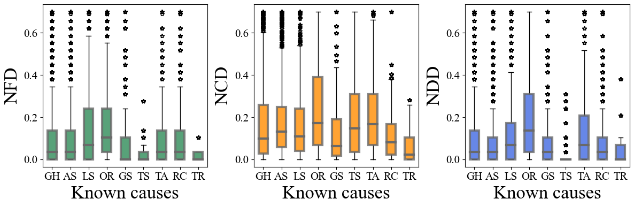


Figure 6: The duration in three stages during congestion.

Traffic volume. The traffic volume describes how many vehicles drive on a road at a specified time. When the traffic congestion occurs, the traffic volume would be decrease. Similar to the traffic speed, there are different patterns in the traffic volume before congestion, during congestion and after congestion. Thus, we select the traffic volume before congestion, during congestion and after congestion as the important features.

3.2.2 *Spatial Features.*

Points of Interests (POIs). POIs, as a location with a certain function, reflects the land use of an area. Since the traffic congestion is usually influenced by the surrounding environment, POI has a correlation with the congestion. Hence, we acquire a real-world POIs data and road networks data for Beijing, and extracts 1, 618, 605 POIs. POIs is divided into 9 categories, including Education, Finance, Shopping, Residence, Entertainment, Hospital, Scenic spots, Transportation and Toll station. In order to extract the semantic information of the location of congestion events, it is necessary to design a reasonable method to map POIs to the road network. We take the midpoint of each road section as the center of the circle and 500 meters as the radius to obtain a circular area. Then, we calculate the number of categories of POIs in the circle. Since the distribution of POI categories in a city is imbalanced, we standardize the number of POIs, and use the maximum value of the standardized POI vector as the functional category of a road segment.

Road type. Different types of roads have different traffic capacity, which has a great impact on the traffic congestion. For example, the road segment with low capacity is more prone to be congested. Hence, we select the road types as the key congestion features.

4 CONGESTION EVENT CAUSE DISCOVERY

In this section,we will describe our two-stage method, deep transfer clustering for spatio-temporal data (ST-DTC) in detail. As shown in Figure 7, we firstly extract features of traffic congestion events and embed them into a feature vector. After that, a classifier will be trained with limited labeled data. In the second stage, we transfer the knowledge learned from labeled data to unlabeled data via transfer clustering. In both the pre-training stage and the clustering stage, we use the reconstruction loss as an auxiliary training objective. In this way, ST-DTC can not only learn useful knowledge for classification task, but also retain the feature of the input data.

4.1 Feature Representation Initialization

Since a traffic congestion event has multiple features introduced in Section 3.2, we have to initialize these features as the effective feature representations for the input of ST-DTC.

From the temporal aspect, for the *start time*, it is represented with hour-of-day, and is encoded into a one-hot vector $x_s^h \in \mathbb{R}^{24}$. We

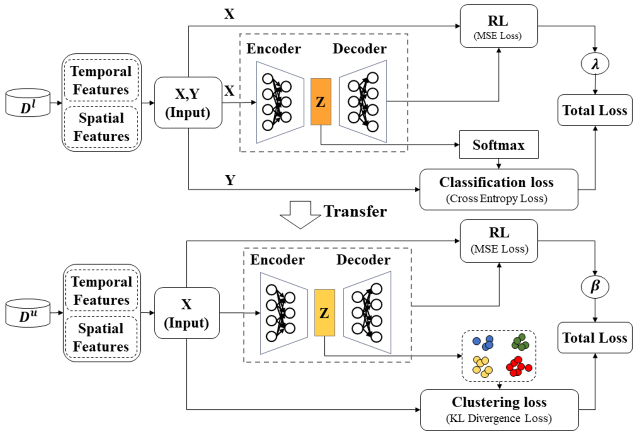


Figure 7: Our clustering framework. Firstly, we train a classifier with limited labeled data. Then, we transfer the knowledge learnt from labeled data to unlabeled data via transfer clustering. (RL : reconstruction loss)

also consider whether a event occurring at weekdays or weekends, and encode day-of-week into $\boldsymbol{x}^d_s \in \mathbb{R}^2$ using one-hot coding. After that, we concatenate them into a vector $\boldsymbol{x}_s = [\boldsymbol{x}^h_s, \boldsymbol{x}^d_s] \in \mathbb{R}^{24+2}$, where [,] denotes the concatenation operator. For *three stages of the congestion*, their duration is normalized and represented as a vector $\boldsymbol{x}_d \in \mathbb{R}^3$. For the *traffic condition*, we embed the proportion of four traffic condition level during a traffic congestion event into a vector $\boldsymbol{x}_p \in \mathbb{R}^4$. For the *traffic speed* and *traffic volume*, we encode the speed and volume before congestion, during congestion and after congestion into $\boldsymbol{x}_v \in \mathbb{R}^3$ and $\boldsymbol{x}_r \in \mathbb{R}^3$, respectively.

From the spatial aspect, For the *POI* feature, we represent the functional category of a road segment into a one-hot vector $\boldsymbol{x}_o \in \mathbb{R}^9$. For the *road type*, we encode it into a one-hot vector $\boldsymbol{x}_w \in \mathbb{R}^9$.

Finally, the congestion feature representation can be initialized as:

$$\boldsymbol{x} = [\boldsymbol{x}_s, \boldsymbol{x}_d, \boldsymbol{x}_p, \boldsymbol{x}_v, \boldsymbol{x}_r, \boldsymbol{x}_o, \boldsymbol{x}_w]. \tag{1}$$

4.2 Pre-training with Limited Supervised Data

We firstly pretrain a classifier to help the task of clustering rather than considering a fully unsupervised setting, which is motivated by Deep Transfer Clustering (DTC) [5].

Autoencoder. The main network used in our model is Autoencoder [23]. Both of our encoder and decoder are fully connected layer. The encoder map the congestion feature representation $\boldsymbol{x}$ to a latent feature space $z \in \mathbb{R}^d$, and d is smaller than the dimension of input data x. The latent features z represent the most important component of input data. Then, the decoder learns to reconstruct $\bar{x}$ from its latent features.

$$z = \sigma(w^e \cdot \boldsymbol{x} + b^e), \tag{2}$$

$$\bar{\boldsymbol{x}} = \sigma(w^d \cdot \boldsymbol{x} + b^d), \tag{3}$$

where $\sigma(\cdot)$ is the activation,w^e and b^e are the parameters of the encoder, and w^d and b^d are the parameters of the decoder.

Objective of pre-training. The objective of pre-training stage in ST-DTC is as follows,

$$L_{pretrain} = -\frac{1}{N}\sum_{i=1}^{N}\log \eta^{l}_{y_i}(z^l_i) + \lambda\sum_{i=1}^{N}(x^l_i - \bar{x}^l_i)^2, \tag{4}$$

where x^l_i means the ith labeled sample,z^l_i and $\bar{x}^l_i$ are its latent features and reconstruction respectively. The first term is the classification loss, which is classic cross entropy(CE) loss. The second term is the reconstruction loss, which is mean square error(MSE). The second term reduces the difference of input data x^l_i and its reconstruction $\bar{x}^l_i$. λ is the weight of the reconstruction loss.

4.3 Transfer Clustering

After pre-training the model with limited supervised data, we transfer the pre-trained model parameters to the second stage.

The second stage is a clustering task extended from Deep Embedded Clustering (DEC) [26]. DEC is the first time to optimize dimensionality reduction and clustering simultaneously using deep neural networks. Given initial cluster centers $\{\mu_1, \cdots, \mu_k\}$ and latent features z, DEC uses a student's t-distribution to measure the similarity between cluster center μ_i and data point x_i as follows,

$$q_{ij} = \frac{(1 + ||z_i - \mu_j||^2/\alpha)^{-\frac{\alpha+1}{2}}}{\sum_k (1 + ||z_i - \mu_k||^2/\alpha)\frac{\alpha+1}{2}}, \tag{5}$$

where α is the degree of freedom of the Student's t-distribution. Following DEC, we also let $\alpha = 1$ for all experiments. q_{ij} can be regarded as the probability of assigning sample i to cluster j. Then KL-divergence loss is used to guide model training.

$$KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \tag{6}$$

$$p_{ij} = \frac{q^2_{ij}/f_j}{\sum_k q^2_{ij}/f_k}, \tag{7}$$

where p_{ij} is an auxiliary distribution and $f_k = \sum_i q_{ij}$ is soft cluster frequencies.

Objective of clustering. DEC exceeds the baseline model greatly, however, its loss function is represented in clustering loss only. A recent study found that the model combining the two tasks works better [27]. Thus, in the clustering stage, the model was optimized by the reconstruction loss and clustering loss simultaneously, and the objective is as follows,

$$L_{train} = KL(P||Q) + \beta\sum_{i=1}^{N}(x^u_i - \bar{x}^u_i)^2, \tag{8}$$

where x^u_i means the ith unlabeled sample, $\bar{x}^u_i$ is its reconstruction. The first term is the clustering loss, which is a Kullback–Leibler (KL) divergence loss between the soft assignments q_i and the auxiliary distribution p_i. The second term is the reconstruction loss, which is mean square error(MSE). β is the weight of the reconstruction loss. We let $\beta = 1$ for all experiments.

5 EXPERIMENTS RESULTS

In this section, we evaluate our system by the extensive experiments. First, we present the experiment results compared to different baselines. After that, the novel congestion causes are analyzed in detail.

5.1 Baselines

We firstly compare our method (ST-DTC) with unsupervised clustering methods, including K-means (KM) [15], DEC [26] and DCN [27]. K-means is a classic clustering algorithm, it iteratively assigned each object to its nearest cluster center by computing the distance. DEC is the first time to optimize dimensionality reduction and clustering simultaneously using deep neural networks. However, DEC trains the model by using clustering loss only. DCN expanded the DEC by optimizing the decoder and clustering simultaneously.

We also compare our method with semi-supervised clustering methods, including KCL [6], MCL [7] and DTC [5]. The idea of semi-supervised clustering methods is to aid the discovery of novel categories by leveraging a certain number of known categories. KCL trains a classifier with limited labeled data by considering pairwise similarity. MCL improves KCL to train the classifier using a probabilistic-graphical-model-based loss function. However, neither KCL nor MCL has training process for the clustering stage. Therefore, DTC combines DEC training methods in clustering stage.

5.2 Evaluation Metrics

We use the conventional clustering accuracy (ACC), normalized mutual information (NMI) [22] and adjusted Rand index (ARI) [19] to evaluate the clustering performance of our approach. All the metrics are valued in the range of [0, 1] and higher values mean better performance.

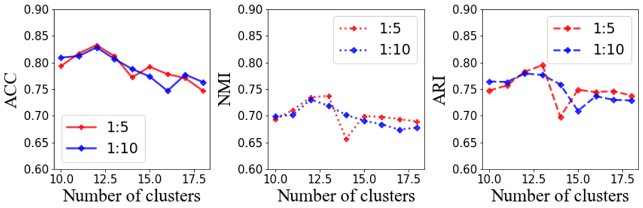
5.3 Experimental Settings

We introduce our experimental settings in terms of data preparation, network architectures and training configurations.

Data preparation. Firstly, we split our dataset into labeled data and unlabeled data. The labeled data contain 70% samples of all known causes events. The unlabeled data contain 30% samples of known causes events and other unknown congestion events. In our dataset, the sample size of some categories of causes was very small. For example, we only have 15 samples for TR (traffic restriction). For small data, we apply the oversampling method to increase the number of samples for both pre-training phase and clustering stage. In pre-training stage, for small sample data, each sample is replicated 10 times. In clustering phase, for small sample data, each sample is replicated 6 times.

Network architectures. The deep learning model is implemented with Pytorch. We use a fully connected encoder with 3 layers. All of the layers have *Relu* activation function and dropout rate of 0.1. The number of hidden units is (128, 64, 12) in three fully-connected layers. The decoder also is a fully connected network with 3 layer. The number of hidden units of the decoder is (12, 64, 128).

Training configurations. In the pre-training stage, we train the feature extractor (MLP) with a batch size of 128, and the fully connected layer has a dropout layer. Adam optimizer is selected with a learning rate of 0.005 for 5 epochs. The weight of the reconstruction loss term is 0.1. In the clustering stage, We fine-tune the pre-trained model with a batch size of 1280. Adam optimizer is also selected with a learning rate of 0.0005 for 20 epochs. The weight of the reconstruction loss term is 1.0.



(a) ACC (b) NMI (c) ARI

Figure 8: The accuracy of different numbers of clusters.

Table 2: The clustering results on two datasets with different ratios of sample of known causes and unknown causes.

Ratio (known:unknown)	1:5			1:10		
	ACC	NMI	ARI	ACC	NMI	ARI
KM	0.3661	0.3523	0.2740	0.4201	0.3610	0.3312
DEC	0.5531	0.4404	0.4698	0.4738	0.3943	0.3358
DCN	0.4106	0.3596	0.2499	0.3964	0.3494	0.2417
KCL	0.6102	0.5241	0.6009	0.5502	0.5061	0.5709
MCL	0.6073	0.5280	0.6093	0.5829	0.5734	0.6578
DTC	0.7950	0.7022	0.7459	0.7981	0.6946	0.7579
ST-DTC	**0.8326**	**0.7347**	**0.7828**	**0.8282**	**0.7307**	**0.7794**

5.4 Determining the Number of Clusters

We compare performance of known categories of congestion causes to determine the optimal number of clusters. Figure 8 shows the value of three metrics for the given number of clusters. From the results, we find that the optimum number of clusters is around 12.

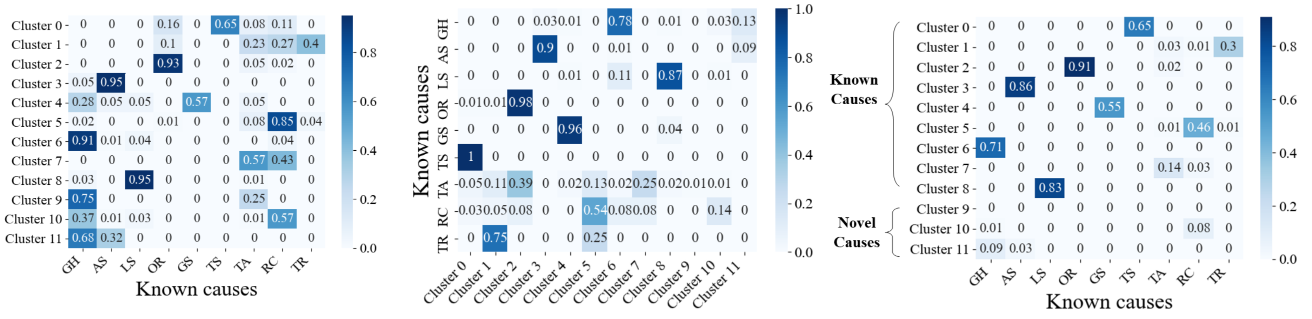
5.5 Overall Performance

We summarize our results in Table 2. Our results on all datasets are averaged over 10 runs. Our method achieves the best results and outperforms other baselines. For baselines, unsupervised clustering methods (KM, DEC, DCN) perform worse than semi-supervised clustering methods (KCL, MCL, DTC). Compare with DTC, the accuracy of our model is improved by around 5%. The pre-training stage in DTC is optimized by classification loss using the labeled data. However, this makes the latent features overly-specialize for the labeled data, and thus DTC may provides a poor initialized representation of the novel congestion causes. Moreover, the clustering stage in DTC is optimized by clustering loss only, which fails to capture the original characteristics of the input data. Since ST-DTC adds the idea of self-supervision (i.e., autoencoder), the model can retain its own important feature while learning classification task and clustering task. Therefore, ST-DTC is more robust and can achieve better performance.

5.6 Cluster Analysis

In this section, we will analyze the clustering results in detail. The dataset with labeled data and unlabeled data ratio of 1 : 5 was selected as a case. All of the following analysis are based on the test dataset.

Latent features. The clustering result of testing dataset is represented in Figure 10. We use a t-distributed stochastic neighbor embedding (TSNE) [10] to visualize the clusters. TSNE can show the learned latent features of ST-DTC in two dimension space. The color represents different clusters, and finally we obtain 12 clusters. In Figure 10, these clusters are separated well, which proves that our model is valid. The larger the number of scatter points, the



(a) T_k

(b) T_c

(c) T_d

Figure 9: The distribution of the congestion events with known causes on clusters. T_k is the intra-cluster distribution of known causes. T_c is the inter-cluster distribution of each known cause. T_d is the final decision matrix, $T_d = T_k \times T_c^{\mathrm{T}}$.

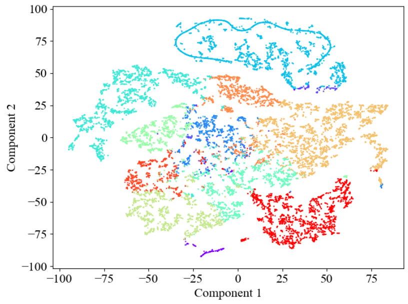


Figure 10: TSNE visualization of clusters.

larger the size of the cluster. We can see that the size of each cluster is unbalanced, which is in line with reality. The causes inducing traffic congestion are also diverse and the size of corresponding events unbalanced in urban environment. For example, the size of periodic congestion events induced by some causes (i.e., attending and leaving school) is usually larger than that of non-periodic congestion events induced by other causes (i.e., traffic accidents).

Known and novel causes. Our test dataset contains both known and novel congestion causes, and we want to analyze which clusters are known or novel causes in Figure 10.

To quantify which clusters correspond to known congestion causes, we need to consider the intra-cluster and inter-cluster distribution of each known cause. Here we compute two matrices T_k and T_c based on the clustering result in Figure 9(a) and Figure 9(b). T_k is the intra-cluster distribution of each known cause. T_c is the inter-cluster distribution of each known cause. Then, based on the cross of T_k and T_c, we get the final decision matrix T_d.

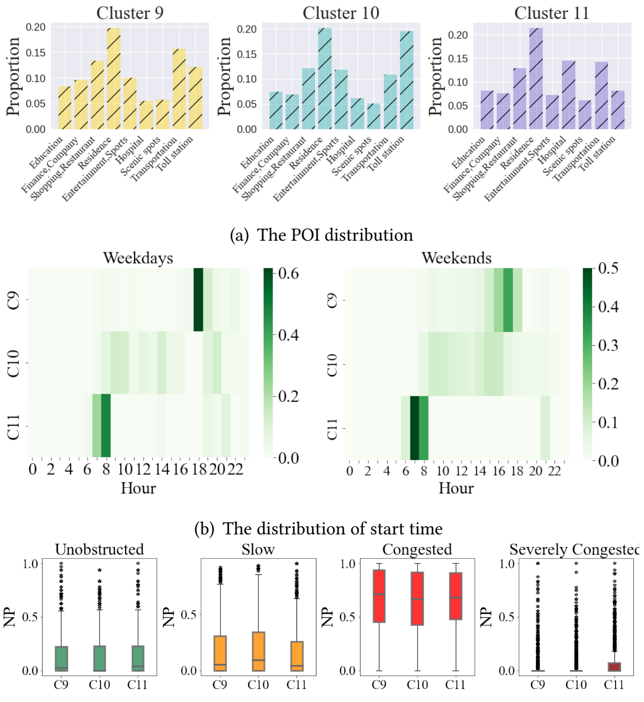
$$T_k(ij) = \frac{N_{kc}}{\sum_{k=1}^{K} N_{kc}}, \tag{9}$$

$$T_c(ij) = \frac{N_{kc}}{\sum_{c=1}^{C} N_{kc}}, \tag{10}$$

where N_{kc} means the number of samples the kth known cause assigned to cth cluser, $k = 1, 2, 3, \cdots, K$, K is the number of known causes, and $c = 1, 2, 3, \cdots, C$, C is the number of clusters. In our experiment, $K = 9$ and $C = 12$. By considering the intra-cluster and inter-cluster distribution of each known causewe obtain the decision matrix T_d as follows,

$$T_d = T_k \times T_c^{\mathrm{T}}. \tag{11}$$

The matrix T_d is showed with a heatmap in Figure 9(c). Each known cause has a pair of corresponding cluster, our clustering model separates the known causes well even with the challenges of small samples. Furthermore, we discover three novel causes independent of known causes, including Cluster 9, Cluster 10 and Cluster 11. These novel causes correspond to different spatio-temporal patterns.



(a) The POI distribution

(b) The distribution of start time

(c) Traffic condition level

Figure 11: The statistics of the congestion events with novel causes on clusters.(C9, C10, C11 mean Cluster 9, Cluster 10 , Cluster 11). NP means normalized proportion.

5.7 Novel Congestion Cause Analysis

In this section, we will analyze the traffic features of novel congestion causes, which is mainly made from the time, spatial attributes and congestion characteristics of traffic congestion events.

POIs feature. The POIs feature of three novel causes is visualized in Figure 11(a). POIs feature of a congestion event is represented with a one-hot vector, and we count the proportion of each POI categories in each novel cause. We find that the distribution of Cluster 9, Cluster 10 and Cluster 11 are basically similar, which proves the complication of congestion causes. Specifically, this result shows congestion events occur induced by different causes even in the similar surrounding environment.

The start time. The distribution of start time of the congestion events with three novel causes is visualized in Figure 11(b). Obviously, we observe different temporal patterns. Most of the congestion events in Cluster 11 occurs on morning at both weekdays and weekend. However, congestion events usually occur at afternoon.

The stop-and-go trend. The stop-and-go trend is an essential feature of urban traffic congestion, which has been mentioned in Section 3. We calculated the proportion of four road conditions in a complete congestion process, namely unobstructed, slow, congested and severely congested in Figure 11(c). These congestion events have higher proportion of slow state in Cluster 10, which means they are more likely to be stop-and-go. Different from the other two causes, the congestion events in Cluster 11 are more likely to be severely congested.

6 SYSTEM DEPLOYMENT

To embed our model to the map application, we release the model internally in forms of the package. After that, we deploy the model package to the machine learning platform, and the application programs will automatically run our model by the designed interfaces.

6.1 System Interface

To offer better user experience, we design three views in the system interface, as shown in Figure 12.

Request View: The user requests the starting location and destination. Next, the driving mode should be selected.

Map View: This view shows a recommendation route consisting of multiple road segments between the starting location and destination. The colors on the road segments represent different traffic conditions, where the red color means congested, and the green is unobstructed. A message icon on a road segment with the red color shows the cause of traffic congestion, as well as other congestion information including the average driving speed.

Navigation View: In this view, the user is using the navigation service for driving. Similar to *map view*, A message icon appears on the congested road segment with the red color, which shows the cause of traffic congestion, as well as other congestion information including the congestion length and the congestion duration.

6.2 Industrial Deployment and Results

We deploy our system on the machine learning platform named AI Studio. The deployment consists of offline training and online inference phases.

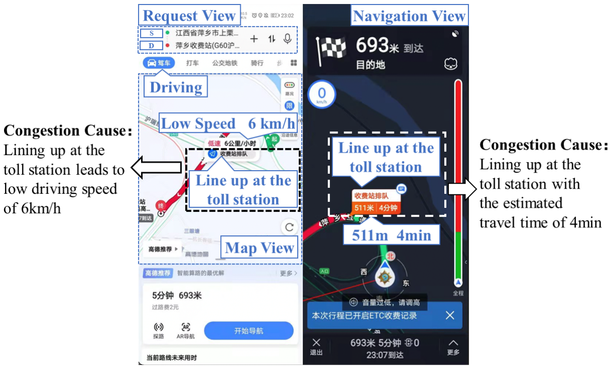


Figure 12: System interface. 'S' and 'D' represent staring location and destination respectively.

Offline Training: During this phase, the model parameters are learned based on the labeled congestion event data, which are updated periodically.

Online Inference: In this phase, a user inputs a request to look up the driving routes. After that, the system extracts all the congested road segments along the route, and queries the traffic data storage to obtain the recent time traffic conditions of these road segments. Then, it runs the transfer clustering model to explore the congestion causes on the road segments. Finally, the results of all the congested road segments are returned and displayed in the *map view* or *navigation view*.

Results: In online environment, our model can discover 25 congestion causes, which obtains high user satisfaction. With this function, users can plan their trip in advance. Even when users are experiencing the traffic congestion, it is helpful to alleviate their anxieties and improve user experience.

7 RELATED WORK

Causal inference of traffic congestion. Many researchers focus on studying the causes of traffic congestion [3, 13, 18, 21]. Bao et al. [2] investigated traffic congestion caused by the demand from residential area to the suburb resorts during public holidays, especially when the toll-exemption policy is performed. Authors in [21] used the k-means algorithm to identify the most important factors affecting traffic congestion based on real traffic data, and obtained several insightful findings, e.g., the early evening peak congestion is more likely to be caused by the number of bus stops. Chawla [3] designed a mining and optimization framework to detect the cause of traffic anomalies like traffic congestion in road traffic flow. Unlike them, our proposed method not only accurately discovers known causes, but also effectively identifies novel ones, which can be applied to practical scenarios of traffic management.

Traffic condition prediction. Understanding traffic condition plays an essential part in studying the causes of traffic congestion in urban environment. Many researchers propose algorithms to predict traffic conditions such as the average speed, traffic volume, and traffic states [4, 14, 29]. For example, Zhang et al. [29] combined weather condition data to predict traffic flow based on the gated recurrent unit (GRU). Recently, several works utilized Convolutional neural network (CNN) or graph convolutional network (GCN) with the recurrent models to explore the spatio-temporal dependencies to further improve the traffic prediction performance. Chen et al. [4] proposed a deep convolutional neural network to model periodic traffic data for short-term traffic congestion prediction. Wang et al. [24] also considered the periodicity of traffic congestion, and designed a periodic spatial-temporal neural network to forecast traffic condition. Different from them, we focus on exploring the congestion causes based on the in-depth analysis of traffic data.

Clustering methods. There are two types of clustering methods related to our work, including unsupervised method and semi-supervised method. Unsupervised clustering methods aim to learn the patterns from unlabelled data. They include partitioning methods [15], hierarchical methods [8] and density-based methods [9]. Recently, deep learning methods are introduced to capture complex dependencies for clustering [1]. However, these methods do not take account into how to utilize limited labeled data as the useful prior knowledge for clustering. Due to the importance of prior knowledge, the semi-supervised clustering with the aid of some labeled data has attracted the attention of researchers. For example, Han et al. [5] presented the deep transfer clustering approach to discover novel visual categories. Lin et al. [12] designed the constrained deep adaptive clustering method with cluster refinement to discover new intents in the dialogue system. Inspired by them, we propose an effective approach by transferring the knowledge from limited labeled traffic data to simultaneously discover known and novel causes of traffic congestion.

8 CONCLUSION

We propose a novel and systematic approach to discover the known and novel causes of traffic congestion based on the limited labeled data. We evaluate our model with the extensive experiments, and the results show that our model achieves better performance in terms of three metrics. Finally, we deploy our system in production environment, and the online results show that our model has ability to discover more congestion causes. When pushing the congestion cause information to users by the map application, it can bring great benefits. For example, when users are experiencing the traffic congestion, their anxieties can be alleviated by knowing the congestion cause information. To conclude, it is an important and practical solution in real-world map services.

9 ACKNOWLEDGEMENTS

This work was supported in part by The National Key Research and Development Program of China under grant 2020AAA0106000, the National Nature Science Foundation of China under U20B2060, 61971267, 61972223.

REFERENCES

[1] Reza Asadi and Amelia Regan. 2019. Spatio-temporal clustering of traffic data with deep embedded clustering. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Prediction of Human Mobility*. 45–52.

[2] Yue Bao, Feng Xiao, Zaihan Gao, and Ziyou Gao. 2017. Investigation of the traffic congestion during public holiday and the impact of the toll-exemption policy. *Transportation Research Part B: Methodological* 104 (2017), 58–81.

[3] Sanjay Chawla, Yu Zheng, and Jiafeng Hu. 2012. Inferring the root cause in road traffic anomalies. In *2012 IEEE 12th International Conference on Data Mining*. IEEE, 141–150.

[4] Meng Chen, Xiaohui Yu, and Yang Liu. 2018. PCNN: Deep convolutional networks for short-term traffic congestion prediction. *IEEE Transactions on Intelligent Transportation Systems* 19, 11 (2018), 3550–3559.

[5] Kai Han, Andrea Vedaldi, and Andrew Zisserman. 2019. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8401–8409.

[6] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. 2017. Learning to cluster in order to transfer across domains and tasks. *arXiv preprint arXiv:1711.10125* (2017).

[7] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. 2019. Multi-class classification without multi-class labels. *arXiv preprint arXiv:1901.00544* (2019).

[8] Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32, 3 (1967), 241–254.

[9] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek. 2011. Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1, 3 (2011), 231–240.

[10] Geoffrey Hinton Laurens van der Maaten. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[11] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. 2017. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926* (2017).

[12] Ting-En Lin, Hua Xu, and Hanlei Zhang. 2020. Discovering new intents via constrained deep adaptive clustering with cluster refinement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8360–8367.

[13] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan, and Xie Xing. 2011. Discovering spatio-temporal causal interactions in traffic data streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1010–1018.

[14] Xiaolei Ma, Zhimin Tao, Yinhai Wang, Haiyang Yu, and Yunpeng Wang. 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies* 54 (2015), 187–197.

[15] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 281–297.

[16] PR Newswire. 2021. New Harris Survey Reveals Chronic Nasal Congestion Is Surprisingly Widespread: Nearly 1 in 4 Americans With Nasal Congestion Experience Symptoms Almost Every Day. (2021).

[17] Zheyi Pan, Yuxuan Liang, Weifeng Wang, Yong Yu, Yu Zheng, and Junbo Zhang. 2019. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1720–1730.

[18] Mingyu Pi, Hanbyul Yeon, Hyesook Son, and Yun Jang. 2019. Visual cause analytics for traffic congestion. *IEEE transactions on visualization and computer graphics* (2019).

[19] William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66, 336 (1971), 846–850.

[20] David Schrank, Bill Eisele, and Tim Lomax. 2021. Urban mobility report 2021. (2021).

[21] Jinchao Song, Chunli Zhao, Shaopeng Zhong, Thomas Alexander Sick Nielsen, and Alexander V Prishchepov. 2019. Mapping spatio-temporal patterns and detecting the factors of traffic congestion with multi-source data fusion and mining techniques. *Computers, Environment and Urban Systems* 77 (2019), 101364.

[22] Alexander Strehl and Joydeep Ghosh. 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research* 3, Dec (2002), 583–617.

[23] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11, 12 (2010).

[24] Tiange Wang, Zijun Zhang, and Kwok-Leung Tsui. 2021. PSTN: Periodic Spatial-temporal Deep Neural Network for Traffic Condition Prediction. *arXiv preprint arXiv:2108.02424* (2021).

[25] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. 2019. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121* (2019).

[26] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*. PMLR, 478–487.

[27] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *international conference on machine learning*. PMLR, 3861–3870.

[28] Bing Yu, Haoteng Yin, and Zhanxing Zhu. 2017. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875* (2017).

[29] Da Zhang and Mansur R Kabuka. 2018. Combining weather condition data to predict traffic flow: a GRU-based deep learning approach. *IET Intelligent Transport Systems* 12, 7 (2018), 578–585.