C-Watcher: A Framework for Early Detection of High-Risk Neighborhoods Ahead of COVID-19 Outbreak

Congxi Xiao[1,2*], Jingbo Zhou[2†*], Jizhou Huang[2*], An Zhuo[2], Ji Liu[2], Haoyi Xiong[2], Dejing Dou[2†]
[1]University of Science and Technology of China, [2]Baidu Inc., China
{v_xiaocongxi, zhoujingbo, huangjizhou01, zhuoan, liuji04, xionghaoyi, doudejing}@baidu.com

Abstract

The novel coronavirus disease (COVID-19) has crushed daily routines and is still rampaging through the world. Existing solution for nonpharmaceutical interventions usually needs to timely and precisely select a subset of residential urban areas for containment or even quarantine, where the spatial distribution of confirmed cases has been considered as a key criterion for the subset selection. While such containment measure has successfully stopped or slowed down the spread of COVID-19 in some countries, it is criticized for being inefficient or ineffective, as the statistics of confirmed cases are usually time-delayed and coarse-grained. To tackle the issues, we propose C-Watcher, a novel data-driven framework that aims at screening every neighborhood in a target city and predicting infection risks, prior to the spread of COVID-19 from epicenters to the city. In terms of design, C-Watcher collects large-scale long-term human mobility data from Baidu Maps, then characterizes every residential neighborhood in the city using a set of features based on urban mobility patterns. Furthermore, to transfer the firsthand knowledge (witted in epicenters) to the target city before local outbreaks, we adopt a novel adversarial encoder framework to learn "*city-invariant*" representations from the mobility-related features for precise early detection of high-risk neighborhoods, even before any confirmed cases known, in the target city. We carried out extensive experiments on C-Watcher using the real-data records in the early stage of COVID-19 outbreaks, where the results demonstrate the efficiency and effectiveness of C-Watcher for early detection of high-risk neighborhoods from a large number of cities.

1 Introduction

The novel coronavirus disease (COVID-19), which has been officially announced as a pandemic by the World Health Organization (WHO), is perhaps the most serious public health emergency over the past decades. The coronavirus continues to spread around the globe, which challenges the governments and medical systems all over the world.

While metropolitan-wide lockdown has demonstrated its effectiveness as a nonpharmaceutical intervention in several countries, the cost of such measures, including unemployment, economic crash, and social anxiety, makes it a tough decision on behalf of administrators. A compromised solution is to place containment measures onto a subset of areas in a city to stop or slow down the spread of COVID-19 while minimizing the social and economic cost. To precisely distinguish the high-risk areas from the city, the spatial distribution of confirmed cases has been used as the key criterion for the potential containment measures in a data-driven fashion.

While the spatial statistics of confirmed cases work, the time consumption and the granularity of data acquisition significantly lower the efficiency and effectiveness of such methods. For example, the incubation period of COVID-19 is around 5–6 days on average, but it could last as long as 14 days. During such a period, a community or neighborhood would have been already invaded by a small number of asymptotic carriers who might not be with any clinical symptoms. When a small number of cases being confirmed, the community and its surrounding neighborhoods may have fallen into COVID-19 for a long time. Furthermore, though the mobility of confirmed patients is usually well restricted, the asymptomatic carriers with no symptoms would still spread the virus to where he/she has gone. When fine-grained mobility traces are not available, the administrators can only place containment measures to places in coarse-grained.

To tackle the technical issues, in this paper, we propose C(OVID)-Watcher, a novel framework to support early detection of high-risk residential neighborhoods for fighting against the spread of COVID-19. Our intuition is to incorporate human mobility data, so as to (1) characterize the socioeconomic and demographic status of every neighborhood (Borjas 2020; Huang et al. 2020b) based on "how residents move" (Renso et al. 2013) and (2) unfold the spatial interactions (Jiang et al. 2020) and potential influences on COVID-19 caused by the mobility of massive asymptotic carriers. Specifically, C-Watcher includes a mobility data-driven machine learning model that screens every neighborhood in a target city and predicts the infection risks, prior to the spread of COVID-19 from epicenters to the city.

In addition to the use of mobility-related features, we also hope to generalize the evidence already witted in the epicenter for screening the risk of neighborhoods in the target city, prior to or in the early stage of local outbreaks. To achieve

[*]C. Xiao, J. Zhou and J. Huang contributed equally to the paper. This work was done when C. Xiao was an intern at Baidu Inc.

[†]J. Zhou and D. Dou are the corresponding authors.

the goal, a core component of C-Watcher is a novel cross-city transfer learning model that transfers knowledge about COVID-19 infections from the epicenter to the target city, while cities are quite different in a large number of domains, ranging from living, foods, transportation, and residences.

All in all, we have made three contributions as follows.

- Through extensive data analytics, we explore a set of empirical features related to long-term/regular human mobility patterns (before the COVID-19). With such long-term mobility features, the socioeconomic and demographic status, as well as the spatial interactions among neighborhoods could be well characterized. In this way, one can easily distinguish high-risk neighborhoods from the urban area and predict the potential risks for infection, with respect to the two factors.
- While these features are with certain discriminative information for risk prediction, they also involve some city-specific characteristics. For example, the popular choices for transport modes in different cities vary. Such city-specific characteristics burden the use of mobility-related features to transfer the knowledge obtained in the epicenter to the target city. To generalize the knowledge transfer, C-Watcher adopts a novel adversarial encoder-decoder framework to learn the "city-invariant" representations from the mobility-related features for prediction.
- To validate C-Watcher, we collect and construct real-world datasets for high-risk neighborhood detection based on the publicly available information from the web and human mobility traces from the largest online map service in China. We conduct extensive experiments for evaluation. The results demonstrate that C-Watcher can accurately predict the potential risk of massive residential neighborhoods in a large number of Chinese cities. With large datasets, C-Watcher makes insightful suggestions on preventing the epidemic of COVID-19 alike for different residential neighborhoods via feature importance.

2 Notations and Related Work

In this section, we first introduce the basic notations used throughout this paper, and then we formally formulate the research problem for early detection of high-risk neighborhoods. Last, we review the studies that are relevant to our work with the most related work discussed.

Notations and Formulation. We use $\boldsymbol{n}$ to denote the features of a residential neighborhood which will be presented in Section 3, and use y to denote the binary label meaning whether the neighborhood is high risky ($y = 1$) or not ($y = 0$). The detection problem can be defined as: $f(\boldsymbol{n}) \rightarrow y$ where the function $f(\cdot)$ can be any machine learning model like Multi-Layer Perceptron (MLP).

The objective of C-Watcher is to make early detection of high-risk neighborhoods without epidemic outbreaks. Instead of relying on the confirmed infection cases to make a prediction which deemed to be time-delayed like (Fu et al. 2020), we assume that the COVID-19 epidemic only outbroke in epicenter cities (such as Wuhan in China) and no prior knowledge of confirmed cases, spreading trend or known hazard neighborhoods in target cities can be referred to. Such a cross-city prediction problem of latent high risky residential neighborhoods can be formulated as:

$$f_{cross}(\boldsymbol{n}^T \mid \{(\boldsymbol{n}_i^E, y_i^E)\}) \rightarrow y^T \tag{1}$$

where $\boldsymbol{n}^T$ and y^T denote the features and binary label of a residential neighborhood in the target city. $\boldsymbol{n}_i^E$ and y_i^E denote the features and label of a neighborhood from epicenter cities set. Hereafter, we omit subscript i for simplicity. The $f_{cross}(\cdot)$ is a cross-city transfer learning model which is trained without ground-truth information in the target city.

Related Work. Aiming to fight against the COVID-19 pandemic, researchers in the computer science community carried out many studies from several perspectives recently. For instance, Huang et al. (2020b) exhibit that user transportation-related behaviors in China have indeed been impacted by the containment measures during the COVID-19 pandemic. There are also a few studies (Huang et al. 2020c; Xiong et al. 2020; Liu et al. 2020b) investigating the human mobility, the local economy, and the information acquisition during the COVID-19 outbreak in China while most of these studies remain at city level.

Some studies also demonstrate the effectiveness of mobility data for controlling the spread of COVID-19. Vollmer et al. (2020) exploit a Bayesian semi-mechanism model with mobility data to show the effectiveness to slow down the spread of the virus by constraints on individual movements and social interactions. Based on the integration of mobility data and the global epidemic model (Balcan et al. 2009), a study also reveals the effectiveness of fine-grained targeted mobility control policies towards the COVID-19 pandemic (Hao et al. 2020). The mobility data can also be integrated with compartmental models in epidemiology (like Susceptible-Exposed-Infected-Recovered (SEIR) model) (Ghamizi et al. 2020) to better predict the epidemic dynamics.

Discussion. From the problems and methodologies perspectives, the most relevant work to our study includes (Fu et al. 2020) and (Xu et al. 2019; Peng and Qi 2019; Mai, Hu, and Xing 2020). Compared to (Fu et al. 2020), which smooths the confirmed cases of infections over spatial domains and predicts hazard areas during the COVID-19 outbreaks using simple spatial features like distance, C-Watcher system tackles the time delay and coarse-grained granularity issues and can early detect the high-risk residential neighborhoods even before the outbreaks, through leveraging features derived from long-term/regular human mobility patterns. In terms of methodologies, though a great number of algorithms have been proposed for adversarial representation learning Makhzani et al. (2015), adversarial metric learning (Xu et al. 2019) and cross-modalities (Mai, Hu, and Xing 2020), our work is the first to study the city-invariant representation learning through Generative Adversarial Networks (Goodfellow et al. 2014) in the context of urban computing and COVID-19 prediction.

3 Features for Neighborhood Detection

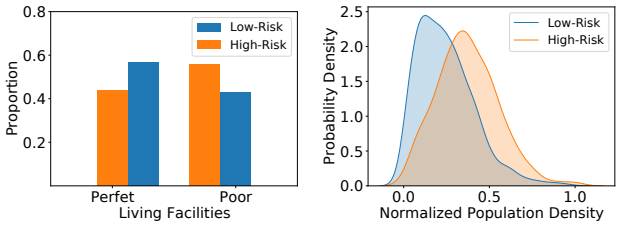
In this section, we present how to construct features from mobility data to characterize a residential neighborhood for early risk detection. We first introduce the data source used in our framework, and then three groups of constructed features are briefly discussed which are Point of Interest (POI) radius features (see Section 3.1), demographic features (see Section 3.2) and transportation-related features (see Section 3.3), respectively. More details about the feature construction can be found in the Appendix of (Xiao et al. 2020).

The feature construction is mainly based on three data sources: POI basic property data, user profile data and human mobility data. POI basic property data contains the basic information of a POI, such as name, coordinates and types, which provides many semantic information for a POI (Huang et al. 2020a; Yuan et al. 2020; Hu et al. 2020). This data enable us to analyze the spatial relationship between neighborhoods and different types of POIs, such as hospitals, schools and bus stops (Li et al. 2020). The user profile data are obtained from a user profile platform of Baidu which can return profile features for almost all internet users in China, such as gender, age and educational level. Human mobility data, collected from Baidu Maps in China, record search and transportation behaviors of the map users.

3.1 POI Radius Features

Here we introduce how to compute a group of POI radius features for a residential neighborhood based on POI basic property data. The intuition for this feature group is that basic living facilities around a residential neighborhood may have a correlation with the probability of its residents being infected by COVID-19. For example, a neighborhood lacking basic living facilities may face a high risk, for the residents may passively go further away for basic living needs and face greater infection risks. Moreover, neighborhoods with poor living facilities often lack good property management, which may also lead to high infection risks. To describe these living facilities related characteristics, we construct 15 POI radius features. Each of them is defined as the shortest distance between the neighborhood and one certain type of POIs. All used POI types are listed in the Appendix of (Xiao et al. 2020).

Meanwhile, we define an additional binary feature to directly represent the perfect degree of living facilities. The value of this feature will be assigned as "perfect" if a set of basic living facilities (e.g. hospital, bus stop and so on) are all within $1km$ of the given neighborhood. Otherwise, it is assigned "poor". The list of basic living facilities is also shown in the Appendix of (Xiao et al. 2020). We collect the high-risk and low-risk neighborhoods data in Wuhan city in China which is officially announced by the local government. Figure 1(a) presents the ratio distribution of high-risk and low-risk neighborhoods grouping by this "perfect-poor" facility label in Wuhan data. As we can see from Figure 1(a), for the neighborhoods with feature value as "perfect", the ratio of low-risk neighborhoods and high-risk ones is $0.57 : 0.44$; whereas the ratio of them for "poor" ones is $0.43 : 0.56$. It indicates that more high-risk neighborhoods have poor living facilities, while the low-risk neighborhoods are just the opposite.



(a) (b)

Figure 1: Features of living facilities and population density visual analysis.

3.2 Demographic Features

Next, we present the demographic features of a residential neighborhood. At first, given that the COVID-19 is easy to transmit in a person-to-person way (Liu et al. 2020c), it is necessary to take into account population density for infection risks prediction. As Figure 1(b) illustrates, on average high-risk neighborhoods do have a higher population density than low-risk neighborhoods in Wuhan city. We also compute average commute distance as a feature for each neighborhood since residents with long commutes have high infection risks.

Moreover, different groups of residents may face different risk levels in a neighborhood. For example, old people and children are easier to be infected. And residents with higher educational levels may pay more attention to scientific prevention. Hence, we construct 11 features based on the distribution of residents according to different human attributes. We present each of these features as a vector of histogram statistics of residents' distribution. The full list of such attributes is provided in the Appendix of (Xiao et al. 2020).

3.3 Transportation-Related Features

We also extract features of transportation-related behaviors from human mobility data to help predict infection risks. There have been some studies to prove that transportation-related behaviors have a close relationship with COVID-19 contagion spreading (Huang et al. 2020b). The transportation-related behaviors typically are recognized as the origin-transportation-destination (OTD) information(Xu et al. 2020; Xu et al. 2016; Liu et al. 2020a). Thus, we consider detailed features from the perspectives of T (transportation), OD (origin & destination venues) and OTD (origin-transportation-destination pattern). All the features are extracted from the search and transportation data of Baidu Maps in a certain time period. Previous studies have shown that map search behavior is a leading indicator and predictor for crowd dynamics (Zhou, Pei, and Wu 2018).

A vector in which the value of each element equals the corresponding ratio of transportation means, mainly including walk, bicycle, public transit and private vehicle, is used to depict the "T feature" for a residential neighborhood. The

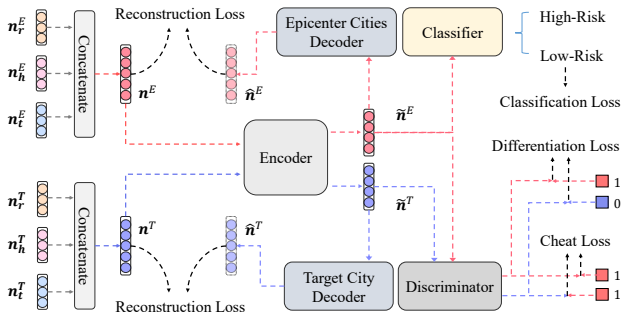


Figure 2: Illustration of cross-city transfer learning model of C-Watcher.

"OD feature" consists of types of visit venues and the distance between origin and destination venues. We classify the destination venues according to their types (e.g. hospital, restaurant, hotel, and school) and compute the proportion of each type. We also extract origin-destination distance and categorize it into different distance buckets. The proportions of different distance buckets for the neighborhood are also formed as a feature vector. Moreover, since the OTD (origin-transportation-destination) patterns most directly reflect human mobility, we collect the top-20 hottest travel patterns from all the cities in our dataset, which is a triplet tuple composed of the type of origin venue (residential area), means of transportation and type of destination venue. The histogram distribution of these top-20 OTD travel patterns of each neighborhood is treated as "OTD" features. More details about the transportation-related features can be found in the Appendix of (Xiao et al. 2020).

4 Cross-City Transfer Learning

In this section, we present the cross-city transfer learning model, which is a core component of C-Watcher, to improve the performance of early detection of high-risk neighborhoods via transferring the knowledge about COVID-19 infection from the epicenter to the target city. Usually, discrepancies always exist between different cities. Thus, we intend to learn city-invariant knowledge applicable to both epicenters and target cities, instead of those characteristics unique to epicenter cities.

4.1 Overview

An overview of our proposed cross-city transfer learning model with four components is given in Figure 2. The first component is a neural network encoder used to learn the representation of a neighborhood on the basis of three groups of features introduced in Section 3. However, the discrepancy of input distributions in different cities may lead to the gap between encoded representations of epicenter cities and target cities, which may severely disrupt the detection ability in target cities. Thus, we adopt adversarial learning by adding a discriminator component to identify whether the output of the encoder belongs to the target city or not. In addition, to ensure that the embedded feature of the encoder still keeps the ability to depict the residential neighborhoods, we exert two decoders which recover features of epicenter cities and target city respectively from the output of the encoder. Moreover, to achieve the prediction goal of C-Watcher, a classifier is also added to optimize the learned representations space and make it more related to the prediction of COVID-19 infection risks.

In the following sections, we present our cross-city transfer learning model in detail by introducing how each component works. Here comes the notation of model input first. For both residential neighborhoods in epicenter cities and target cities, we have a feature vector consist of three groups:

$$\begin{aligned} \boldsymbol{n}^E &= cat(\boldsymbol{n}_r^E, \boldsymbol{n}_h^E, \boldsymbol{n}_t^E) \\ \boldsymbol{n}^T &= cat(\boldsymbol{n}_r^T, \boldsymbol{n}_h^T, \boldsymbol{n}_t^T) \end{aligned} \tag{2}$$

where $\boldsymbol{n}^E$ denotes all of the features of a residential neighborhood in epicenter cities. $\boldsymbol{n}_r^E$, $\boldsymbol{n}_h^E$ and $\boldsymbol{n}_t^E$ respectively denote the POI radius features, demographic features and transportation-related features of that residential neighborhood in epicenter cities. $\boldsymbol{n}^T$, $\boldsymbol{n}_r^T$, $\boldsymbol{n}_h^T$ and $\boldsymbol{n}_t^T$ similarly denote the corresponding features of residential ones in target city. The function $cat(\cdot)$ is the concatenating operation.

4.2 City-Invariant Representation Learning

Since learning unique characteristics of neighborhoods in epicenter cities brings little benefit for early risk detection in target cities, we propose a city-invariant representation learning method, which is inspired by the multi-mode adversarial representation learning methods (Mai, Hu, and Xing 2020; Makhzani et al. 2015). Here the encoder is a transformer of data distribution. Given the input vectors $\boldsymbol{n}^E$ and $\boldsymbol{n}^T$, we use $\widetilde{\boldsymbol{n}}^E$ and $\widetilde{\boldsymbol{n}}^T$ to denote the outputs of the encoder. Similar to (Mai, Hu, and Xing 2020) and (Makhzani et al. 2015), the distributions transformation from the inputs to encoded representations can be presented as:

$$\begin{aligned} p(\widetilde{\boldsymbol{n}}^E, \Phi_e) &= \int_{\boldsymbol{n}^E} e(\widetilde{\boldsymbol{n}}^E | \, \boldsymbol{n}^E, \Phi_e)\, p(\boldsymbol{n}^E)\, d\boldsymbol{n}^E \\ p(\widetilde{\boldsymbol{n}}^T, \Phi_e) &= \int_{\boldsymbol{n}^T} e(\widetilde{\boldsymbol{n}}^T | \, \boldsymbol{n}^T, \Phi_e)\, p(\boldsymbol{n}^T)\, d\boldsymbol{n}^T \end{aligned} \tag{3}$$

where $p(\cdot)$ denotes the data distribution and $e(\cdot\,, \Phi_e)$ represents the encoding distribution. Φ_e are parameters of the encoder, determining the projection space where the distributions of input data $p(\boldsymbol{n}^E)$ and $p(\boldsymbol{n}^T)$ are transformed into that of encoded representations $p(\widetilde{\boldsymbol{n}}^E, \Phi_e)$ and $p(\widetilde{\boldsymbol{n}}^T, \Phi_e)$.

In general, $p(\widetilde{\boldsymbol{n}}^E, \Phi_e)$ and $p(\widetilde{\boldsymbol{n}}^T, \Phi_e)$ are different distributions characterizing different cities. It means that some features helpful in predicting COVID-19 infection risk may be unique to neighborhoods in epicenter cities unless we impose constraints on the encoder. To this end, we use adversarial learning to narrow the discrepancies between distributions $p(\widetilde{\boldsymbol{n}}^E, \Phi_e)$ and $p(\widetilde{\boldsymbol{n}}^T, \Phi_e)$ by adding a discriminator to distinguish whether the neighborhood comes from epicenter cities or the target city. In this way, the discriminator needs to do a binary classification task, in which it takes the encoded representations as inputs and aims to identify the inputs $\widetilde{\boldsymbol{n}}^E$ from epicenter cities as true but the inputs $\widetilde{\boldsymbol{n}}^T$ from

target cities as false, while the encoder tries its best to confuse the discriminator to classify both of them as true. We can formulate the function of discriminator as:

$$\begin{aligned} D(\widetilde{\boldsymbol{n}}^E, \Phi_D) &\rightarrow 1 \\ D(\widetilde{\boldsymbol{n}}^T, \Phi_D) &\rightarrow 0 \end{aligned} \tag{4}$$

where $D(\cdot, \Phi_D)$ denotes the function of the discriminator which can be an MLP model that outputs the probability from 0 to 1. On the contrary, the encoder competes against the discriminator by:

$$\begin{aligned} D(\widetilde{\boldsymbol{n}}^E, \Phi_D) &\rightarrow 1 \\ D(\widetilde{\boldsymbol{n}}^T, \Phi_D) &\rightarrow 1 \end{aligned} \tag{5}$$

For this adversarial learning procedure, we use binary cross entropy (BCE) to define the loss function:

$$\mathcal{L}_{al} = \mathcal{L}_{diff}(\widetilde{\boldsymbol{n}}^E, \widetilde{\boldsymbol{n}}^T) + \mathcal{L}_{ch}(\widetilde{\boldsymbol{n}}^E, \widetilde{\boldsymbol{n}}^T) \tag{6}$$

$$\mathcal{L}_{diff} = -[log(D(\widetilde{\boldsymbol{n}}^E)) + log(1 - D(\widetilde{\boldsymbol{n}}^T)] \tag{7}$$

$$\mathcal{L}_{ch} = -[log(D(\widetilde{\boldsymbol{n}}^E)) + log(D(\widetilde{\boldsymbol{n}}^T)] \tag{8}$$

where $D(\widetilde{\boldsymbol{n}}^E)$ is used to represent $D(\widetilde{\boldsymbol{n}}^E, \Phi_D)$ in simplicity and so does $D(\widetilde{\boldsymbol{n}}^T)$. The differentiation loss $\mathcal{L}_{diff}$ guides discriminator to predict $\widetilde{\boldsymbol{n}}^E$ as true (epicenter cities) but $\widetilde{\boldsymbol{n}}^T$ as false (target city), while the encoder tries to learn features that are common between epicenter cities and target city to hinder discriminator from distinguishing successfully, under the effects of cheat loss $\mathcal{L}_{ch}$. The adversarial procedure will finally reach an equilibrium situation where the discriminator could no longer distinguish whether the encoded representations come from epicenter cities or target city, then the encoder is able to extract "city-invariant" features from raw inputs $\boldsymbol{n}^E$ and $\boldsymbol{n}^T$. In this case, discrepancies between cities decrease and the experience which helps predict infection risks in epicenter cities can make more sense in target cities.

4.3 Embedding Space Constraints

A problem about city-invariant representation learning is that, if no regulations and restrictions are imposed on the embedding space of the encoder, the encoded representations of epicenter cities and target cities may only be similar in distribution but fail to retain useful information for identifying high-risk neighborhoods. We solve this problem with multi-task learning strategy by additionally exerting an auto encoder-decoder features reconstruction component, as well as a COVID-19 infection risks prediction component.

The reconstruction component consists of two decoders (one for residential neighborhoods in epicenter cities and another one for residential neighborhoods in the target city) which take the encoded representations as inputs. The decoding operation can also be considered as a distribution transformation like encoding:

$$\begin{aligned} p(\widehat{\boldsymbol{n}}^E, \Phi_d^E) &= \int_{\widetilde{\boldsymbol{n}}^E} d^E(\widehat{\boldsymbol{n}}^E | \, \widetilde{\boldsymbol{n}}^E, \Phi_d^E)\, p(\widetilde{\boldsymbol{n}}^E)\, d\widetilde{\boldsymbol{n}}^E \\ p(\widehat{\boldsymbol{n}}^T, \Phi_d^T) &= \int_{\widetilde{\boldsymbol{n}}^T} d^T(\widehat{\boldsymbol{n}}^T | \, \widetilde{\boldsymbol{n}}^T, \Phi_d^T)\, p(\widetilde{\boldsymbol{n}}^T)\, d\widetilde{\boldsymbol{n}}^T \end{aligned} \tag{9}$$

where $\widehat{\boldsymbol{n}}^E$, $\widehat{\boldsymbol{n}}^T$ denote the reconstructed outputs of decoders from $\widetilde{\boldsymbol{n}}^E$ and $\widetilde{\boldsymbol{n}}^T$ respectively, and $d^E(\widehat{\boldsymbol{n}}^E | \, \widetilde{\boldsymbol{n}}^E, \Phi_d^E)$ represents the epicenter cities decoder function with parameters Φ_d^E, while $d^T(\widehat{\boldsymbol{n}}^T | \, \widetilde{\boldsymbol{n}}^T, \Phi_d^T)$ is similar but for the target city. Aiming to approximate decoded representations to the original inputs ($\widehat{\boldsymbol{n}}^E \rightarrow \boldsymbol{n}^E$ and $\widehat{\boldsymbol{n}}^T \rightarrow \boldsymbol{n}^T$), we use mean square error to define reconstruction loss function:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec}^E + \mathcal{L}_{rec}^T \tag{10}$$

$$= \| \, \widehat{\boldsymbol{n}}^E, \boldsymbol{n}^E \|_2 + \| \, \widehat{\boldsymbol{n}}^T, \boldsymbol{n}^T \|_2 \tag{11}$$

Optimized by the reconstruction loss above, the encoder-decoder framework ensures that the embedding space is still characterizing a residential neighborhood.

Moreover, considering that our ultimate objective is to detect latent high-risk neighborhoods, we add a classifier to identify COVID-19 infection risks in epicenter cities upon the learned encoded representations. The classification problem can be defined as :

$$C(\widetilde{\boldsymbol{n}}^E, \Phi_c) \rightarrow y^E, \; y^E \in \{0, 1\} \tag{12}$$

where $C(\cdot, \Phi_c)$ denotes the function of MLP classifier with parameter Φ_c. This is also a binary classification task and we use BCE to define the classification loss function:

$$\mathcal{L}_{cl} = -y^E log(C(\widetilde{\boldsymbol{n}}^E, \Phi_c)) - (1 - y^E) log(1 - C(\widetilde{\boldsymbol{n}}^E, \Phi_c)) \tag{13}$$

The classification loss transmits the known information carried by label y^E to encoder and classifier, which is COVID-19 infection risks of neighborhoods in epicenter cities. It achieves the goals to restrict the encoded representations to be instructive in high-risk neighborhood identification.

All in all, loss functions generated from all the three components of discriminator, decoders and classifier will act on the encoder and optimize the embedding space in our proposed cross-city transfer model. The total loss function can be expressed qualitatively as:

$$\mathcal{L} = \lambda_{diff}\mathcal{L}_{diff} + \lambda_{ch}\mathcal{L}_{ch} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cl}\mathcal{L}_{cl} \tag{14}$$

In model training, the adversarial model is optimized in an alternate mode. We use differentiation loss $\mathcal{L}_{diff}$ to optimize the discriminator first to improve its discriminatory ability to neighborhoods in both epicenter cities and that target city, which also leads to the rise of cheat loss. Then we apply cheat loss $\mathcal{L}_{ch}$ combined with $\mathcal{L}_{rec}$ and $\mathcal{L}_{cl}$, to guide the encoder to optimize its parameters in a direction where demands to learn city-invariant, informative and risk-discriminative features are all taken into consideration. Together with the encoder, the decoders and classifier update themselves based on $\mathcal{L}_{rec}$ and $\mathcal{L}_{cl}$, respectively.

Reference City Validation Mechanism. Another problem of C-Watcher is how to select the best hyperparameters to train the model. Here we build a reference city validation mechanism to tune hyperparameters. The illustrated diagram is shown in Figure 3. Reference city in our paper can be epicenter cities, and can also be some cities with COVID-19 outbreak but not so serious as epicenters. We train the C-Watcher model on epicenter cities set, and use ground truth data of the reference city as validation data to choose the

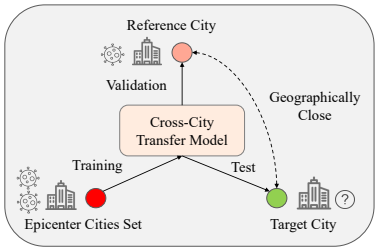


Figure 3: Diagram of reference city validation mechanism.

hyperparameters. Then we evaluate the early detection performance in target cities geographically close to that reference city. In this case, we ensure that our trained model to detect latent high-risk neighborhoods in a target city with best hyperparameters, without any prior information related to COVID-19 confirmed cases and spreading trend.

5 Experiments

5.1 Datasets and Settings

Dataset construction. The constructed datasets simulate a common outbreak pattern in a country. In the scenarios, there is a set of epicenter cities in the country (like Wuhan in China), and a few reference cities (see Section 4.3) which have some confirmed cases. The C-Watcher can be trained on the epicenter cities and reference cities datasets, and then be used to make early detection of high-risk residential neighborhoods in the rest of cities in the country.[1]

All datasets are built based on 16 cities in China which consists of one dataset from epicenter city, 5 evaluation datasets from selected reference cities, and 10 test datasets from other cities. The epicenter city dataset is constructed based on Wuhan, which has the largest number of confirmed cases in China and is well-recognized as the epicenter of the COVID-19 outbreak in China. The five selected reference cities, Shenzhen, Changsha, Chengdu, Shanghai and Zhengzhou, are key cities in their provinces and they are also evenly situated in different geographical regions of China. For each reference city, we also construct two test datasets from two cities geographically closed to them. The full list of the test cities is in the Appendix of (Xiao et al. 2020). The POI data and user profile data of all the cities are both collected by the first week of March 2020. The human mobility data are collected from January 1, 2020 to March 3, 2020.

We also make a great effort to build the ground-truth dataset. For the Wuhan dataset, we manually collected all the high-risk residential neighborhoods (released on February 24, 2020) and low-risk residential neighborhoods (released on March 6, 2020) which are officially published by the local government. After data cleaning and feature alignment, there are 336 high-risk neighborhoods and 715 low-risk neighborhoods. The statistics of high-risk neighborhoods in other cities are listed in Appendix of (Xiao et al. 2020). For datasets of other cities, we label the neighborhoods with at least one confirmed case as high-risk while others as low-risk, based on the public COVID-19 patients dataset by (Fu et al. 2020).

In order to tune hyperparameters for baselines, we split Wuhan dataset into three folds as train, validation and test data by a 0.7:0.15:0.15 ratio. The hyperparameters tuning for C-Watcher is done by reference city validation mechanism (see Section 4.3).

Baselines. Since we are the first to study the COVID-19 high-risk neighborhoods early detection problem, there is no direct competitor of C-Watcher. Thus, we compare C-Watcher with classical machine learning methods of Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), XGBoost (XGB) and Lasso Logistic Regression (Lasso-R). We use the dataset from epicenter city to train the baselines, and make a prediction on the datasets of test cities.

Metrics. Since the detection of high-risk neighborhoods is an imbalance binary classification task (high-risk neighborhoods are much less than low-risk ones), we mainly evaluate the performance by AUC (Area under the ROC Curve), which reflects model performance within different discrimination thresholds (Manning, Schütze, and Raghavan 2008; Fu et al. 2020). In addition, we also calculate the p-value by pairwise t-test between baselines and C-Watcher to show the statistical significance of the evaluation results.

Optimization and hyperparameters tuning. We optimize C-Watcher by Adam optimizer. The main hyperparameters of C-Watcher, including weights of the loss function (λ_{ch}, λ_{diff}, λ_{rec} and λ_{cl}), learning rate and hidden size of the neural network of each component are determined by grid search method, with batch size fixed as 64.

5.2 Performance Evaluation of Early Detection

We evaluate the performance of C-Watcher and its baselines for early detection of high-risk neighborhoods on test datasets of the 10 target cities. In Table 1, the overall column shows the average AUC of the 10 cities. We can see that C-Watcher can improve the AUC by 8.18% over the best baseline (SVM and MLP). We also conduct pairwise t-test between the C-Watcher and each baseline. The p-values in Table 1 demonstrate that C-Watcher can achieve significantly better performance than other baselines.

We also show the prediction performance on test datasets of five target cities in Table 1. Each target city corresponds to one reference city. We can see that the improvement by C-Watcher over baselines in different cities is different. For example, the improvement by C-Watcher over the best baseline on Shaoyang is 14.57% (i.e., C-Watcher (0.6433) vs. SVM (0.5615)); but the one by C-Watcher over the best baseline on Xuchang is about 0% (the AUC of C-Watcher is almost the same with other baselines). A possible reason for such different performances is that some geographically closed cities are not similar, thus the reference city cannot help to select the best hyperparameters for transfer learning. We leave this problem as a further research investigation. We put the prediction performance of all the ten cities in the Appendix of (Xiao et al. 2020).

[1]The code can be found at https://github.com/PaddlePaddle/Research/tree/master/ST_DM/AAAI2021-CWatcher/.

	Overall		Huizhou	Shaoyang	Lianyungang	Xuchang	Chongqing
	AUC	P-value	AUC				
SVM	0.5999	0.0005	0.7049	0.5615	0.6728	**0.7330**	0.5693
XGB	0.5810	0.0018	0.6266	0.5190	0.6182	0.7067	0.4901
Lasso-R	0.5853	0.0006	0.6364	0.5410	0.6515	0.7195	0.5718
MLP	0.5963	0.0005	0.6995	0.5594	0.6850	0.7278	0.5438
C-Watcher	**0.6490**	–	**0.7352**	**0.6433**	**0.7218**	0.7312	**0.6142**

Table 1: Early detection performance comparison between C-Watcher and baselines on cross-city datasets. The "target city - reference city" relationship are "Huizhou - Shenzhen", "Shaoyang - Changsha", "Lianyungang - Shanghai", "Xuchang - Zhengzhou" and " Chongqing - Chengdu".

5.3 Feature Importance

Here we conduct a feature importance analysis to discuss possible characteristics of neighborhoods leading to the high risk for infection. We use Lasso Logistic Regression (Lasso-R) on epicenter Wuhan dataset to select the top-20 important features according to the absolute coefficient value, which is illustrated in Figure 4. The full name of each feature is listed in the Appendix of (Xiao et al. 2020). The feature importance analysis reveals several insightful and interesting points for preventing the epidemic. For POI radius features, except the effect of perfect and poor living facility of a neighborhood (which are denoted by "P:PFLF" and "P:PRLF" in Figure 4), the coefficient of "P:RTS" indicates that the long distance to a train station can reduce the risk of the neighborhood. For the demographic features, except the high population density (denoted by "D:PD"), the long average commute distance (denoted by "D:ACD") also increases the risk of the neighborhood. For the transportation-related features, we find that the percentage of travelling on walk (denoted by "T:TW") can reduce the risk of the neighborhood by a large margin. We believe such analysis can help us identify factors for high-risk neighborhoods, and provide insightful suggestions on preventing the epidemic of COVID-19 in future.

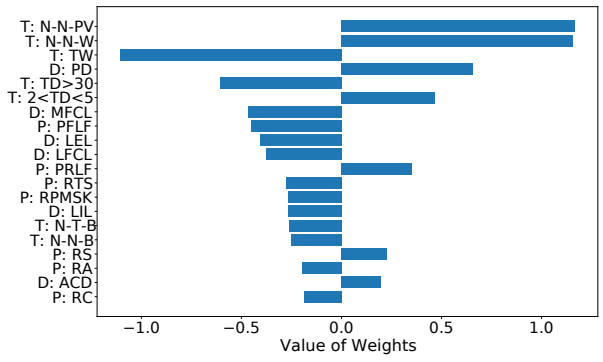


Figure 4: The top-20 most important features for high-risk neighborhoods detection.

5.4 Effectiveness of Feature Groups

In this section, we verify the effectiveness of 3 groups of hand-crafted features. In specific, we separately evaluate the performance of each group of features in detecting high/low-risk neighborhoods, then we compare them with the performance of taking all 3 groups of features together as inputs. All the comparative experiments for feature effectiveness are conducted by MLP on the epicenter Wuhan dataset. As we can see from Table 2, all the three groups of features can positively classify the high-risk and low-risk neighborhoods. More importantly, the combination of these three groups certainly improves the model's overall performance, which proves complementary among the three groups of features.

Feature groups	AUC
POI Radius	0.8033
Demographic	0.7579
Transportation-Related	0.7414
All three Groups	**0.8458**

Table 2: Detection performance comparison of MLP with different feature groups on Wuhan dataset.

6 Conclusion

In this paper, we study the problem of predicting infection risks of COVID-19 in urban neighborhoods. We first construct a set of features incorporating human mobility data to characterize the demographic/socioeconomic status and spatial interactions of a residential neighborhood, then propose C-Watcher, a data-driven framework based on these features to early detect high-risk neighborhoods in a city ahead of local COVID-19 outbreaks. To improve infection risks identification in target cities, C-Watcher adopts adversarial learning algorithms that learn "city-invariant" features to boost generalizing knowledge witted in epicenter and build a reference city validation mechanism for hyperparameters selection. We conduct extensive experiments upon real-world data in the early stage of COVID-19 outbreaks from China to demonstrate the advantages of C-Watcher to early detect high-risk neighborhoods across cities and analyze the importance and effectiveness of explored features.

Acknowledgments

We thank all reviewers for insightful comments. This work is supported in part by National Key R&D Program of China (No. 2018YFB1402600), and in part by grant from the National Natural Science Foundation of China (Grant No. 91746301 and No. 71531001). Part of experiments in this paper was carried out using anonymous data and secure data analytics provided by Baidu Data Federation Platform (Baidu FedCube). For data accesses and usages, check http://fedcube.baidu.com/page_en.html.

References

Balcan, D.; Colizza, V.; Gonçalves, B.; Hu, H.; Ramasco, J. J.; and Vespignani, A. 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* 106(51): 21484–21489.

Borjas, G. J. 2020. Demographic determinants of testing incidence and COVID-19 infections in New York City neighborhoods. Technical report, National Bureau of Economic Research.

Fu, Z.; Wu, Y.; Zhang, H.; Hu, Y.; Zhao, D.; and Yan, R. 2020. Be Aware of the Hot Zone: A Warning System of Hazard Area Prediction to Intervene Novel Coronavirus COVID-19 Outbreak. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2241–2250.

Ghamizi, S.; Rwemalika, R.; Cordy, M.; Veiber, L.; Bissyandé, T. F.; Papadakis, M.; Klein, J.; and Le Traon, Y. 2020. Data-driven Simulation and Optimization for Covid-19 Exit Strategies. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3434–3442.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems(NIPS)*, 2672–2680.

Hao, Q.; Chen, L.; Xu, F.; and Li, Y. 2020. Understanding the Urban Pandemic Spreading of COVID-19 with Real World Mobility Data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3485–3492.

Hu, R.; Lu, X.; Liu, C.; Li, Y.; Liu, H.; Gu, J.; Ma, S.; and Xiong, H. 2020. Why we go where we go: Profiling user decisions on choosing POIs. In *roceedings of the 29th International Joint Conference on Artificial Intelligence*, 3459–3465.

Huang, J.; Wang, H.; Fan, M.; Zhuo, A.; and Li, Y. 2020a. Personalized Prefix Embedding for POI Auto-Completion in the Search Engine of Baidu Maps. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2677–2685.

Huang, J.; Wang, H.; Fan, M.; Zhuo, A.; Sun, Y.; and Li, Y. 2020b. Understanding the Impact of the COVID-19 Pandemic on Transportation-Related Behaviors with Human Mobility Data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3443–3450.

Huang, J.; Wang, H.; Xiong, H.; Fan, M.; Zhuo, A.; Li, Y.; and Dou, D. 2020c. Quantifying the economic impact of COVID-19 in Mainland China using human mobility data. *arXiv preprint arXiv:2005.03010* .

Jiang, P.; Fu, X.; Van Fan, Y.; Klemeš, J. J.; Chen, P.; Ma, S.; and Zhang, W. 2020. Spatial-temporal potential exposure risk analytics and urban sustainability impacts related to COVID-19 mitigation: A perspective from car mobility behaviour. *Journal of Cleaner Production* 123673.

Li, S.; Zhou, J.; Xu, T.; Liu, H.; Lu, X.; and Xiong, H. 2020. Competitive Analysis for Points of Interest. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1265–1274.

Liu, H.; Li, Y.; Fu, Y.; Mei, H.; Zhou, J.; Ma, X.; and Xiong, H. 2020a. Polestar: An Intelligent, Efficient and National-Wide Public Transportation Routing Engine. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2321–2329.

Liu, J.; Wang, X.; Xiong, H.; Huang, J.; Huang, S.; An, H.; Dou, D.; and Wang, H. 2020b. An Investigation of Containment Measures Against the COVID-19 Pandemic in Mainland China. *arXiv preprint arXiv:2007.08254* .

Liu, Y.; Ning, Z.; Chen, Y.; Guo, M.; Liu, Y.; Gali, N. K.; Sun, L.; Duan, Y.; Cai, J.; Westerdahl, D.; et al. 2020c. Aerodynamic analysis of SARS-CoV-2 in two Wuhan hospitals. *Nature* 582(7813): 557–560.

Mai, S.; Hu, H.; and Xing, S. 2020. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 164–172.

Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* .

Manning, C. D.; Schütze, H.; and Raghavan, P. 2008. *Introduction to information retrieval*. Cambridge University Press.

Peng, Y.; and Qi, J. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 15(1): 1–24.

Renso, C.; Baglioni, M.; de Macedo, J. A. F.; Trasarti, R.; and Wachowicz, M. 2013. How you move reveals who you are: understanding human behavior by analyzing trajectory data. *Knowledge and information systems* 37(2): 331–362.

Vollmer, M.; Mishra, S.; Juliette, H.; et al. 2020. Using mobility to estimate the transmission intensity of COVID-19 in Italy: a subnational analysis with future scenarios. Imperial College London. 2020.

Xiao, C.; Zhou, J.; Huang, J.; Zhuo, A.; Liu, J.; Xiong, H.; and Dou, D. 2020. C-Watcher: A Framework for Early Detection of High-Risk Neighborhoods Ahead of COVID-19 Outbreak. *arXiv preprint arXiv:2012.12169* .

Xiong, H.; Liu, J.; Huang, J.; Huang, S.; An, H.; Kang, Q.; Li, Y.; Dou, D.; and Wang, H. 2020. Understanding the collective responses of populations to the COVID-19 pandemic in Mainland China. *medRxiv* .

Xu, T.; Zhu, H.; Xiong, H.; Zhong, H.; and Chen, E. 2020. Exploring the Social Learning of Taxi Drivers in Latent Vehicle-to-Vehicle Networks. *IEEE Transactions on Mobile Computing* 19(8): 1804–1817.

Xu, T.; Zhu, H.; Zhao, X.; Liu, Q.; Zhong, H.; Chen, E.; and Xiong, H. 2016. Taxi driving behavior analysis in latent vehicle-to-vehicle networks: A social influence perspective. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1285–1294.

Xu, X.; He, L.; Lu, H.; Gao, L.; and Ji, Y. 2019. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web Journal* 22(2): 657–672.

Yuan, Z.; Liu, H.; Liu, Y.; Zhang, D.; Yi, F.; Zhu, N.; and Xiong, H. 2020. Spatio-Temporal Dual Graph Attention Network for Query-POI Matching. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 629–638.

Zhou, J.; Pei, H.; and Wu, H. 2018. Early warning of human crowds based on query data from baidu maps: Analysis based on shanghai stampede. In *Big data support of urban planning and management*, 19–41. Springer.