Transfer Learning-Based Outdoor Position Recovery with Cellular Data

Zhang, Yige; Ding, Aaron Yi; Ott, Jorg; Yuan, Mingxuan; Zeng, Jia; Zhang, Kun; Rao, Weixiong

DOI
10.1109/TMC.2020.2968899

Publication date
2021

Document Version
Final published version

Published in
IEEE Transactions on Mobile Computing

Transfer Learning-Based Outdoor Position Recovery With Cellular Data

Yige Zhang, Aaron Yi Ding, Jörg Ott, Mingxuan Yuan, Jia Zeng, *Senior Member, IEEE*, Kun Zhang, and Weixiong Rao, *Member, IEEE*

Abstract—Telecommunication (Telco) outdoor position recovery aims to localize outdoor mobile devices by leveraging measurement report (MR) data. Unfortunately, Telco position recovery requires sufficient amount of MR samples across different areas and suffers from high data collection cost. For an area with scarce MR samples, it is hard to achieve good accuracy. In this paper, by leveraging the recently developed transfer learning techniques, we design a novel Telco position recovery framework, called TLoc, to transfer good models in the carefully selected source domains (those fine-grained small subareas) to a target one which originally suffers from poor localization accuracy. Specifically, TLoc introduces three dedicated components: 1) a new coordinate space to divide an area of interest into smaller domains, 2) a similarity measurement to select best source domains, and 3) an adaptation of an existing transfer learning approach. To the best of our knowledge, TLoc is the first framework that demonstrates the efficacy of applying transfer learning in the Telco outdoor position recovery. To exemplify, on the 2G GSM and 4G LTE MR datasets in Shanghai, TLoc outperforms a non-transfer approach by 27.58 and 26.12 percent less median errors, and further leads to 47.77 and 49.22 percent less median errors than a recent fingerprinting approach NBL.

Index Terms—Cellular data, outdoor position, transfer learning, data driven approach

1 Introduction

Recent years we have witnessed the ever-growing size and complexity of telecommunication (Telco) networks to process 1000-fold growth in the amount of traffic and 100-fold increase in the number of users [20]. Telco operators have to manage heterogeneous networks (including 2G-4G and upcoming 5G networks), composed of macro cells, small cells, and distributed antenna systems. The growing demands and heterogeneous networks require an automated approach to network control and management, instead of error-prone manual network management and parameter configuration. To enable the automated network control and management, the *outdoor locations* of mobile devices are important for Telco operators to *1*) pinpoint location hotspots for capacity planning, *2*) identify gaps in radio frequency spatial coverage, and *3*) locate users in emergency situations (E911) [20]. Moreover, the locations of mobile devices are widely used to understand mobility patterns [36] and optimize many third-party applications such as urban planning and traffic forecasting [6].

Outdoor locations of mobile devices can be recovered from cellular Measurement record (MR) data [9]. MR samples are generated when mobile devices make phone calls and access data services. MR samples contain connection states (e.g., signal strength) between mobile devices and connected base stations. After the locations of mobile devices are recovered, we tag the MR samples by the associated geo-locations, generating the so-called geo-tagged MR samples.

In literature, various position recovery algorithms via cellular MR samples have been developed. Google MyLocation [2] approximates outdoor locations by the positions of cellular towers connected with mobile devices. This method suffers from median errors of hundreds and even thousands of meters. More recently, data-driven approaches have attracted intensive research interests in both academia and Telco industry [4], [14], [17], [29], [34], [40]. These approaches leverage geo-tagged MR samples to build the mapping from MR samples to associated locations, and the mapping is then used to localize the mobile devices in non-geo-tagged samples. For example, the fingerprinting approach [14] builds a histogram of MR signal strength (i.e., fingerprint database) for each divided grid cell in the areas of interest, and the Random Forest (RaF)-based approach [40] maintains the mapping function between MR features (i.e., MR signal strength) and position labels. When enough amount of training geo-tagged MR samples are used, the data-driven algorithms achieve the median error of $20 \sim 80$ meters [13], [40].

A key concern of the data-driven methods mentioned above is requiring sufficient geo-tagged MR samples to build the accurate mapping from MR samples to associated locations. Nevertheless, collecting sufficient geo-tagged MR samples across the distributed areas of an urban city incurs rather high cost. It is not rare that an area of interest suffers from insufficient geo-tagged MR samples. If we have scarce geo-tagged MR samples for such an area, the position recovery precision in that area could be very low. For example in

- *Y. Zhang and W. Rao are with the Tongji University, Shanghai 200092, China. E-mail: {1610832, wxrao}@tongji.edu.cn.*
- *A.Y. Ding is with the Department of Engineering Systems and Services, TU Delft, 2628 CD Delft, Netherlands. E-mail: Aaron.Ding@tudelft.nl.*
- *J. Ott is with the Faculty of Informatics, Technical University of Munich, 80333 München, Germany. E-mail: ott@in.tum.de.*
- *M. Yuan and J. Zeng are with the Noahs Ark Lab, Huawei, Hong Kong. E-mail: {mingxuan.yuan, jia.zeng}@huawei.com.*
- *K. Zhang is with the Philosophy Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA. E-mail: kunz1@andrew.cmu.edu.*

Manuscript received 1 Aug. 2019; revised 7 Dec. 2019; accepted 10 Jan. 2020. Date of publication 23 Jan. 2020; date of current version 2 Apr. 2021. (Corresponding author: Weixong Rao.)
Digital Object Identifier no. 10.1109/TMC.2020.2968899

a recent work NBL [20], though over 100 TB GPS-tagged cellular signal data in an American city are collected by 4 million users from Jan 2016 to July 2016, the median localization errors in rural areas are still as high as around 750 meters due to insufficient samples.

In this paper, targeting Telco operators, we design a transfer learning-based cellular position recovery approach, called TLoc, to accurately localize mobile devices in those areas with scarce data samples. The general idea of TLoc is as follows. First, we divide an entire area into fine-grained small subareas, namely *domains*. For each domain, we then maintain the mapping from MR samples within this domain to their associated positions. Next for the target domains suffering from low precision, we transfer good mappings from appropriately selected source domains to target ones via transfer learning. In this way, we greatly improve the localization accuracy in target domains. Though transfer learning has been used for indoor WiFi-based localization [22], [37], [38], we believe that indoor WiFi and outdoor cellular localization differs significantly. Thus, those indoor localization approaches are not expected to perform well in our case (they will be evaluated in our experiment) due to the following challenges. *First*, given the cellular outdoor localization, designing a proper position coordinate space is the prerequisite to enable knowledge transfer across two domains. This, unfortunately, cannot be achieved by using outdoor GPS longitude and latitude coordinates: the different GPS position (i.e., position label) for every area makes it impossible to share position labels across distributed domains, and hence hard to perform knowledge transfer. *Second*, given a large number of domains, it is challenging to select the best source domains for a target one. In contrast, due to the small area and rather limited domains in an indoor environment, it is straightforward for indoor localization to select source domains. Thus, trivial effort on source domain selection is employed for indoor WiFi-based localization [22], [37], [38].

To tackle the challenges above, TLoc builds the following components. First, unlike absolute GPS coordinates, we use a *relative coordinate space* for position recovery. Under this coordinate space, the mobile devices even in two distributed domains can still share the same relative positions, facilitating the transfer across two domains. Second, based on the relative position space, we design an effective distance metric to measure the similarity between domains. The metric incorporates the distribution of the signal strength of MR samples, relative position information, and non-serving base station deployment information. Finally, by adapting an existing structured transfer learning (STL) method [26], we build a Random Forest (RaF)-based position recovery model for each domain and then perform model transfer from appropriately chosen source domains to target ones. As a summary, this paper makes the following contributions.

- To the best of our knowledge, TLoc is the first method to plausibly leverage transfer learning for cellular outdoor localization. Unlike the fingerprinting-based and machine learning-based approaches [13], [17], [40], TLoc mitigates high efforts to collect a large quantity of training samples across an entire area. Moreover, our evaluation empirically verifies that the idea of TLoc can generally benefit other approaches (e.g., fingerprinting-based approaches) to achieve better precision by re-using MR samples from source domains to target ones.
- We design a novel approach to divide an entire urban area into small domains by the proposed relative coordinate space. Based on the divided domains, we define a distance metric for measuring domain similarity to select appropriate source domains effectively for a target one. By adapting a recent structured transfer learning (STL) scheme [26] for a RaF regression model [40], TLoc leads to much better position recovery precision than those non-transfer models.
- Our extensive evaluation validates that TLoc greatly outperforms both state-of-the-arts and the variants of TLoc. For example, on two 2G GSM and 4G LTE MR datasets, TLoc outperforms the recent fingerprinting approach NBL [20] by 47.77 and 49.22 percent less median errors, respectively, and leads to 27.58 and 26.12 percent less median error when compared with the non-transfer RaF algorithm [40], respectively.

TABLE 1
Mainly Used Short Names/Symbols and Notations

Notation/Symbol	Meaning
Telco	Telecommunication
MR	Measurement Report
TL	Transfer Learning
STL	Structure Transfer Learning
SVR	Supported Vector Regression
RSSI	Received Signal Strength Indicator
RaF	Random Forest
D, s	Domain (divided small areas), MR Sample
$\mathbf{S}_T$ and $\mathbf{S}_S$	Target and Source Data Set
$F_d()$, $F_i()$	MR Features dependent (resp. independent) upon locations
$L()$	Recovered Location
dis_{mr}^{rssi}, dis_{mr}^{sig}	Weighted Histogram Distance of RSSI (resp. SignalLevel)
dis_{mr}	Overall MR Feature Distance
dis_{pos}	Relative Position Distance
$dist(D, D')$	Domain Distance between two domains D and D'

The rest of this paper is organized as follows. Section 2 reviews the background and related work. Section 3 gives the general idea of TLoc and the proposed relative coordinate space. After that, Section 4 defines the distance metric to measure domain similarity, and Section 5 adapts the STL model [26] for TLoc. Section 6 evaluates TLoc and Section 7 finally concludes the paper. Table 1 summarizes the main acronyms and notations used in the paper.

2 Background and Related Work

In this section, first we give the background of Measurement Report (MR), Random Forests (RaF), and transfer learning, and second review the literature in terms of outdoor position recovery and selection of source domains.

Measurement Report (MR) Data. MR samples are used to record the connection states between mobile devices and nearby base stations in a cellular network. Table 2 gives an

TABLE 2
A 2G GSM MR Sample Collected by an Android Device

MRTime xxx	IMSI xxx	RNCID 6188	BestCellID 26050	NumOfBS 7
$RNCID_1$ 6188	$CellID_1$ 26050	$AsuLevel_1$ 18	$SignalLevel_1$ 4	$RSSI_1$ -77
$RNCID_2$ 6188	$CellID_2$ 27394	$AsuLevel_2$ 16	$SignalLevel_2$ 4	$RSSI_2$ -81
$RNCID_3$ 6188	$CellID_3$ 27377	$AsuLevel_3$ 15	$SignalLevel_3$ 4	$RSSI_3$ -83
$RNCID_4$ 6188	$CellID_4$ 27378	$AsuLevel_4$ 15	$SignalLevel_4$ 4	$RSSI_4$ -83
$RNCID_5$ 6182	$CellID_5$ 41139	$AsuLevel_5$ 16	$SignalLevel_5$ 4	$RSSI_5$ -89
$RNCID_6$ 6188	$CellID_6$ 27393	$AsuLevel_6$ 9	$SignalLevel_6$ 3	$RSSI_6$ -95
$RNCID_7$ 6182	$CellID_7$ 26051	$AsuLevel_7$ 9	$SignalLevel_7$ 3	$RSSI_7$ -95

example of 2G GSM MR samples collected by an Android device. It contains a unique number (IMSI: International Mobile Subscriber Identity), connection time stamp (MRTime), up to 7 nearby base stations (identified by RNCID and CellID) [25], signal measurements such as AsuLevel and SignalLevel, and a radio signal strength indicator (RSSI). SignalLevel indicates the power ratio (typically logarithm value) of the output signal of the device and the input signal. AsuLevel, i.e., Arbitrary Strength Unit (ASU), is an integer value proportional to the received signal strength measured by the mobile phone. Among the up-to 7 base stations, one of them is selected as the primary serving base station to provide communication and data transmission services for the mobile device. Previous work on cellular localization [13], [40] might ignore the use of serving base station. Unlike these works, we will carefully exploit serving base stations as the base of TLoc.

Besides 2G GSM MR samples, we also collect 4G LTE MR samples by frontend Android devices. They both follow the same data format. Nevertheless, due to the limitation of Android API, frontend Android devices cannot acquire the identifiers (RNCID_2 $\sim$ 7 and CellID_2 $\sim$ 7) of non-serving base stations from 4G LTE networks. Nevertheless, the signal measurements associated with the missed base stations can be still collected.

Finally, Telco operators can collect MR samples via backend base stations except the frontend MR samples above by Android mobile phones. Nevertheless, their data formats are different [13]. First, besides RSSI, the backend 4G MR samples provided by Telco operators contain RSRP and RSRQ which do not appear in the frontend 4G MR samples. Second, backend 2G MR samples contain RxLev, the received signal strength on Absolute Radio Frequency Channel Number (ARFCN) [13]. The previous work [12] shows that RxLev is exactly equal to the RSSI value, and we thus treat RxLev equally as RSSI. Now, to make sure that we have proper knowledge transfer between frontend and backend MR datasets, we perform knowledge transfer only for those MR feature items (e.g., RSSI) that appear within all datasets. For example, we transfer the knowledge from the RSSI (or RxLev) items in backend 2G MR samples to the RSSI items in frontend 2G samples. Yet, we do not transfer knowledge for such MR features as RSRP and RSRQ.

Random Forest (RaF) is an ensemble method for classification, regression, and other learning tasks. It constructs a multitude of decision trees (DTs) [1] during the training phase and outputs either the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. RaF avoids the overfitting of DTs to their training set. Specifically, DTs that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, i.e., low bias but very high variance. RaFs are a way of averaging multiple deep DTs, trained on different parts of the same training set, with the goal of reducing the variance. This greatly boosts the performance in the final model, at the expense of a small increase in the bias and loss of interpretability.

Transfer learning aims at improving the learning in a new task through proper transfer of knowledge from a related task that has already been learned. Those machine learning algorithms such as RaFs are designed to address a *single* task. In contrast, transfer learning attempts to leverage individual tasks by developing methods to transfer knowledge learned in one or more source tasks to a related target task. Transfer learning is frequently used due to expensive cost or impossibility to re-collect the needed training data and rebuild the models. Transfer learning approaches include *Model Transfer*, *Instance Transfer* (or data sample transfer), *Features Transfer*, and *Relational knowledge-Transfer*. We refer interested readers to the detailed survey of transfer learning [23].

TLoc mainly utilizes a recent model transfer scheme, i.e., the structure transfer learning (STL) [26] in decision tree (DT)-based model to transfer knowledge from multiple source domains to the target one. Specifically, DTs for similar problems (in various domains) exhibit a certain extent of structural similarity. However, the scale of the features used to construct RaF and their associated decision thresholds are likely to differ from various problems. Thus, the DTs trained on source domains are adapted to the target one by discarding all numeric threshold values in the original DTs and working top-down, and then selecting a new threshold for a node with a numeric feature using the subset of target examples that reach this node.

Recall that general transfer learning frameworks (such as instance transfer) require training examples from source domains for domain adaptation. Instead, the STL scheme can directly adapt the already trained models from source domains to target ones. This unique property is particularly useful for the scenario that cannot directly leverages source examples for domain adaption, for whatever reason, e.g., storage capacity or data privacy. Thus, TLoc can comfortably adapt a given source model to a target domain relying on a relatively small training set from the target. The experimental results show that multi-source transfer in STL has the better precision than the single-source transfer.

Outdoor Position Recovery. In the literature, cellular outdoor position recovery techniques are broadly classified into two categories: *1)* measurement-based methods [18], *2)* data-driven methods. Measurement-based methods frequently

adopt absolute point-to-point distance estimation or angle estimation from cellular signals to calculate mobile device locations. Examples of measurement-based techniques include Angle of arrival (AOA), Time of Arrival (TOA) [8], and Received signal strength (RSS)-based single source localization [30]. Nevertheless, information related to AOA and TOA is highly error prone in cellular systems, and measurement-based techniques suffer from high localization errors, typically with the median error of hundreds of meters [8], [24]. In addition, as shown in previous work [24], 4G LTE MR samples typically have signal strength from at most two cells, namely, the serving cell and the strongest neighboring cell. Triangulation-based localization approaches thus do not perform well because they require signal strength from at least three cells.

Fingerprinting-based and machine learning-based algorithms generally belong to the data-driven methods. They both leverage collected historical data samples for outdoor position recovery. *Fingerprinting methods* [14], [17] have been reported to have better performance than measurement-based approaches. For example, in the offline survey phase, the classic work, CellSense [14], first divides the area of interest into smaller cells and constructs a fingerprint database, e.g., a vector histogram of RSSI on each cell. When given a query (i.e., an input RSSI feature), the online prediction phase then searches the fingerprint database to find the location that has the maximum probability given the received signal strength vector in the query. An average of the k most probable fingerprint cells, weighted by the probability of each location, can be used to obtain a better estimate of locations. In addition, a better CellSense-hybrid technique consists of two phases: the rough estimation phase first uses the standard probabilistic fingerprinting technique to obtain the most probable cell in which a user may be located, and the estimation refinement phase then uses a k-nearest neighbor approach to estimate the closest fingerprint point, in the signal strength space, to the current user location inside the cell estimated in phase one.

AT&T researchers recently studied the fingerprinting-based outdoor localization problems [20], [24]. In particular, the authors in NBL [20] extended CellSense [14] similarly using two stages. In an offline stage, NBL developed radio frequency coverage maps based on a large-scale crowdsourced channel measurement campaign. Then, in an online stage, a localization algorithm quickly matches the input radio frequency measurements to coverage maps. By assuming a Gaussian distribution of signal strength within each divided grid, NBL maintains the mean value and standard deviation of signal strength of each neighboring cell tower for the samples in the grid. The online stage computes the predicted location by using either Maximum Likelihood Estimation (MLE) or Weighted Average (WA). The median errors in the 4G LTE network reported by NBL are around 80 and 750 meters in urban and rural areas, respectively.

Machine learning approaches leverage machine learning models such as Random Forests, Support Vector Regression (SVR), Gradient Boosting Decision Tree (GBDT), and artificial neural network (ANN) to build the mapping from MR features (which are extracted from MR samples and engineering parameter data of connected base stations) to device positions [13], [40]. When given an MR record without position information, machine learning models then predict the associated location. As shown in [40], the authors proposed a context-aware coarse-to-fine regression (CCR) model (implemented by a two-layer RaFs). The CCR model takes as input 258 dimensional coarse features and 34 dimensional fine-grained contextual feature vectors. Thus, beyond strength indicators frequently used by fingerprinting approaches, those context-aware and coarse-to-fine features such as moving speed enable CCR to outperform the classic fingerprinting approaches with slightly 14 percent lower median errors. In a very recent deep learning-based outdoor cellular localization system, namely DeepLoc [27], a data augmentor is used to handle data noise issue and to provide more training samples. With help of the samples, a deep learning model is trained for better localization result.

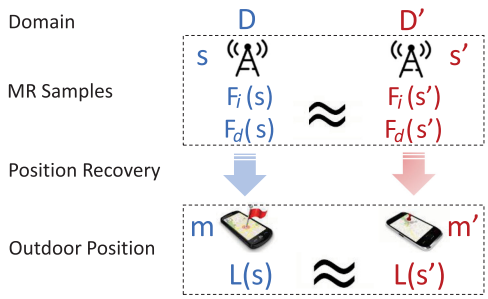


Fig. 1. General idea of TLoc.

Source Domain Selection. Given a number of diverse source domains, to successfully perform knowledge transfer, one may need to select a certain number of source domains that bear essential similarity to the considered target domain. Some previous works in transfer learning studied the general source domain selection problem. For example, an information theoretic framework was developed [3] to rank source convolution neural networks (CNNs) and select the top-k CNNs for the target learning task by understanding the source-target relationship. A restricted boltzman machine (RBM) was also used [5] to select source domains in the context of reinforcement learning. Some works instead did not take domain selection into account and focused on instance selection from available source domains [19]. In addition, many existing transfer learning methods suppose that source domains are provided in advance by default.

Compared to the above methods, TLoc gives a meaningful distance metric to determine the domain similarities for cellular position recovery task. Unlike TLoc, the previous work [3] focuses on the selection of pre-trained CNN models which can be intuitively treated as learning tasks, and [19] selects source instances.

3 System Design

3.1 General Idea

We first give the general idea of TLoc to perform model transfer across different domains. In Fig. 1, we consider that two mobile devices m and m' in two distributed domains D and D' (with $D \neq D'$) generate the MR samples s and s', respectively. Suppose that we are using a RaF regression model to recover the outdoor locations $L(s)$ and $L(s')$ for the samples s and s', respectively. The outdoor locations are frequently represented by GPS coordinates [13], [20], [24],

[40]. Given the two distributed domains $D \neq D'$, the MR samples s and s' within the two domains indicate that the corresponding *RNC/CellID* and GPS positions are different, indicating $s \neq s'$ and $L(s) \neq L(s')$.

Inside MR samples, we note that there exist two types of features: 1) those ID-alike features $F_d()$ dependent upon located domains such as *RNCID* and *CellID*, and 2) those numeric features $F_i()$ independent upon located domains such as *AsuLevel*, *SignalLevel* and *RSSI*. Due to the distributed domains D and D', $F_d(s) \neq F_d(s')$ holds. Nevertheless, when the two samples s and s' contain very similar cellular signal strength (including *AsuLevel*, *SignalLevel* and *RSSI*), it is highly possible to have $F_i(s) \simeq F_i(s')$. The similar cellular signal strength gives us a hint: we would like to modify the representation of the features $F_d()$ and locations $L()$ to ensure that $F_d(s) \simeq F_d(s')$ and $L(s) \simeq L(s')$ hold. When both $F_i(s) \simeq F_i(s')$ and $F_d(s) \simeq F_d(s')$ hold, we then could have the similar MR samples $s \simeq s'$ and the roughly equal positions $L(s) \simeq L(s')$. Based on this representation, we next perform knowledge transfer across two similar MR samples $s \in D \simeq s' \in D'$ as follows: if $s \simeq s'$ holds, we estimate the position $L(s) \leftarrow L(s')$ via the position $L(s')$. In general, we extend the idea of TLoc from similar samples to similar domains. Gvien the two similar domains $D \simeq D'$, we infer $s \in D \simeq s' \in D'$, and then estimate $L(s) \leftarrow L(s')$ via the available position $L(s')$.

3.2 Relative Coordinate Space

To perform the knowledge transfer above, we introduce a relative coordinate space to represent $L()$ and $F_d()$, such that $F_i(s) \simeq F_i(s')$ and $F_d(s) \simeq F_d(s')$ hold for two samples s and s' within two distributed domains D and D'.

1) *Representation of* $L()$: We first represent $L()$ by transforming original GPS coordinates to relative coordinates as follows. For the MR samples having a certain base station as their serving stations, the mobile devices generating such MR samples are highly possible to be located around the serving base station. Thus, based on serving base stations, we divide a large urban area of interest (e.g., either a university campus or an entire city) into fine-grained *small subareas* (or equivalently we use the term *domains* D that are frequently used in the transfer learning community). That is, based on serving base stations in MR samples, the MR samples having the same serving base stations belong to the same domains. For every domain, we design a *relative coordinate space* for all MR samples within the domain. We use Fig. 2 as an example to represent the relative coordinates. In this figure, we assume that those MR samples belonging to the same domain D (a.k.a having the same serving base station BS) are all within a circle and BS is the center. The radius R of this circle is equal to the maximal distance between the positions of BS and MR samples. Given this center BS in the coordination space, we convert the original GPS coordinate (x_0, y_0) of BS into a relative one $(0, 0)$. For a MR sample $s \in D$ with the GPS coordinate $(x + x_0, y + y_0)$, its relative coordinate becomes (x, y). In this way, we can compute the relative coordinates of all MR samples by referring BS as the center of this coordination space.

Until now, we show the key point of the relative coordination space as follows. Let us consider another domain D', where the GPS position of the serving base station (i.e., the center) belonging to D' is (x_1, y_1). For one MR sample $s' \in D'$ with the GPS coordinate $(x + x_1, y + y_1)$, its relative coordinate becomes (x, y). Here, though the two samples $s \in D$ and $s' \in D'$ are originally with different GPS coordinates $(x + x_0, y + y_0)$ and $(x + x_1, y + y_1)$, they now share the exactly same relative coordinates (x, y) under their own domains. In this way, we can perform the transfer from D to D'. That is, when both cellular signal strength in MR samples (a.k.a MR features) and relative position (labels) refer to serving base stations, the transfer across domains becomes possible. Moreover, for a large amount of MR samples across an entire area, we can group the MR samples by their serving base stations to build the associated domains and relative coordination space. After the big area is divided into small domains, for each domain (and the associated serving base station), we can learn an individual mapping from MR samples within this domain to their relative positions. The key is that the mapping is adaptively learned by the data-driven fashion, even if the transmitting power, cellular signal coverage, and bandwidth of serving base stations are unavailable. Thus, even for two base stations (with various transmitting power, cellular signal coverage, and bandwidth) located at the exactly same locations, we could establish two corresponding mappings from the MR samples generated by an individual base station to the associated MR positions.

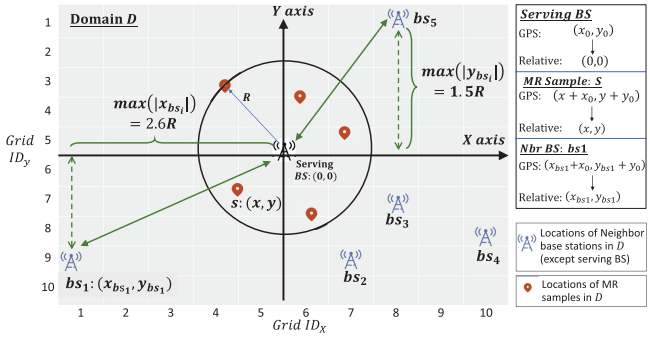


Fig. 2. Relative coordinate space.

Note that the relative coordinate space above requires the GPS coordinate of serving base stations. Telco operators can easily obtain the GPS coordinates of base stations because base stations are deployed by Telco operators themselves.

2) *Representation of* $F_d()$: For a certain MR sample s, we convert MR features $F_d(s)$, such as *RNCID* and *CellID* of a neighboring base station, into meaningful IDs which are independent upon the associated domains. Specifically, depending upon all the neighbouring base stations appearing inside the MR samples within a domain, we determine a rectangle area covered by these neighboring base stations. As shown in Fig. 2, the width (resp. height) of the rectangular is equal to $2 \times \max(|x_{bs_i}|)$ (resp. $2 \times \max(|y_{bs_i}|)$), where we have $\max(|x_{bs_i}|) = x_{bs_1}$ and $\max(|y_{bs_i}|) = y_{bs_5}$. Then, we evenly divide the rectangle into $g \times g$ small grids (we have $g = 10$ in this figure). In this way, each neighboring base station is located within a certain grid and we replace its *RNCID* and *CellID* by the associated grid IDs $Grid\_ID_x$ and $Grid\_ID_y$. For example, we represent the two base stations bs_1 and bs_5 by the grid IDs $(1, 9)$ and $(8, 1)$, respectively. The representation of $F_d(s)$ above offers the following

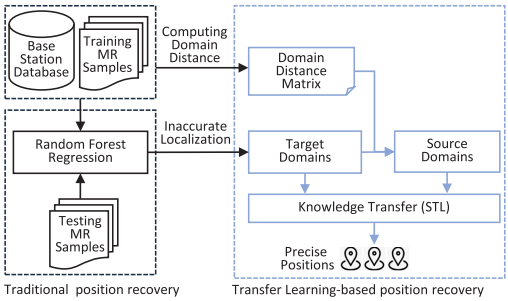


Fig. 3. System overview.

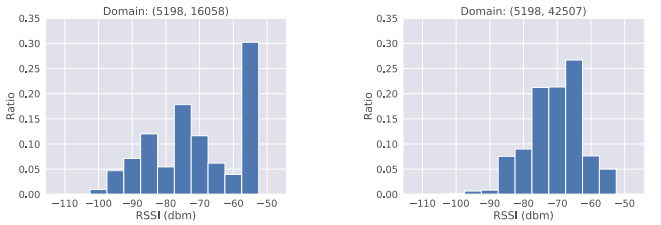


Fig. 4. Histograms of two domains in terms of RSSI from serving base station (where 5198 indicates RNC ID and 16058/42507 indicates cell ID).

advantage: the grid IDs are now independent upon domains and $F_d(s) \simeq F_d(s')$ holds for two MR samples $s \in D$ and $s' \in D'$.

3.3 System Overview

Following the general idea above, we introduce three following components of TLoc (see Fig. 3): a traditional position recovery model (e.g., a Random Forest-based regression model), a matrix to maintain the pairwise domain similarity, and the transfer learning component for those target domains suffering from inaccurate position accuracy (caused by insufficient position labels for MR training samples).

Let us consider the following scenario in a big area, where the geo-tagged MR samples are distributed unevenly across this area. To this end, we follow Section 3.2 to divide the entire big area into multiple smaller areas (a.k.a *domains*) and represent MR samples and associated positions under the associated relative coordinate space. Among the divided domains, due to the uneven distribution of geo-tagged MR samples, some of the domains could be with sufficient samples, and a regression-based position recovery model thus works very well. Yet other domains may contain scarce geo-tagged MR samples, the trained position recovery model usually suffers from poor localization accuracy [13].

To this end, TLoc adapts the recent transfer learning scheme STL [26] to improve the localization accuracy in the domains suffering from poor prediction precision, e.g., with a median error higher than a given threshold. We treat such domains as *target domains*. Based on the developed distance metric (Section 4), we choose those top-k domains that *1)* are most similar to a target domain and *2)* are with low localization errors. Such top-k domains are called the *source domains* of the target one. Finally, we *transfer* the recovery models from the top-k source domains to the target one using an adapted STL technique (Section 5).

4 Domain Distance

Since the position recovery model essentially maintains the mapping from MR features, e.g., $F_i(s)$ and $F_d(s)$, to MR positions $L(s)$, we thus define the domain distance by two parts: *1)* the distance in terms of MR features and *2)* the distance in terms of MR positions $L(s)$. In this section, we first give the detail for each of the two parts and next give the domain distance by integrating the two parts.

4.1 MR Feature Distance

To measure the similarity of MR features between two domains, the distance metric takes into account three following aspects: *1)* the general approach to compute the distance of those MR features $F_i(s)$ involving cellular signal strength, *2)* the distance by introducing the weight of up to seven base stations, and *3)* the overall distance involving the refinement of three specific signal strengths (*RSSI*, *AsuLevel* and *SignalLevel*).

4.1.1 Distance of Cellular Signal Strength

First, to compute the distance of cellular signal strength between two domains, we exploit a histogram structure to capture the overall distribution of cellular signal strength in a certain domain, and next compute the histogram distance. Fig. 4 plots the histograms of two example domains to capture the distribution of RSSI from serving base stations. The x-axis is the RSSI value and y-axis indicates the ratio of the MR samples having RSSI values falling inside a RSSI interval against total MR samples in the domain. To compute the histogram distance, we choose three frequently used metrics: probabilistic likelihood [14], [33], Kullback-Leibler Divergence [21], and p-norm distance [31]. Among the three metrics, we empirically find that the p-norm distance with $p = 3$ leads to the best result. Formally, for a domain D (resp. D'), we denote by $h_{D,j}$ (resp. $h_{D',j}$) the MR sample rate in y-axis for the jth RSSI interval in x-axis. When each histogram contains r RSSI intervals, we compute the histogram distance between D and D' by

$$dis_{hist}(D, D') = \left(\sum_{j=1}^{r} (|h_{D,j} - h_{D',j}|)^p \right)^{\frac{1}{p}}. \quad (1)$$

4.1.2 Weighted Distance of Cellular Signal Strength

We note that each MR sample contains up to seven base stations sorted by descending order of cellular signal strength. These stations contribute differently to the distance in Equation (1), due to various signal strength caused by these stations.

Specifically, each domain D contains a set of MR samples. For all (neighboring) base stations appearing in these MR samples, we group such stations by their order index in MR samples: the 1st group contains only one serving base station with the strongest signal strength, the 2nd group contains the neighboring base stations of 2nd order index (i.e., RNCID_2, CellID_2) in each MR sample. In this way, we have up to seven groups of base stations. Each group contains a list of base stations, denoted by l_i with $i = 1, \ldots, 7$. In this way, we improve Equation (1) by introducing a weight w_i for each l_i

$$dis_{mr}(D, D') = \sum_{i=1}^{7} w_i \times dis^i_{hist}(D, D'). \quad (2)$$

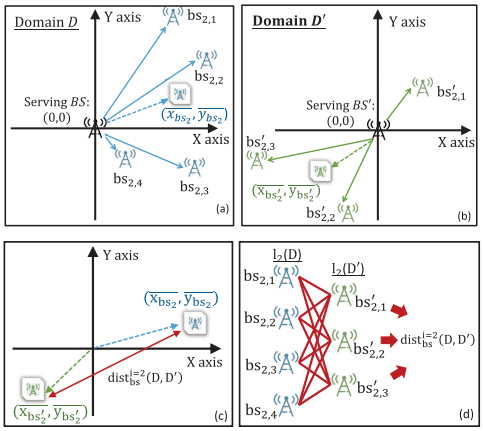


Fig. 5. Distance of BS locations between domains.

In Equation (2), $dis^i_{hist}(D, D')$, computed by Equation (1), is the histogram distance for the MR signal associated with the ith lists $l_i(D)$ and $l_i(D')$ in two domains D and D', and w_i is the weight of the ith group.

We give the general idea of computing the weight w_i as follows. Recall that Section 3.2 transforms the neighboring base stations (identified by *RNCID* and *CellID*) into grid IDs. Such grid IDs approximate the positions of neighboring stations within each domain: the nearest (resp. farthest) base stations contribute to the strongest (resp. weakest) cellular signal strength. We leverage these grid IDs to compute the weight w_i. As shown in Fig. 5, we have the 2nd base station lists in two domains D and D', denoted by $l_2(D)$ and $l_2(D')$, which contain 4 member stations $bs_{2,1} \dots bs_{2,4}$ and 3 stations $bs'_{2,1} \dots bs'_{2,3}$, respectively. Based on the distance $dis^{i=2}_{bs}(D, D')$ of the two lists $l_2(D)$ and $l_2(D')$, , we define the normalized weight w_i as follows:

$$w_i = \frac{e^{dis^i_{bs}}}{\sum_{j=1}^{7} e^{dis^j_{bs}}}. \tag{3}$$

To compute the item dis^i_{bs} above, as shown in Fig. 5c, we exploit the average position of the 4 stations in $l_2(D)$, denoted by $(\overline{x_{bs_i}}, \overline{y_{bs_i}})$, and one of the 3 stations in $l_2(D')$, denoted by $(\overline{x_{bs'_i}}, \overline{y_{bs'_i}})$. After that, we compute the euclidean distance between the two average positions

$$dis^i_{bs}(D, D') = [(\overline{x_{bs_i}} - \overline{x_{bs'_i}})^2 + (\overline{y_{bs_i}} - \overline{y_{bs'_i}})^2]^{\frac{1}{2}}. \tag{4}$$

Nevertheless, the average positions above might lose the geographical characteristics of base stations. Thus, as an improvement to compute the item dis^i_{bs}, as shown in Fig. 5d, we first calculate the pairwise distance between the base stations in $l_i(D)$ and $l_i(D')$, and then compute the average of the pairwise distance

$$dis^i_{bs}(D, D') = \frac{\sum_{n=1}^{|l_i(D')|} \sum_{m=1}^{|l_i(D)|} ed(bs_{i,m}, bs'_{i,n})}{|l_i(D)||l_i(D')|}. \tag{5}$$

In Equation (5), $ed(\cdot)$ indicates the euclidean Distance of two base stations $bs_{i,m}$ and $bs'_{i,n}$, whose positions are represented by grid IDs under the relative coordinate space.

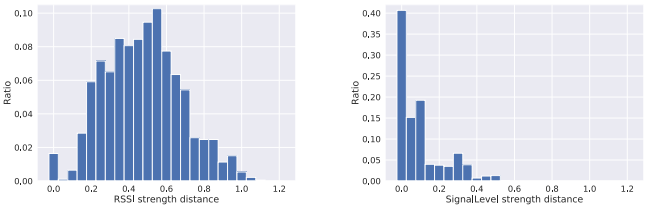


Fig. 6. Distribution of RSSI (left) and SignalLevel (right) histogram distance between pairwise domains.

Note that the 4G LTE MR samples collected by Android API may miss the IDs of neighbour base stations (see Section 2), and we cannot leverage the positions of base stations, required by Equation (5), to compute the weight w_i. To overcome this issue, we could approximate $w_i = \frac{1/i}{\sum_{j=1}^{7} 1/j}$, such that w_i is inverse to the index number i. This approximation makes sense: the index number i essentially indicates the signal strength of base stations, and the weight w_1 regarding to the 1st index (i.e., the serving base station) consequently contributes most to the overall distance.

4.1.3 Overall Distance of MR Features

Third, recall that MR samples in Table 2 contain three types of signal strength: *RSSI*, *AsuLevel* and *SignalLevel*. For such signal strength, we might first follow Equation (2) to compute three associated histogram distance such as $dis^{rssi}_{mr}(D, D')$ and next sum the three weighted distance as the overall distance. However, the sum may not provide a sensible overall measure if the three types of cellular signal strength are heavily dependent. In fact this is the case because *AsuLevel* is a scaling value of *RSSI*, i.e., in 2G GSM data set, $AsuLevel = (RSSI + 113)/2$ [25], and dis^{asu}_{mr} can be treated as a linear transformation of dis^{rssi}_{mr}. Among the three types of signal strength, we thus take into account the independent contribution of *RSSI* and *SignalLevel* to compute the overall distance of MR features.

In terms of *SignalLevel*, we follow Equation (2) to compute its histogram distance $dis^{sig}_{mr}(D, D')$. As shown in Fig. 6, the distribution of $dis^{sig}_{mr}(D, D')$ in our datasets significantly differs from the one of $dis^{rssi}_{mr}(D, D')$: around 40 percent *SignalLevel* distance values are 0.0 and more than 85 percent (resp. 95 percent) are smaller than 0.1 (resp. 0.3). The numbers indicate that the majority of *SignalLevel* feature values in the datasets are zeros and the output and input signals are equal (see Section 2 for the meaning of *SignalLevel*). Thus, we could assign a small weight for *SignalLevel* and among the overall distance, the distance of *SignalLevel* contributes less than the one of *RSSI*. Since *RSSI* distance plays a key role in the overall distance, we use the *average Pearson coefficient* c between *RSSI* and *SignalLevel* as the weight of *SignalLevel*.

Based on the intuition above, we compute the overall distance of MR features as follows:

$$\begin{aligned} dis_{mr}(D, D') &= \frac{1 \times dis^{rssi}_{mr}(D, D') + c \times dis^{sig}_{mr}(D, D')}{1 + c} \\ dis^{rssi}_{mr} &= \sum_{i=1}^{7} w_i \times dis^i_{hist\_rssi}(D, D') \\ dis^{sig}_{mr} &= \sum_{i=1}^{7} w_i \times dis^i_{hist\_sig}(D, D'). \end{aligned} \tag{6}$$

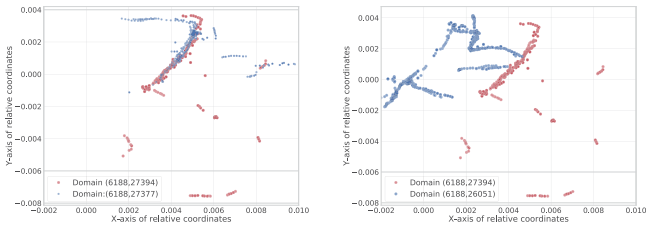


Fig. 7. Trajectory distance of the left figure is smaller than the right one, where 6188 is RNCID and 27394 is CellID.

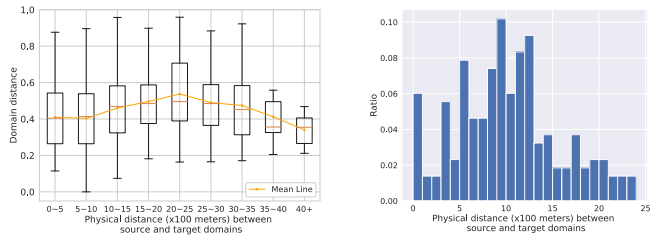


Fig. 8. Domain selection (Jiading 2G). Left: Domain distance; Right: Top-k source domains.

In the equation above, $dis_{mr}^{rssi}(D, D')$ (resp. $dis_{mr}^{sig}(D, D')$) is the weighted histogram distance between D and D' for *RSSI* (resp. *SignalLevel*) using Equation (2), and c is the average Pearson coefficient between *RSSI* and *SignalLevel*.

4.2 Relative Position Distance

Besides MR features, we also compute the distance of MR positions (labels) between two domains. Since we have represented MR positions by relative ones, we compute the distance by relative positions. In addition, instead of using discrete MR positions, we connect such positions into moving trajectories.

For one mobile device (identified by IMSI), we have a series of relative positions corresponding to the neighbouring MR samples sorted by the MR time stamp. In the case where the timestamp gap between any two neighbouring MR samples exceeds a threshold (e.g., 60 minutes), we divide the MR series into multiple short ones. A short MR series then becomes an associated moving trajectory. The trajectories are useful for understanding the overall spatio-temporal mobility patterns of mobile devices. Thus, we compute the distance of the trajectories, instead of MR positions, between two domains.

Given two trajectories T and T', we compute the Frechet distance [10]: $dis(T, T') = min[max_{t\in T, t'\in T'} dis(t, t')]$, where t and t' indicate the sample points in trajectories T and T', respectively. If an euclidean distance is used to compute $dis(t, t')$, then the sub-item $max_{t\in T, t'\in T'} dis[t, t']$ computes the *maximum distance*, and the item $min[max_{t\in T, t'\in T'} dis(t, t')]$ finds the minimal one among the maximum distance.

In addition, each domain may contain multiple trajectories. Thus, we compute the average of the sum of pairwise trajectory distance

$$dis_{pos}(D, D') = \frac{\sum_{T\in D, T'\in D'} dis(T, T')}{|D| \times |D'|}, \tag{7}$$

where $|D|$ and $|D'|$ indicate the trajectory count in domains D and D', respectively. $dis_{pos}(D, D')$ indicates the average distance between any two trajectories in D and D'. As shown in Fig. 7, the trajectory distance between two domains (6188, 27394) and (6188, 27377) is smaller than the one between two domain (6188, 27394) and (6188, 26051).

4.3 Source Domain Selection by Domain Distance

We now integrate the two distance of MR features and positions above to define the overall domain distance

$$dist(D, D') = w_{mr} \times dis_{mr} + w_{pos} \times dis_{pos}. \tag{8}$$

In the equation above, the weights w_{mr} and w_{pos} with $0 \leq w_{mr}, w_{pos} \leq 1.0$ and $w_{mr} + w_{pos} = 1.0$ measure the importance of dis_{mr} and dis_{pos}, respectively. By default we set $w_{mr} = w_{pos} = 0.5$. Our evaluation will show that such parameters can be effectively tuned according to the amount of labelled samples in source and target domains.

Given the defined metric above, we are interested in how the similar domains are also physically close. To this end, for our Jiading 2G GSM data set, we plot Fig. 8 (left) to give the average domain distance under various physical distance between domains. The $x-$axis indicates the interval of the physical distance, and the y-axis is the average domain distance within the interval. This figure indicates that two physically closer domains, e.g., the physical distance is smaller than 2.5 km, are more similar. Moreover, two domains, though rather far away, still have chance to be similar.

Next, Fig. 8 (right) plots the physical distance between top $k = 3$ source domains and a target one, where $x-$axis is the interval of physical distance between source and target domains, and $y-$axis is the rate of source domains. We find that the distance between most source and target domains is smaller than 2.5 km, consistent with Fig. 8.

From Fig. 8, we find that the needed source domains for a target one are physically close. In addition, some far-away domains are useful for a target one (In Section 6, we select source domains across different areas which leads to the best transferring results). Thus, we compute the pairwise domain distance for the top-k most similar source domains for the target ones. Nevertheless, when the count of divided domains is a large number, the pairwise domain distance involves non-trivial computing overhead. Thus, for higher efficiency, we apply the locality sensitive hash (LSH) technique [15] to approximately find the top-k source domains for a target one. Our experiment will investigate the trade-off between approximation precision and computation efficiency.

5 Structure Transfer in Random Forests

In this section, we give the detail of the proposed transfer learning framework on a Random Forest (RaF) regression model. We consider the labelled MR samples (denoted by $\mathcal{S}_T$) in a target domain and those (denoted by $\mathcal{S}_S$) in the top-k source domains. A simply way is to mix the data samples from $\mathcal{S}_T$ and $\mathcal{S}_S$, and then apply a classic RaF algorithm [7]. However, this approach cannot differentiate source domains from the target one, and thus does not work very well.

To solve the issue above, we adapt the recent structure transfer learning (STL) [26] (its general idea refers to Section 2) to solve a regression model that differs from the classification problem in the original STL work [26]. Fig. 9 gives the work

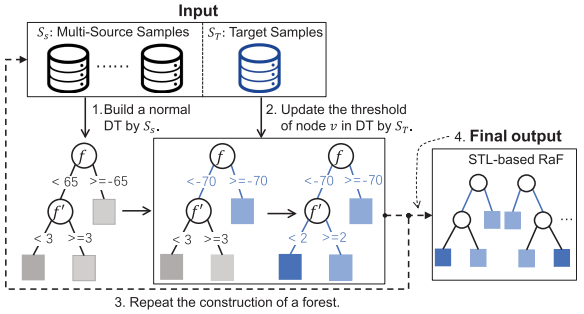


Fig. 9. Details of structure transfer in random forest.

flow of STL. The input of STL is the labelled MR samples in $\mathcal{S}_T$ (target domain) and $\mathcal{S}_S$ (multi-source domains), and the output is a transferred random forest model which is adaptive to the target domain. Specifically, we first use the data samples from $\mathcal{S}_S$ (i.e., those k selected source domains) to determine the feature f which can perform the split at each node v in a certain decision tree (DT). Next, we re-calculate the node split thresholds f_ϕ by using only the data from $\mathcal{S}_T$. For example, still in Fig. 9, the original threshold of feature v at node v is -65 computed by source samples $\mathcal{S}_S$ only. Then the threshold is optimized to -70 by the target samples $\mathcal{S}_T$. In this way, STL works top-down to select a new threshold for each node, and finally generates a random forest with transferred DTs.

Ideally, a desirable threshold yields high similarity between the distributions transferred from source domains to the target one. The purpose is that the threshold is adaptive to the target domain. Meanwhile, this similarity is restricted to "informative" thresholds where, for any sufficiently small $\epsilon > 0$, the information gain (IG) of threshold x is larger than the IG of any other $x' \in (x-\epsilon, x+\epsilon)$ in the ϵ-neighbourhood of x. It means that the thresholds are local maximums of IG. We thus formulate the threshold selection as an optimization problem

$$\begin{aligned} &\mathrm{Max}_x DG\big(Q_v^t, f, x, P_{v,left}(f), P_{v,right}(f)\big) \\ &s.t. \quad x \in \mathbb{R}, \forall x' \in (x-\epsilon, x+\epsilon): \\ &\qquad -[h(x)-mean(y)]^2 \leq -[h(x')-mean(y)]^2. \end{aligned} \tag{9}$$

In the equation above, Q_v^t denotes the samples of target task t at node v, $h(x)$ (resp. $h(x')$) is the prediction of x (resp. x'), $mean(y)$ is the mean of label y, and DG is the Jensen-Shannon divergence gain defined on Kullback-Leibler divergence and mean distribution. In addition, $P_{v,left}(f)$ and $P_{v,right}(f)$ indicate that label distribution of two subsets (left and right) split on the feature f at node v. The optimization problem in Equation (9) uses DG to quantify distributional similarity and information gain criterion computed by $-[h(x)-mean(y)]^2$ to measure a threshold's informative value. Thus, the solution of this optimization problem maximizes the defined similarity DG to make sure that the optimal decision threshold f_ϕ is adaptive enough to the target domain, thus leading to a better decision threshold f_ϕ.

6 Evaluation

6.1 Datasets and Counterparts

Datasets. In Table 3, we mainly use seven data sets collected at two cities in China: Shanghai and Urumqi. The data sets in Shanghai are sampled from three areas: 1) a university campus in the sub-urban area *Jiading*, 2) another university campus in the urban area *Siping*, and 3) several main roads in the core urban area *Xuhui*. The average physical distance of the three areas is around 15-37 km. In each of the three areas in Shanghai, we have two data sets containing MR records collected from 2G GSM and 4G LTE networks. The data sets in *Xuhui* were sampled from backend cellular towers, and the data sets in *Jiading* and *Siping* campus were collected by our developed Android mobile app via frontend Android API. In addition, to generally validate the performance of TLoc in various cities, we collect a 2G GSM MR data set by our app in *Urumqi*, where only 2G GSM cellular network is available. Since the *Urumqi* dataset contains a relatively small quantity of MR samples, we by default evaluate TLoc on the Shanghai data sets without special mention.

TABLE 3
Statistics of Used Data Sets (BSs: Base Stations)

	Jiading: 2/4G	*Siping*: 2/4G	*Xuhui*: 2/4G	*Urumqi*: 2G
# of samples	15954/10372	6723/4953	13404/7755	7645
Route Len: km	94.1/52.1	24.6/15.5	26.4/12.7	17.3
# of samples/sec	2∼3	2∼3	1	2∼3
BS density	25.85/29.43	27.16/34.67	28.18/37.12	18.31
# of serving BSs	61/44	51/42	21/16	39

For the mobile phones installed with our app, mobile users holding these mobile phones moved around the road network inside the campus. The app then collected MR samples and GPS coordinates. Specifically, when collecting MR samples from a cellular network, the mobile app switches on GPS receivers and records the current GPS coordinates. The collected GPS coordinates are used as the location ground truth. Note that the GPS coordinates collected by mobile phones may contain noise. We thus employ the data cleaning techniques including map-matching to mitigate the effect of noise [39].

Counterparts. We compare TLoc against four previous works and two variants of TLoc (see Table 4).

1) We first implement the classic fingerprinting-based approach CellSense [14] and a very recent improvement work NBL [20]. NBL assumes a prior Gaussian probability of signal strength in divided cell grids. We note that the reasonable size of cell grids in NBL involves the following trade-off: each cell grid should be great enough to contain sufficient MR samples, and yet an excessive size of the grid could alternatively

TABLE 4
Counterparts

Counterpart	Description	Source Selection
NBL [20]	Recent fingerprinting method	No transfer
CellSense [14]	Classical fingerprinting method	No transfer
DeepLoc [27]	Recent deep neural network method	No transfer
Non-Transfer [40]	Random Forest regression	No transfer
MTL [28]	Multi-task learning in Random Forest	No src selection
SVR-Transfer [37]	Transfer Learning in SVR	No src selection
TLoc	Our approach	Auto-selection

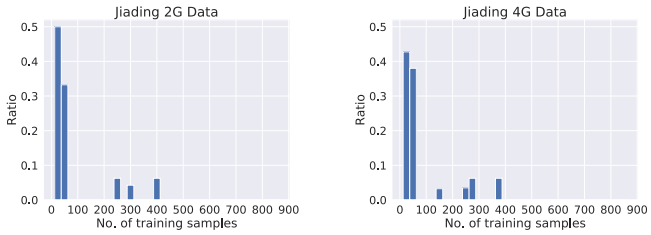


Fig. 10. Number of training samples in those domains with low-accuracy.

lead to higher localization error (because the center of a greater grid, which is used to approximate the positions of all samples within the grid, leads to a higher error).

2) The previous work CCR [40] implements a pure RaF-based regression model and has demonstrated better localization accuracy than other existing works including the classic work CellSense. Since CCR does not perform knowledge transfer and TLoc performs knowledge transfer on top of RaF-based regressor, we thus name it Non-Transfer in Table 4. In addition, we also implement a recent deep neural network-based localization approach, namely DeepLoc [27]) as one of the non-transfer learning approaches.

3) We are interested in how the adapted STL model is comparable with other transfer learning techniques. Consider that multi-task learning (MTL) is widely used in the transfer learning community [28], and Supported Vector Regression (SVR) has been used for indoor WiFi localization [37]. We thus develop two variants of TLoc by using MTL and SVR as the alternative transfer learning techniques. For the three transfer learning-based approaches (STL, MTL, and SVR), we follow TLoc to divide a big area of interest (where each MR data set was sampled) into smaller domains and perform knowledge transfer from source domains to target ones.

We tune the key parameters of the aforementioned counterparts as follows. First, according to CellSense [14] and DeepLoc [27], we carefully tune the grid size of $50 \times 50m^2$ for the best localization precision. In addition, following [37], we use the radial basis function (RBF) kernel in SVR-Transfer. Since Non-Transfer, MTL, and TLoc are all RaF-based approaches, we follow the previous works including a benchmark [13] and Non-transferr [40] to carefully tune the following parameters of RaFs: *1)* the number of trees is set to 200 (to achieve a good trade-off between accuracy and time cost), *2)* the number of used features when looking for the best split is set to $\sqrt{n_f}$ (e.g., $n_f = 44$ is the number of total features in 2G GSM MR datasets), and *3)* nodes are expanded until all leaves are pure.

Following the work [13], we adopt the following criteria to empirically determine whether or not a certain domain is treated as a target one. A domain is considered as a target domain, if *1)* the median error of this domain is greater than 30 meters for 4G LTE data or 40 meters for 2G GSM data, and *2)* the number of training samples within this domain is smaller than a threshold, $\tau = 50$. From the localization result of non-transfer CCR in each domain (we use *Jiading* datasets for illustration), we find that *1)* the number of training samples in each domain is between the interval from 22 to 864 and *2)* the localization median error is between 8.3 to 86.3 meters. Moreover, we find a strong correlation between the localization error and the quantity of training samples. That is, among those domains with median errors greater than the aforementioned thresholds (i.e., the so-called target domains with low accuracy), 85 percent of them contain 50 or even fewer (labelled) MR samples. Fig. 10 plots the distribution of the number of training samples in the domains of low accuracy. Thus, we empirically set $\tau = 50$ for target domains, such that the majority of available domains have improved localization performance by TLoc.

During the evaluation, we adopt 10 times 5-fold cross validation to choose 80 percent training and 20 percent testing data from each data set [16], and compare the prediction result of the testing data against ground truth. We compute the prediction error by the euclidean distance between prediction result and ground truth.

Table 5 lists the values of the key parameters in our experiments. We use the default values for the baseline experiments, and vary their values in some appropriate range for sensitivity study. Given the experimental settings above, we mainly evaluate TLoc to study *1)* how TLoc performs against the counterparts (Section 6.2), *2)* how TLoc is generally beneficial to various transfer learning approaches and other localization schemes (Section 6.3), *3)* how to meaningfully select source domains (Section 6.4), *4)* how to design an effective measurement of domain distance (Section 6.5), and finally *5)* how TLoc is sensitive to some key parameters such as the number of source/target MR samples (Section 6.6). After that, we visualize the localization result (Section 6.7) and give the discussion (Section 6.8).

TABLE 5
Key Parameters

Parameter	Default Values
Transfer techniques in RaF	Structure transfer
Top k source domains	$k = 3$
Localization threshold of a target domain (meters)	40 (2G)/ 30 (4G)
τ: Num. of MR samples of a target domain	50
% of used MR samples in target/source domains	80/100
Domain distance weights	$w_{mr} = w_{pos} = 0.5$

6.2 Baseline Study

We first report the position recovery errors of seven position recovery approaches. In Table 6. We show the median, mean and 90 percent errors (denoted by 50 percent, M_e, and 90 percent) in cases of using 2G and 4G network data, respectively. From Table 6, we have the following findings.

First, TLoc achieves the least errors among the seven approaches on all data sets. For example, in *Siping* 2G GSM dataset, the median error of NBL, CellSense, DeepLoc, Non-Transfer, MTL, SVR-Transfer, and TLoc algorithms is 42.8, 44.9, 35.5, 37.5, 34.3, 78.4 and 28.8 meters, respectively. Such result indicates that TLoc outperforms the Non-Transfer approach by 23.2 percent. Similar situation occurs on other data sets. Among the seven algorithms, the three RaF-based algorithms, including TLoc, MTL, and Non-Transfer, lead to better accuracy than SVR-based and fingerprint-based algorithms. Moreover, Non-Transfer,

TABLE 6
Baseline Experiment

Dataset	Jiading(2G)			Jiading(4G)			Siping(2G)			Siping(4G)			Xuhui(2G)			Xuhui(4G)			Urumqi(2G)		
	50%	Mean	90%	50%	Mean	90%	50%	Mean	90%	50%	Mean	90%	50%	Mean	90%	50%	Mean	90%	50%	Mean	90%
NBL [20]	53.8	67.4	188.8	51.8	69.3	179.5	42.8	63.0	298.3	43.2	64.9	256.7	45.9	59.0	216.8	32.2	52.4	191.6	58.3	70.2	213.6
CellCense [14]	55.4	68.7	181.1	55.6	70.6	176.4	44.9	65.7	275.4	45.8	66.3	262.6	44.7	60.2	221.3	34.9	55.5	184.3	59.7	71.4	198.7
DeepLoc [27]	37.8	47.3	175.3	37.2	48.9	184.5	35.5	44.7	219.9	38.7	49.6	267.5	31.2	40.3	210.5	27.4	39.8	180.1	44.2	62.9	175.6
No-Transfer [40]	38.8	47.6	109.8	35.6	46.5	100.9	37.5	42.8	119.5	35.8	41.4	113.7	30.0	40.2	113.4	20.0	34.1	98.3	45.2	64.3	132.3
MTL [28]	34.3	44.4	80.2	32.1	42.7	79.4	34.3	40.6	89.4	32.2	40.1	77.9	28.8	38.9	80.3	19.5	33.7	96.6	38.3	60.5	99.7
SVR-Transfer [37]	78.4	90.3	79.8	91.8	47.2	167.4	78.4	88.2	145.3	74.5	85.7	159.7	59.3	70.3	152.2	44.8	60.2	149.7	68.9	81.4	150.7
TLoc	**28.1**	**40.2**	**72.3**	**26.3**	**39.6**	**69.8**	**28.8**	**39.7**	**69.2**	**23.2**	**37.4**	**67.4**	**27.7**	**37.5**	**72.5**	**18.9**	**32.4**	**69.5**	**35.4**	**49.1**	**92.8**

i.e., the RaF-based localization approach, indicates the comparable localization accuracy to DeepLoc.

Second, the 4G LTE data sets exhibit lower errors than the 2G GSM data sets. For example, in *Siping* 4G LTE data set, TLoc has the median error of 23.20 meters, 16.88 percent lower localization error when compared to *Siping* 2G GSM dataset. By carefully checking the database of base stations, we find that the 4G LTE base stations are deployed more densely than the 2G GSM stations. In addition, *Siping* campus is located at the urban areas in Shanghai with denser deployment of base stations than *Jiading* campus in suburban areas in Shanghai. Thus, the localization errors on *Siping* data sets, including 2G GSM and 4G LTE, are smaller than those on *Jiading* data sets.

Third, in terms of the localization performance of TLoc against the two fingerprint-based methods CellSense and NBL, TLoc consistently outperforms the two fingerprint-based methods on all data sets. In addition, NBL and CellSense exhibit very similar curves on all data sets, though NBL leads to slightly lower errors than CellSense. This result is consistent with the one reported by the evaluation of NBL. Note that due to the similar curves between CellSense and NBL, in the rest of this section, we mainly choose NBL as the representative implementation of a fingerprint-based method.

Fourth, although TLoc is used to overcome the data scarcity issue, In Table 6, it is interesting to see how TLoc generally performs in diverse domains, e.g., those with sufficient MR samples (e.g., *Xuhui* dataset) and those located in a city such as *Urumqi*, which has a rather different distribution of base stations from the large urban city *Shanghai*. From the results of *Xuhui* and *Urumqi* data sets, we have two findings. First for the domains in *Xuhui*, TLoc again consistently outperforms Non-Transfer, although relatively small improvement when compared with the results in *Jiading* and *Siping* data sets. Second, for the domains in *Urumqi*, it is not surprising that the localization error for the *Urumqi* data set is much higher than that for the *Xuhui* data set, mainly due to the rather sparse deployment of base stations in *Urumqi*. Nevertheless, in the *Urumqi* data set, TLoc still leads to a significant reduction of localization errors over Non-Transfer.

Finally, in terms of the accuracy of the RaF-based transfer learning approaches, we find that TLoc outperforms MTL in all data sets. It is mainly because MTL learns the tasks for both source and target domains, and yet TLoc adaptively tunes the split thresholds on RaF nodes by the MR samples in target domains. In addition, those three RaF-based algorithms, including TLoc, MTL, and even Non-Transfer, all achieve much better accuracy than SVR-based and fingerprint-based algorithms, consistent with the benchmark [13]. The main reason is that it is hard for SVR to select an appropriate kernel function for the nonlinear feature space of MR samples. Meanwhile the hierarchical tree in RaF works very well to model the spatial structure: from a big area [40] to divided small domains.

6.3 Benefits of TLoc

Benefit to Instance-Based Transfer Learning. Beyond the model-based STL used by TLoc, we believe that the top-k source domains can offer benefits to other transfer learning techniques, e.g., instance-based transfer. To this end, based on the selected source domains, we mix the MR samples from both source and target domains to train a RaF regression model for the target domains. This approach can be intuitively treated as instance-based transfer, namely Ins-Transfer. Fig. 11a plots the results of Non-Transfer, Ins-Transfer and TLoc. Both Ins-Transfer and TLoc lead to lower errors than Non-Transfer. These results verify the benefits of using the top-k similar source domains.

Benefit to Fingerprinting-Based Localization. In this experiment, we explore the potential of applying the techniques developed for TLoc to fingerprinting-based methods, e.g., NBL [20]. Similar to TLoc, we divide the area of interest into multiple domains and perform the representation of MR features and position labels as before. Next, for the MR features and positions within each domain, we follow NBL to perform the fingerprinting-based position recovery. We name the NBL method in relative coordinate space as reNBL. Based on the reNBL, we implement the instance-based transfer, namely Tran-reNBL, by first mixing the training samples from source and target domains and then performing position recovery by reNBL. We compare NBL and the two variants reNBL and Tran-reNBL in Fig. 11b. As shown in this figure, the instance transfer in Tran-reNBL does lead to the lowest localization error among the three methods as expected.

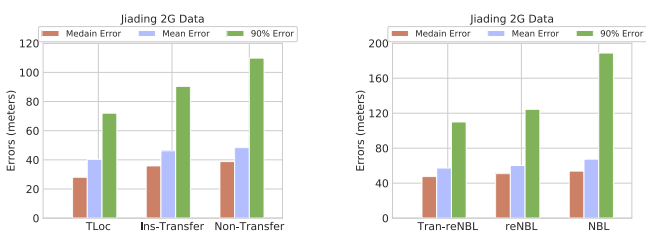


Fig. 11. Benefits of TLoc (from left to right). (a) Instance-based Transfer. (b) Fingerprinting-based localization.

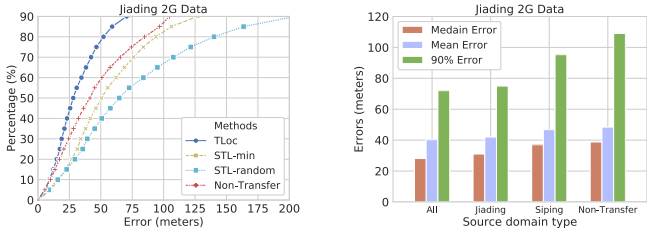


Fig. 12. Source domain selection (from left to right). (a) Four approaches. (b) Areas of source domains.

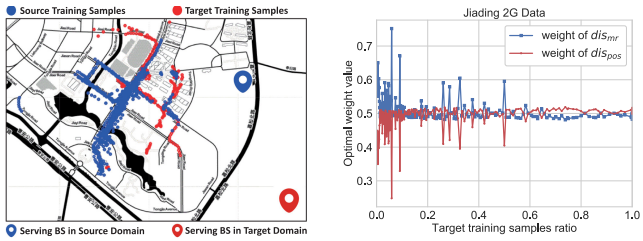


Fig. 13. From left to right: (a) Domain intersection. (b) Weight tuning.

6.4 Source Domain Selection

Domain Selection. First we compare the proposed approach of selecting the top-k most similar source domains against two alternative approaches: *1)* STL_min selects the top-k domains with the least prediction error (achieved by Non-Transfer), and *2)* STL_random randomly selects k source domains. After these source domains are selected, we adopt STL to transfer knowledge from source domains to target ones. As shown in Fig. 12a, both STL_min and STL_random even lead to higher errors than Non-Transfer. The result verifies the necessity of carefully selecting the most similar source domains. Otherwise, those dissimilar source domains, e.g., those selected by STL_min and STL_random even harm the localization accuracy of target domains.

Domain Distance. Motivated by the result above, we are further interested in the effect of selected source domains by various domain distance. In Table 7, the target domains of *Jiading* 2G data set are divided into 5 groups according to the average domain distance of Top-k (= 3) source domains. For each group, we compute the average median errors on target domains before transfer and after transfer. From this table, we have the following findings. *1)* A source domain with lower distance (a.k.a higher similarity) to a target domain leads to a more positive transfer effect with lower localization errors. It means that using similar source domains does improve the localization accuracy of target domains. *2)* When the domain distance is greater than 0.95 (though the proportion of such target domains is trivially 1.7 percent), it indicates the selected source domains are rather dissimilar to the target domain. Such source domains result in a negative transfer effect and higher localization errors, consistent with the result in Fig. 12a. Thus, we can empirically set a pre-defined threshold of domain distance, e.g., 0.95, to prune such dissimilar source domains. In this way, we can ensure that the selected source domains are truly similar to target ones and thus avoid the negative transfer effect of dissimilar source domains.

Areas of Source Domains. Third, we are interested in the areas where selected source domains are located. To this end, we purposely select source domains from 1) all three areas in Shanghai (all), 2) *Jiading* alone, and 3) *Siping* alone. In Fig. 12b, the source domains from all areas lead to the least errors, and Non-Transfer suffers from the highest errors. Specifically, for the target domains in *Jiading* 2G data set, if we select source domains from all three areas in Shanghai, we can find that 11.1 percent selected source domains are from *Xuhui*, 28.4 percent source domains are from *Siping*, and 60.5 percent source domains are from *Jiading*. These numbers indicate that most of source domains and the corresponding target domains are within the same area, but still a small number of source domains are from the two other areas. If we only select the source domains from the same area where the target ones are located, those source domains from other areas could be missed. In addition, as shown in Fig. 12b, the source and target domains within the same area *Jiading* can achieve less errors than those across areas, i.e., the target domains in *Jiading* but the source domains in *Siping*. It is because among those similar source domains for a certain target domain, most of them are within the same area, and a small number of them are from other areas, consistent with Table 7.

Source-Target Domains Within the Same Area. Differing from the experiment in Fig. 12b above, we now evaluate TLoc on the source-target domains within partially overlapping areas. Fig. 13 illustrates an example scenario for two specially chosen domains in our *Jiading* 2G dataset: the MR samples (blue dots) in a certain source domain and those (red dots) in a target domain are partially co-located within the same road segments. Given this scenario, we purposely study various approaches to select MR samples from the source domain, and evaluate the performance of TLoc. From Table 8, we find that simply selecting the source samples only from the overlapping road segments incurs the highest errors. It is mainly because the source and target samples even within the same road segments could exhibit very different signal features and relative position coordinates (because MR samples within the same road segments could be connected to various serving base stations). Instead, via the STL scheme, TLoc adapts the RaF regression model built from the source domain to the target one,

TABLE 7
Source Domain Effects of Varying Domain Distances Between Source and Target Domains: *Jiading* 2G GSM Data Set

Median Error on Target Domain (meters)	Avg. Domain Distance of Source Domains				
	<0.4	0.4-0.6	0.6-0.8	0.8-0.95	>0.95
Before Transfer	49.3	51.3	48.9	51.9	50.4
After Transfer	34.1	35.7	37.7	46.2	51.2
% of Target Domains	15.3	25.6	42.8	14.6	1.7

TABLE 8
Effect of Intersection Between Two Different Domains

Types of Source Samples	Median	Mean	90%
All samples in source	40.6	51.4	108.7
Samples in intersection area	42.4	52.5	110.4
Samples in non-intersection area	41.6	52.9	112.3
Non-Transfer	42.5	53.7	105.2
TLoc with Source Selection	33.8	45.3	94.4

TABLE 9
Trade-off Between Localization Errors and Time Cost

Source Selection Criterion	Localization Error (meters)			Avg. time per Target domain (ms)
	Median	Mean	90%	
Domain Distance	28.1	40.3	72.3	1657
LSH Approximation	32.4	47.7	90.6	362

leading to the least error. This experiment clearly indicates the advantages of TLoc over the approach that simply selects those source domains located at the same road segments as the target ones.

Trade-off Between Localization Errors and Time Cost. First, by varying the number k, we study the effect of the number k on the median error and running time of TLoc (due to space limit, this figure is not shown). The experimental result indicates that a greater number k in general leads to decreased errors, but the curve remains rather stable for $k > 3$. The errors even become slightly higher when k reaches 5. It is mainly because a greater number k indicates less similarity between source and target domains. A dissimilar source domain may lead to negative transfer effect. In terms of the time efficiency of TLoc measured by the training and prediction time, as the number k grows, more training samples are used by the model, leading to more running time. Thus, to balance the trade-off between time efficiency and model accuracy, we by default set $k = 3$.

Next, consider that TLoc requires pairwise domain distance, incurring non-trivial computing overhead. To overcome this issue, we apply the technique of Locality Sensitive Hash (LSH) [15] to efficiently approximate the domain distance. As shown in Table 9, though LSH is only an approximation approach, it can still achieve acceptable localization errors (e.g., 11.1 percent higher median error) while the time cost is greatly reduced by 4.58×.

6.5 Domain Distance

Ablation Study of Domain Distance. Recall that the domain distance is computed by integrating MR feature distance dis_{mr} and relative position distance dis_{pos}, and the MR feature distance dis_{mr} is further computed by the weighted items dis_{mr}^{rssi} and dis_{mr}^{sig}. Thus, to study the effect of each item, Table 10 first uses dis_{mr}^{rssi}, dis_{mr}^{sig}, dis_{mr}, dis_{pos} alone, and then various combinations of these items to compute domain distance for source domain selection. First, using dis_{mr}^{rssi} alone leads to lower errors than using dis_{mr}^{sig} alone, indicating that dis_{mr}^{rssi} makes a major contribution to dis_{mr}. Second, dis_{mr} leads to slightly lower errors than dis_{pos}.

TABLE 10
Ablation Study of Domain Distance: *Jiading* 2G GSM Data Set

Domain Distance	Medain	Mean	90%
dis_{mr}^{rssi}	33.6	50.9	82.7
dis_{mr}^{sig}	39.4	58.2	99.3
dis_{mr}	32.5	46.4	78.7
dis_{pos}	34.3	49.2	82.5
$0.5 * dis_{mr} + 0.5 * dis_{pos}$	28.1	40.2	72.3
$0.67 * dis_{mr} + 0.33 * dis_{pos}$	31.5	44.4	76.2
$0.33 * dis_{mr} + 0.67 * dis_{pos}$	33.4	48.6	79.6

TABLE 11
Effect of Number of Trajectories in Domain Distance: *Jiading* 2G GSM Data Set

No. of Traj in Target	Median Error on Target (meters)		
	Source Selection by $dist(D, D')$	Source Selection by dis_{pos}	Non-Transfer
1-4	42.2	50.1	62.6
5-8	34.2	39.3	55.3
8+	27.8	32.7	45.2

Finally, it is not surprising that source selection by the distance integrating the weighted dis_{mr} and dis_{pos} leads to the best result.

In terms of the weights w_{mr} and w_{pos} (See Equation (8) in Section 4), we study the effect of weight setting on the errors of TLoc. As shown in Table 10, using either dis_{mr} or dis_{pos} alone, i.e., $w_{mr} = 1.0$ or $w_{pos} = 1.0$, cannot lead to the least error. Instead, the equal weights $w_{mr} = w_{pos} = 0.5$ lead to the best result. It makes sense because the position recovery model maps MR features to associated positions. Thus, in general, dis_{mr} and dis_{pos} leads to roughly equal importance for domain distance $dist(D, D')$.

We note that the weight setting should be adaptive to the ratio of MR samples between target domains and source ones. To this end, for a given ratio of MR samples between a target domain and source domains, we empirically tune the weights w_{mr} and w_{pos} which lead to the least prediction error, and plot the weight against the MR sample ratio in Fig. 13b. When the ratio is close to 0.0 (indicating that the target domain has very few labelled MR samples), w_{mr} values are typically greater than w_{pos}. It is because the domain distance mainly depends upon MR features instead of MR positions (due to the ratio equal to 0.0, i.e., very few MR position labels in target domains). As the ratio becomes greater, i.e., more target labelled samples, w_{mr} remains stabilize equal around 0.5, consistent with Table 10.

Number of Trajectories. Recall that relative position distance is dependent on the number of trajectories in domains. Thus, to study the effect of the number of trajectories on localization errors, in Table 11, we divide all target domains into three groups according to the number of trajectories. For each target domain in a group, we compute the distance between this target domain and a certain source domain, then use the distance as the criterion for source domain selection, and finally compute the average median error for all target domains in this group. From this table, the group with more trajectories corresponds to lower localization errors. It is mainly because in our datasets, the group with more trajectories indicates a higher spatial coverage rate of MR samples in target domains. Moreover, more trajectories in target domains indicate more significant contribution of the weight w_{pos} and lead to low errors, which is consistent with the result in Fig. 14a.

6.6 Sensitivity Study

In this section, we vary the values of several key parameters and study the performance of TLoc.

Transfer Learning Techniques in Random Forests. Recall that we adapt the Structure Transfer (STL) technique for model transfer. Besides STL, the previous work [26] proposed two

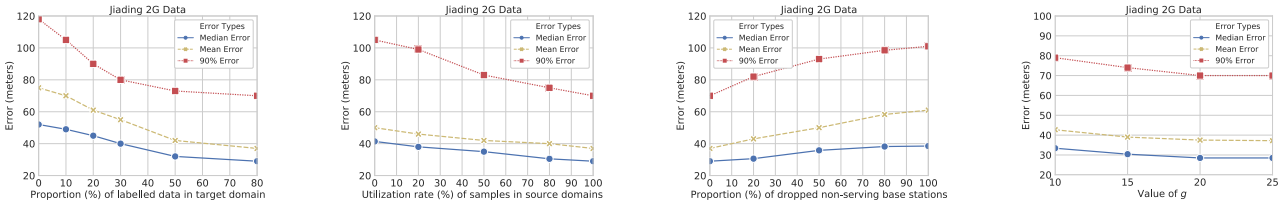


Fig. 14. Sensitivity Study (from left to right): (a-b-c) Proportion of target samples, source samples, and dropped base stations. (d) Effect of grid size g.

other model transfer algorithms: Structure Expansion/Reduction (SER) and MIX. Here, SER searches greedily for locally optimal modifications of each tree structure by trying to locally expand or reduce the tree around individual nodes, and MIX utilizes a majority vote on the decision trees transferred by either STL or SER. As shown in Table 12, we evaluate the effectiveness of these three model transfer techniques. STL leads to the lowest localization errors and SER suffers from the highest errors. It is mainly because the selected source domains are with the highest similarity with the target domain, and STL does not significantly update the DTs trained from source domains. These results are consistent with the previous work [26], where source and target images do share the similar geometric shapes though with various inverted colors and other features. In addition, the running time of STL is much faster than STL and MIX due to the trivial update of the node thresholds in the decision trees of STL. Therefore, we implement our model transfer by STL.

Proportions of Target Samples. First, by varying the proportions of data samples in target domains from $0 \sim 80\%$, we train TLoc and plot the mean, median and 90 percent errors in Fig. 14a. When no data samples are used in target domains, TLoc has to fully leverage the trained models from source domains, leading to the highest errors. When more samples are used in target domains, the errors become gradually smaller. It is mainly because TLoc adapts the models originally trained on source domains towards target domains.

Proportions of Source Samples. Besides the samples in target domains, we also vary the proportion of source data samples from 0 to 100 percent in Fig. 14b. The proportion equal to 0, i.e., the *No-Transfer* approach, suffers from the highest error. More source samples lead to lower errors. Nevertheless, when comparing the sub-Figs. 14a and 14b, we find that TLoc is more sensitive to the data samples in target domains. It makes sense because TLoc performs the position recovery on target domains, and the data samples in target domains thus directly determine the errors of TLoc.

TABLE 12
Effect of Different Transfer Learning Techniques in Random Forest: *Jiading* 2G GSM Data Set

Transfer Learning Techniques	Localization Error (meters)			Avg. Training Time per Target domain (s)
	Median	Mean	90%	
STL	28.1	40.2	72.3	14.2
SER	33.4	48.7	80.4	25.9
MIX	31.1	43.3	77.9	53.4

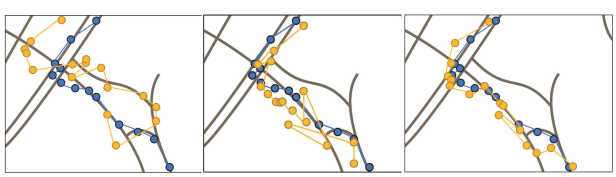
(STL: Structure Transfer, SER: Structure Expansion/Reduction)

Base Stations Density. From Table 3, we find that the base stations in *Jiading* datasets (both 2G GSM and 4G LTE) are much sparser than those in *Siping*. Thus, the localization errors in *Jiading* datasets are slightly higher than those in *Siping* dataset. Moreover, we note that TLoc builds the position recovery model by referring to serving base stations as domain centers. Thus, we randomly choose some non-serving base stations in each MR sample, and reset such base stations and associated cellular signal strength values to be empty. In this way, we drop these base stations from MR samples and vary the density of base stations in MR datasets. In Fig. 14c, the x-axis shows the percentage of dropped base stations and the y-axis gives the median error. A larger dropping rate leads to higher localization error. However, when the total number of dropped base stations rises, the prediction error does not rise sharply. This experimental result indicates that the localization precision of TLoc mainly depends upon serving base stations.

Count g of Divided Grids. Recall that in Section 3.2, we represent the MR features $F_d()$ by grid IDs which require the division of each domain into $g * g$ smaller girds. In Fig. 14d, when the number g of divided grids grows (i.e., smaller grid width/height), the error of TLoc first increases and then remains stable when $g > 20$. The reason is as follows. A smaller g (i.e., larger grid width/height) leads to more neighboring base stations within each divided grid cell, and consequently incurs the coarser-grained representation of $F_d()$. It results in higher errors. Instead, a greater g divides a domain into more cells with smaller grid width/height, and thus leads to lower errors. The tuning of g involves the aforementioned trade-off and we empirically set $g = 20$ by default.

6.7 Localization Visualization

Finally we visualize the positions recovered by three RaF-based algorithms (non-Transfer, MTL, and TLoc) on a randomly selected domain in *Jiading* 2G GSM data set. We choose these approaches is mainly because they lead to the top-3 best results. As shown in Fig. 15, the blue dots represent the GPS position labels (as ground truth) and orange ones represent the prediction result of each algorithm. For each algorithm, we connect the recovered positions into



(a) Non-Transfer (b) MTL (c) TLoc

Fig. 15. Visualization result. Blue: Ground truth; Orange: Predicted position.

a moving trajectory. By observing the two moving directions which are parallel and vertical to road segments, we find that the *non-Transfer* algorithm leads to the largest significant shift in both horizontal and vertical directions. Instead, TLoc can achieve the least shift and the trajectory recovered by TLoc roughly matches the road segments.

6.8 Discussion

Changes in Base Stations. Recall that the relative coordination space of TLoc takes a serving base station as the center of a domain. Though the changes of base stations are not frequent, it is not rare that the software and/or hardware of base stations are updated. We show that how TLoc is adaptive to the update. First, we consider the case that a base station is moved to a new location. We then have two sets of MR samples, denoted by S and S', generated by the base station in the previous and new positions, respectively. To make sure that TLoc works, one simply way is to leverage the new MR samples S' to train a new RaF regression model. Nevertheless, if the number of new MR samples S' is trivial, the new model does not work very well. This is exactly the same challenge that we expect to address in this paper. To this end, we could re-use the RaF regression model which was trained by the previous MR samples S, and transfer this previously trained model from S to S'. Specifically, given the model trained from S, we can follow the general idea of structured transfer learning (STL) in Section 5, and re-select node thresholds in decision trees (DTs) by these new MR samples S'. In this way, the previous model is transferred to the new dataset S'. Second, due to hardware upgrade (e.g., the update from 2G base stations to 4G ones), the signal transmission power of the station might become significantly different. Given such a scenario, the new MR samples generated by the updated station do not follow the original mapping from MR features to relative positions, and TLoc treats the station as a completely new one and has to re-train the location model based on new MR samples S'.

Upcoming 5G Network. With the coming of 5G communication, it is highly expected that 5G base stations are much densely deployed than 2G GSM and 4G LTE stations. Nevertheless, we believe that TLoc can still bring benefits to Telco operators due to the following observations. *1*) Nowadays a Telco operator typically maintains heterogeneous cellular networks mixed by 4G LTE, 3G WCDMA, and/or 2G GSM technologies. With the deployment of 5G network in near future, it is expected that Telco operators could still maintain heterogeneous networks. Thus, TLoc can still work to recover the positions of mobile devices using non-5G cellular networks. *2*) Even for a 5G network, it is highly possible that 5G base stations deployed in rural areas could be much sparse than those in urban areas. TLoc still has chance to work well in rural areas.

7 Conclusion and Future Work

In this paper, we study the problem of cellular outdoor position recovery in the areas with insufficient geo-tagged MR samples, and design a transfer learning-based position recovery framework, namely TLoc. The contributions of our work include *1*) the proposed relative coordinate space to represent MR features and positions, *2*) the distance metric to measure the similarity of domains, and *3*) a transfer learning-based position recovery framework by adapting the STL approach. Our extensive evaluation validates that TLoc outperforms two state-of-the-art methods (CCR and NBL) and the variants of TLoc. As TLoc is a first stepping stone to explore transfer-learning for Telco outdoor position recovery, the promising results motivate the following future work, e.g., deep neural network (DNN)-based position recovery [35] empowered by transfer learning techniques [11], [32].

Acknowledgments

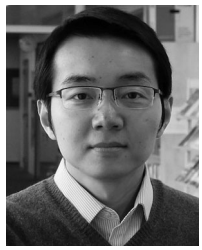
This work was supported in part by the National Natural Science Foundation of China (Grant No. 61572365, No. 61772371, and No. 61972286).

References

[1] Decision tree learning, 2019. [Online]. Available: https://en.wikipedia.org/wiki/Decision_tree_learning
[2] Google maps for mobile, 2020. [Online]. Available: https://play.google.com/store/apps/details?id=com.location.test
[3] M. J. Afridi, A. Ross, and E. M. Shapiro, "On automated source selection for transfer learning in convolutional neural networks," *Pattern Recognit.*, vol. 73, pp. 65–75, 2018.
[4] H. Aly and M. Youssef, "Dejavu: An accurate energy-efficient outdoor localization system," in *Proc. 21st ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2013, pp. 154–163.
[5] H. B. Ammar *et al.*, "An automated measure of MDP similarity for transfer in reinforcement learning," in *Proc. Workshops 28th AAAI Conf. Artif. Intell.*, 2014, pp. 31–37.
[6] R. A. Becker *et al.*, "A tale of one city: Using cellular network data for urban planning," *IEEE Pervasive Comput.*, vol. 10, no. 4, pp. 18–26, Apr. 2011.
[7] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[8] A. Chang and J. Chang, "Robust mobile location estimation using hybrid TOA/AOA measurements in cellular systems," *Wireless Pers. Commun.*, vol. 65, no. 1, pp. 1–13, 2012.
[9] C. Costa and D. Zeinalipour-Yazti, "Telco big data: Current state & future directions," in *Proc. 19th IEEE Int. Conf. Mobile Data Manag.*, 2018, pp. 11–14.
[10] M. de Berg, O. Cheong, M. J. van Kreveld, and M. H. Overmars, *Computational Geometry: Algorithms and Applications*, 3rd ed. Berlin, Germany: Springer, 2008.
[11] M. Gong, K. Zhang, T. Liu, D. Tao, C. Glymour, and B. Schölkopf, "Domain adaptation with conditional transferable components," in *Proc. 33rd Int. Conf. Int. Conf. Mach. Learn.*, 2016, pp. 2839–2848.
[12] J. Hoy, *Forensic Radio Survey Techniques for Cell Site Analysis*. Hoboken, NJ, USA: Wiley, 2014.
[13] Y. Huang *et al.*, "Experimental study of telco localization methods," in *Proc. 18th IEEE Int. Conf. Mobile Data Manag.*, 2017, pp. 299–306.
[14] M. Ibrahim and M. Youssef, "CellSense: An accurate energy-efficient GSM positioning system," *IEEE Trans. Veh. Technol.*, vol. 61, no. 1, pp. 286–296, Jan. 2012.
[15] P. Indyk and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in *Proc. 13th Annu. ACM Symp. Theory Comput.*, 1998, pp. 604–613.
[16] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. 14th Int. Joint Conf. Artif. Intell.*, 1995, pp. 1137–1145.
[17] H. Koshima and J. Hoshen, "Personal locator services emerge," *IEEE Spectr.*, vol. 37, no. 2, pp. 41–48, Feb. 2000.
[18] L. J. Caffery and G. L. Stuber, "Overview of radiolocation in CDMA cellular systems," *IEEE Commun. Mag.*, vol. 36, no. 4, pp. 38–45, Apr. 1998.
[19] D. Lin, X. An, and J. Zhang, "Double-bootstrapping source data selection for instance-based transfer learning," *Pattern Recognit. Lett.*, vol. 34, no. 11, pp. 1279–1285, 2013.

[20] R. Margolies *et al.*, "Can you find me now? Evaluation of network-based localization in a 4G LTE network," in *Proc. IEEE INFOCOM*, 2017, pp. 1–9.
[21] P. W. Mirowski *et al.*, "Probability kernel regression for WiFi localisation," *J. Location Based Serv.*, vol. 6, no. 2, pp. 81–100, 2012.
[22] S. J. Pan, D. Shen, Q. Yang, and J. T. Kwok, "Transferring localizastion models across space," in *Proc. 23rd Int. Conf. Artif. Intell.*, 2008, pp. 1383–1388.
[23] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
[24] A. Ray, S. Deb, and P. Monogioudis, "Localization of LTE measurement records with missing information," in *Proc. IEEE INFOCOM*, 2016, pp. 1–9.
[25] H. Rizk, M. Torki, and M. Youssef, "CellinDeep: Robust and accurate cellular-based indoor localization via deep learning," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2305–2312, Mar. 2019.
[26] N. Segev, M. Harel, S. Mannor, K. Crammer, and R. El-Yaniv, "Learn on source, refine on target: A model transfer learning framework with random forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1811–1824, Sep. 2017.
[27] A. Shokry, M. Torki, and M. Youssef, "DeepLoc: A ubiquitous accurate and low-overhead outdoor cellular localization system," in *Proc. 26th ACM SIGSPATIAL Int. Conf. Advances Geographic Inf. Syst.*, 2018, pp. 339–348.
[28] J. Simm, I. M. de Abril, and M. Sugiyama, "Tree-based ensemble multi-task learning method for classification and regression," *IEICE Trans. Inf. Syst.*, vol. E97-D, no. 6, pp. 1677–1681, 2014.
[29] A. Thiagarajan, L. Ravindranath, H. Balakrishnan, S. Madden, and L. Girod, "Accurate, low-energy trajectory mapping for mobile devices," in *Proc. 8th USENIX Conf. Netw. Syst. Des. Implementation*, 2011, pp. 267–280.
[30] R. M. Vaghefi, M. R. Gholami, and E. G. Ström, "RSS-based sensor localization with unknown transmit power," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 2480–2483.
[31] C. Wu, J. Xu, Z. Yang, N. D. Lane, and Z. Yin, "Gain without pain: Accurate WiFi-based localization using fingerprint spatial gradient," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, 2017, Art. no. 29.
[32] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.
[33] M. Youssef and A. K. Agrawala, "The horus WLAN location determination system," in *Proc. 3rd Int. Conf. Mobile Syst. Appl. Serv.*, 2005, pp. 205–218.
[34] M. Yuan *et al.*, "OceanST: A distributed analytic system for large-scale spatiotemporal mobile broadband data," *Proc. VLDB Endowment*, vol. 7, no. 13, pp. 1561–1564, 2014.
[35] Y. Zhang, W. Rao, K. Zhang, M. Yuan, and J. Zeng, "PRNet: Outdoor position recovery for heterogenous telco data by deep neural network," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 1933–1942.
[36] K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma, "Explaining the power-law distribution of human mobility through transportation modality decomposition," *Sci. Rep.*, vol. 5, 2015, Art. no. 9136.
[37] V. W. Zheng, S. J. Pan, Q. Yang, and J. J. Pan, "Transferring multi-device localization models using latent multi-task learning," in *Proc. 23rd Int. Conf. Artif. Intell.*, 2008, pp. 1427–1432.
[38] V. W. Zheng, E. W. Xiang, Q. Yang, and D. Shen, "Transferring localization models over time," in *Proc. Int. Conf. Artif. Intell.*, 2008, pp. 1421–1426.
[39] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 38:1–38:55, 2014.
[40] F. Zhu *et al.*, "City-scale localization with telco big data," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 439–448.

Yige Zhang received the BSc degree in software engineering from Tongji University, Shanghai, China, in July 2016. She is currently working toward the PhD degree in the School of Software Engineering, Tongji University, China since September 2016. Her research interests include mobile computing and machine learning.



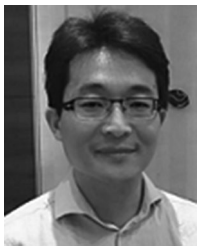
Aaron Yi Ding received the MSc and PhD degrees (with distinction) from the Department of Computer Science (Birthplace of Linux), University of Helsinki, Helsinki, Finland. He is currently an assistant professor with the Department of Engineering Systems and Services, TU Delft. Prior to joining TU Delft, he has worked with TU Munich (2016-2018), Germany, with the University of Helsinki (2007-2016), Finland, with Columbia University (2014), and with the University of Cambridge (2013), United Kingdom. His PhD was supervised by Prof. Sasu Tarkoma and Prof. Jon Crowcroft with the University of Cambridge. He has been awarded Best Paper of ACM EdgeSys, ACM SIGCOMM Best of CCR, and Nokia Foundation Scholarships.

Jörg Ott received the diploma and doctoral (Dr-Ing) degrees in computer science from TU Berlin, Berlin, Germany, in 1991 and 1997, respectively, and the diploma degree in industrial engineering from TFH Berlin, Berlin, Germany, in 1995. He holds the chair of connected mobility with the Department of Informatics, Technical University of Munich. He is also adjunct professor for networking technology with Aalto University. His research interests include network architecture, (Internet) protocol design, and networked systems, with a focus on (mobile) decentralized services and cloudless applications. One particular focus has been delay-tolerant networking and distributed computing for Internet services in remote areas, sensing and management in underground mines, and operation in similarly demanding situations. He has contributed to standardization of Internet protocols since 1993. He has cofounded four companies in Germany and Finland on performance enhancement and secure satellite communications, mobile Internet access, continuous quality and performance measurements for cloud-based multimedia services, and networking in challenging environments.

Mingxuan Yuan received the PhD degree from the Hong Kong University of Science and Technology, Hong Kong. He is currently a principal researcher of Noahs Ark Lab, Huawei. Before joined Huawei, he worked with HKUST as a postdoc researcher. His research interests include spatiotemporal data analytics and enterprise operation optimization models. He has led several projects in telecommunication data mining and supply chain optimization.

Jia Zeng (Senior Member, IEEE) received the BEng degree from the Wuhan University of Technology, Wuhan, China, in 2002, and the PhD degree from the City University of Hong Kong, Hong Kong, in 2007. He is currently the chief scientist for enterprise intelligence (e.g., supply chain management) with Huawei Noahs Ark Lab, Hong Kong. He is also an adjunct professor with the School of Computer Science and Technology, Soochow University, Suzhou 215006, China. His research interests include machine learning and big data applications. He is a member of the CCF and ACM.

Kun Zhang received the PhD degree from The Chinese University of Hong Kong, Hong Kong, in 2005. He is currently an associate professor with the Philosophy Department and an affiliate faculty member with the Machine Learning Department, Carnegie Mellon University. His research interests lie in machine learning and artificial intelligence, especially in causal discovery, causality-based learning, and general-purpose artificial intelligence. He coauthored a Best Student Paper at UAI 2010, received the Best Benchmark Award of the causality challenge 2008, and coauthored a Finalist Best Paper at CVPR 2019. He has served as an area chair or senior program committee member for major conferences in machine learning or artificial intelligence, including NeurIPS, UAI, ICML, AISTATS, AAAI, and IJCAI, and has organized various academic activities to foster interdisciplinary research in causality.



Weixiong Rao (Member, IEEE) received the PhD degree from The Chinese University of Hong Kong, Hong Kong, in 2009. After that, he worked for Hong Kong University of Science and Technology (2010), University of Helsinki (2011-2012), and University of Cambridge Computer Laboratory Systems Research Group (2013) as postdoctor researchers. He is currently a full professor with the School of Software Engineering, Tongji University, China (since 2014). His research interests include mobile computing and spatiotemporal data science, and is a member of the CCF and ACM.