Article

Transfer Learning for Wireless Fingerprinting Localization Based on Optimal Transport

Siqi Bai [ID] **, Yongjie Luo and Qun Wan ***

School of Information and Communication Engineering, University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, Chengdu 611731, China; 201511020304@std.uestc.edu.cn (S.B.); luoyongjie@cdu.edu.cn (Y.L.)
* Correspondence: wanqun@uestc.edu.cn

Abstract: Wireless fingerprinting localization (FL) systems identify locations by building radio fingerprint maps, aiming to provide satisfactory location solutions for the complex environment. However, the radio map is easy to change, and the cost of building a new one is high. One research focus is to transfer knowledge from the old radio maps to a new one. Feature-based transfer learning methods help by mapping the source fingerprint and the target fingerprint to a common hidden domain, then minimize the maximum mean difference (MMD) distance between the empirical distributions in the latent domain. In this paper, the optimal transport (OT)-based transfer learning is adopted to directly map the fingerprint from the source domain to the target domain by minimizing the Wasserstein distance so that the data distribution of the two domains can be better matched and the positioning performance in the target domain is improved. Two channel-models are used to simulate the transfer scenarios, and the public measured data test further verifies that the transfer learning based on OT has better accuracy and performance when the radio map changes in FL, indicating the importance of the method in this field.

Keywords: fingerprinting localization; transfer learning; optimal transport; indoor positioning; adaptive radio map

1. Introduction

In the mode of pervasive computing, people can acquire and process information at any time, any place, and in any way. Location information is essential for pervasive computing. Satellite positioning technology has been able to meet most outdoor location acquisition requirements, and indoor positioning technologies are constantly emerging to get through the "last meter" of positioning technology [1]. According to wireless technologies, the existing methods include Wifi positioning, Bluetooth positioning, ultra-wideband positioning, and lidar positioning, and so on. According to the measurement techniques, the existing methods include time of arrival (TOA) positioning, angle of arrival (AOA) positioning, received signal strength (RSS) positioning [2], Channel State Information (CSI) positioning [3], etc. According to the algorithms, the existing methods include triangulation, direct positioning, fingerprint localization (FL), and so on. Compared to other techniques, the FL can be applied to a complex environment, with wide application scope and easy implementation [2–5].

FL avoids the need for artificial modeling of complex indoor wireless channels and is typically achieved through machine learning techniques such as classification or regression algorithms. As with traditional machine learning applications, FL usually assumes that the training fingerprint data (also called radio map) are sampled from the same distribution as the test data. However, in practice, many factors will cause fingerprint distribution to change. For example, when using RSS fingerprints

as features, changes in Access Point (AP) number/location, positioning environment, or equipment parameters will all lead to changes in channel parameters, resulting in changes in the location fingerprint distribution. As a result, the model's accuracy obtained from the old training data will decrease or even fail. This can be addressed by seeking transfer learning techniques.

Transfer learning studies how to transfer the knowledge from the old data domain to the new data domain to deal with the problem that there is no labeled data (unsupervised) or insufficient labeled data (semi-supervised) in the new domain. Transfer learning is a new paradigm in machine learning technology, which has been successfully applied in many machine learning fields, such as the biomedical field [6,7], language and text recognition field [8], graph neural network field [9], detection field [10], etc., and also in positioning [11].

The existing studies on FL transfer learning mainly consider three scenarios: time transfer [12–14], device transfer [15,16], and space transfer [17]. Time transfer refers to transferring the collected fingerprints or trained model knowledge from one time to another. Device transfer is the transfer from one device to another. Spatial transfer refers to the transfer from one area to another. No matter which kind of transfer is used, the essential learning is that the fingerprint distribution has changed from one state to another. Feature-based transfer learning has recently been verified to be effective in FL problems [18]. However, the frequently used maximum mean difference (MMD) distance does not consider the detailed differences between the two distributions.

In this paper, a novel transfer learning method based on optimal transport (OT) [19] is applied to FL. By introducing the Laplacian regularization and jointly learning mechanism, a smoother mapping function can be learned to improve the algorithm's robustness further. We use both the free space channel model and the multi-wall model to simulate the proposed method's performance and analyze the reason why the OT-based transfer learning performance is good. The performance of the algorithm is further verified on the public measured data set. The results indicate that OT technology is significant to the transfer learning problem in FL.

2. Related Works

2.1. Wireless Fingerprinting Localization

RADAR is one of the earliest wireless fingerprint positioning systems [20]. It adopts the KNN regression positioning method: a classic FL method with a simple algorithm and good robustness. Subsequently, support vector machine [21], decision tree [22], neural network [23], and other machine learning methods are also used in fingerprint positioning. These methods are deterministic methods, which map the location fingerprint to a specific location. Another type of method is called the probability method, which considers the position fingerprint's randomness and maps a fingerprint into the probability density of the position. The usual methods are naive Bayes [24], probability kernel regression [25], conditional random field [26], and so on.

2.2. Transfer Learning in the Wireless Fingerprinting Localization

The research on the transfer learning of wireless FL started at about the same time as the general transfer learning. Yin et al. proposed a time transfer method of radio fingerprint map in 2005 [12], which uses regression analysis to learn the predictive relationship between the received signal strength of mobile devices and the received signal strength of reference points on sparse locations, to update the radio map in real-time according to the predictive relationship. LeManCoR [13] is a kind of adaptive method for fingerprint mapping based on the Manifold co-regularization. The LuMA method [15] takes into account both time transfer and device transfer. In addition, there are studies that consider multi-device transfer [16] and spatial transfer [17] in fingerprint positioning. All the above methods require the target domain to have a small amount of labeled data, so they belong to semi-supervised transfer learning. Furthermore, they all adopt the manifold alignment algorithm. TrHMM [14] is also a time transfer learning method for FL. Unlike the previous method, it is a parameter transfer learning

method based on the Hidden Markov model. None of these methods adopt the latest available transfer learning algorithm.

General feature-based transfer learning methods include Transfer Component Analysis (TCA) [27], Joint Distribution Adaptation (JDA) [28], Balanced boundary Distribution Adaptation (BDA) [29], etc. These methods are types of unsupervised transfer learning, where there is no label information in the target domain. An improved method based on the general feature-based transfer learning has been successfully applied to FL recently [18]. However, the method has an excess of super parameters, and the performance may worsen when the super parameters are not chosen carefully. It is urgent to develop a fingerprint transfer learning method with fewer super parameters and robust performance in practical application.

2.3. Optimal Transport

The theory of optimal transport (OT) originates from what mathematician Gaspard Monge described in 1746 to 1818 as the "Monge problem" about how to move a sand pile to another place and change its shape into a predefined one at minimal cost [19]. It is then used by mathematicians to compare the distance between two probability distributions. In recent years, thanks to approximate solvers' appearance that can be extended to high-dimensional problems, the revolution of OT technology has been initiated. It has been successfully applied to a variety of problems in image science (such as color or texture processing), graphics (for shape processing), or machine learning (for regression, classification, and generation modeling), as well as solving the problem of transfer learning [30].

3. Problem Description

FL is a method of associating location information with its fingerprint and then using parameterized or nonparametric models for location identification. Specific environmental features are the basis for creating fingerprint information. In wireless FL, it uses the wireless signal features to create a fingerprint of the position, also known as a wireless fingerprint. The fingerprints in the interesting area construct a radio map. Various wireless channel measurements can be selected as site-specific signal features, such as TOA, AOA, RSS, CSI, etc. Sometimes, they can be fused to form a higher-dimensional feature space. The signal feature is mapped to the position fingerprint in a preset way. Then, a sample associated with position $\mathbf{u}$ can be expressed as $\mathbf{x}(\mathbf{u}) = (x_1(\mathbf{u}), ..., x_m(\mathbf{u})) \in \mathcal{X}^m$, and its conditional probability density is denoted as $P_{X|U}$, where, $\mathcal{X}^m$ represents the m-dimensional fingerprint space.

Wireless FL usually includes two stages: the offline stage and the online stage. A radio map is first created in the offline phase, which contains data pairs of several coordinates and corresponding fingerprints within the location area. Then, the radio map and the learning algorithm are used to get the position recognition function $g : \mathbf{x} \rightarrow \mathbf{u}$, which maps a fingerprint to its estimated coordinate. In the online phase, the target's estimated coordinate $g(\tilde{\mathbf{x}})$ is obtained according to the online fingerprint $\tilde{\mathbf{x}}$ and the function g.

In transfer learning of FL, it is assumed the data distribution in the offline and online phases comes from different distributions. Mathematically, let the source domain be $\mathcal{D}_s = \left\{ \mathcal{X}_s^m, P_X^s(\mathbf{x}) \right\}$ and the task in the source domain be $\mathcal{T}_s = \left\{ \mathcal{U}_s^d, g_s(\cdot) \right\}$, where the subscript or superscript "s" represents the source domain, and if changed to "t", it represents the target domain. The following text will follow this notation. $\mathcal{X}^m$ represents the m-dimensional fingerprint space, and $\mathcal{U}^d$ represents the d-dimensional position space. In this paper, $d = 2$ is setted, and the superscripts "m" and "d" will be ignored below when it is well defined. $g(\cdot)$ stands for the position recognition function. The joint distribution of the offline phase, $P_{X,U}^s(\mathbf{x}, \mathbf{u})$, is associated with the source domain. Moreover, the joint distribution of the online phase, $P_{X,U}^t(\mathbf{x}, \mathbf{u})$, is associated with the target domain. In transfer learning setting, $P_{X,U}^s(\mathbf{x}, \mathbf{u}) \neq P_{X,U}^t(\mathbf{x}, \mathbf{u})$. Transfer learning of FL studies how to transfer the knowledge about location

from the old radio map (in the source domain) to the new one (in the target domain), so as to make full use of the knowledge of the source domain to optimize the task of FL in the target domain.

In practice, multiple instances of fingerprints in a domain can be observed, either with or without a location label. Assume that the fingerprint–coordinate pair set $D_s = \left\{ \left(\mathbf{x}_i^{(s)}, \mathbf{u}_i^{(s)} \right) \middle| \mathbf{x}_i^{(s)} \in \mathcal{X}_s, \mathbf{u}_i^{(s)} \in \mathcal{U}_s, i = 1, ..., n_s \right\}$ was observed in the source domain, including n_s samples in total. In the target domain, the unlabeled fingerprint set $D_t = \left\{ \mathbf{x}_i^{(t)} \middle| \mathbf{x}_i^{(t)} \in \mathcal{X}_t, i = n_s + 1, ..., n_s + n_t \right\}$ and labeled fingerprint set $D'_t = \left\{ \left(\mathbf{x}_i^{(t)}, \mathbf{u}_i^{(t)} \right) \middle| \mathbf{x}_i^{(t)} \in \mathcal{X}_t, \mathbf{u}_i^{(t)} \in \mathcal{U}_t, i = n_s + n_t + 1, ..., n_s + n_t + n'_t \right\}$ were observed, including n_t and n'_t samples, respectively, and in general we have $n'_t \ll n_s \ll n_t$. When $n'_t = 0$, it is called unsupervised transfer learning; otherwise, it is called semi-supervised transfer learning.

The transfer learning task of FL is to estimate the position recognition function $\hat{g}_t(\mathbf{x}; D_s, D_t, D'_t)$ in the target domain according to the observed fingerprint samples in the source domain and the target domain, so as to minimize the generalization error of the target domain. The generalization error is expressed as follows,

$$E_{\mathbf{x} \in \mathcal{X}_t} \left[\left\| \hat{g}_t(\mathbf{x}; D_s, D_t, D'_t) - g_t(\mathbf{x}) \right\| \right]. \tag{1}$$

When $\mathcal{X}_s = \mathcal{X}_t$, it is called homogeneous transfer learning. This paper considers unsupervised homogeneous transfer learning.

Hypotheses should be made to theoretically guarantee the transfer learning to succeed [31]. The following are hypotheses often used in the transfer learning problem.

- Class imbalance hypothesis: the distribution of labels in the two fields is different, i.e., $P_Y^s(\mathbf{y}) \neq P_Y^t(\mathbf{y})$, but the conditional probability distribution of the feature is the same, i.e., $P_{X|Y}^s(\mathbf{x}|\mathbf{y}) = P_{X|Y}^t(\mathbf{x}|\mathbf{y})$.
- Covariance offset hypothesis: the marginal distribution of the two domains is different, that is, $P_X^s(\mathbf{x}) \neq P_X^t(\mathbf{x})$, but the conditional probability distribution of the label is the same, that is, $P_{Y|X}^s(\mathbf{y}|\mathbf{x}) = P_{Y|X}^t(\mathbf{y}|\mathbf{x})$ (equivalent to the learning function $g_s = g_t = g$).

In wireless FL, the typical scenario that requires transfer learning can be summarized into two cases:

1. The channel parameters on one or more links are changed;
2. The channel parameters of a local region are changed.

Whichever case is considered, the above hypotheses are too strong to be satisfied. First, the class imbalance hypothesis requires that the distribution of fingerprints at each location remain the same. Second, the covariance offset hypothesis requires that the position recognition function be the same. Therefore, the transfer learning algorithm based on these two hypotheses is easy to fail in FL.

In addition, the feature-based transfer learning approach assumes the existence of a pair of mapping functions $\{\varphi_s(\cdot), \varphi_t(\cdot)\}$ that maps features from the source and the target domains to a common latent domain, while the labels remain unchanged. Therefore, the learning function in target domain $g_t(\cdot)$ is approximately replaced by $g_l : \varphi_s(\mathbf{x}_s) \to \mathbf{u}_s$, the position recognition function trained in the latent domain.

4. Transfer Learning for Fingerprinting Localization

4.1. Transfer Component Analysis

Transfer Component Analysis (TCA) [27] is one of the most usual feature-based transfer learning methods, whose principle is to adaptive the marginal distribution of the feature. TCA learns the cross-domain transfer components in the reproducing kernel Hilbert space by minimizing the MMD distance between the source domain and the target domain samples after mapping. Let the number of

samples in the source domain and target domain be n_s and n_t, respectively, and the MMD distance between them is [32]

$$D(\mathbf{X}_s, \mathbf{X}_t) = \left\| \frac{1}{n_s} \sum_{i=1}^{n_s} \varphi(\mathbf{x}_i) - \frac{1}{n_t} \sum_{j=1}^{n_t} \varphi(\mathbf{x}_j) \right\|, \tag{2}$$

where $\mathbf{X}_s$ and $\mathbf{X}_t$ represent the fingerprint sample matrix of source domain and target domain, respectively. However, it is usually highly nonlinear and difficult to directly minimize the MMD distance. The above distance can be converted into kernel function form, so the problem is converted to kernel matrix learning and written as semi-definite program (SDP) in the form of

$$\max_{\mathbf{K} \succ 0} tr(\mathbf{K}\mathbf{L}) - \lambda tr(\mathbf{K}), \tag{3}$$

where $\mathbf{K}$ is the kernel matrix defined on all data. Let $\mathbf{K}_{s,s}$, $\mathbf{K}_{t,t}$, and $\mathbf{K}_{s,t}$, respectively, represent the Gram matrix defined on the source domain, target domain, and cross-domain data, $\mathbf{K}_{ij} = \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j)$; then,

$$\mathbf{K} = \begin{bmatrix} \mathbf{K}_{s,s} & \mathbf{K}_{s,t} \\ \mathbf{K}_{s,t}^T & \mathbf{K}_{t,t} \end{bmatrix}. \tag{4}$$

The elements in matrix $\mathbf{L}$ is calculated as

$$\mathbf{L}_{ij} = \begin{cases} \frac{1}{n_s^2} & \mathbf{x}_i \mathbf{x}_j \in \mathcal{X}_s \\ \frac{1}{n_t^2} & \mathbf{x}_i \mathbf{x}_j \in \mathcal{X}_t \\ -\frac{1}{n_s n_t} & otherwise \end{cases} \tag{5}$$

The results can be constructed with dimension reduction method, that is, solve the first m eigenvalues of $(\mathbf{KLK} + \mu\mathbf{I})^{-1}\mathbf{KHK}$, reducing the computational cost of solving SDP. The TCA method is the base method in feature-based transfer learning, many other methods are extended upon it, such as JDA [28] and BDA [29].

4.2. Optimal Transport For Fingerprint Transfer Learning

4.2.1. Basic Method

In this paper, we implemented the optimal transport (OT) method in the transfer learning of FL. Different from the feature-based method (such as TCA), it is assumed that the drift of the domain is caused by an unknown transformation $T : \mathcal{X}_s \to \mathcal{X}_t$ from the distribution of source domain to the distribution of target domain. In the FL, the corresponding physical interpretation of the drift may be the change of fingerprint acquisition conditions, changes in environmental parameters, changes in noise conditions, or other unknown processes. Let us say that transformation T maintains the condition distribution of the location label in this process, namely,

$$P_t(\mathbf{u}_t \,|\mathbf{x}_t \,) = P_s(\mathbf{u}_s \,|T(\mathbf{x}_s)\,). \tag{6}$$

This means that the transformation maintains the information of the position decision function, so the position estimator in the target domain can be approximated by the estimator, $g_t(\mathrm{T}(\mathbf{x_s}))$, trained after the source domain is mapped to the target domain, as shown in the schematic diagram in Figure 1. Then, the knowledge about the location recognition function is transferred from the old radio map to the new one.
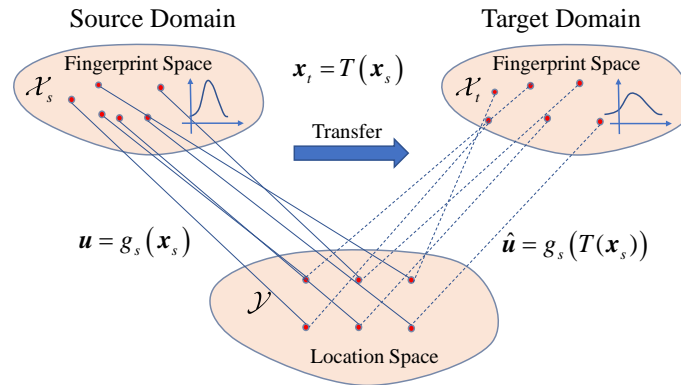
Figure 1. Schematic diagram of the transfer learning in FL based on OT.

From the perspective of probability, T transforms the marginal measure p_X^s of a fingerprint on the source field to the measure of its image, which is represented by $T\#p_X^s$. It is another measure on $\mathcal{X}_t$, which satisfies

$$T\#p_X^s(\mathbf{x}) = p_X^s(T^{-1}(\mathbf{x})). \tag{7}$$

If $T\#p_X^s = p_X^t$, T is called a transport map from p_X^s to p_X^t. Under this assumption, $\mathbf{X}_t$ comes from the same probability distribution as $T\#p_X^s$. Therefore, the principle of solving the problem of transfer learning in FL through OT is the same as that in [16]:

1. The probability measures p_X^s and p_X^t are estimated using $\mathbf{X}_s$ and $\mathbf{X}_t$.
2. Find a transport map T, from p_X^s to p_X^t.
3. The labeled sample $\mathbf{X}_s$ is transported with T, and then the target domain estimator $g_t(\cdot)$ is trained with the transformed samples.

The key point is to find the right transport T. OT finds T by minimizing the transport cost $C(T)$:

$$C(T) = \int_{\mathcal{X}_s} c(\mathbf{x}, T(\mathbf{x}))dp_X^s(\mathbf{x}), \tag{8}$$

where the cost function $c : \mathcal{X}_s \times \mathcal{X}_t \to R^+$ is a distance function in the feature space $\mathcal{X}$. $C(T)$ can be interpreted as the total energy required to move the fingerprint probabilistic mass p_X^s to the fingerprint probabilistic mass $T\#p_X^s$. The solution to the OT problem defined by The Monge problem is

$$\begin{aligned} T_0 &= \arg\min_{T} \int_{\mathcal{X}_S} c(\mathbf{x}, T(\mathbf{x}))dp_X^s(\mathbf{x}) \\ &s.t. T\#p_X^s = p_X^t \end{aligned}. \tag{9}$$

The problem in (9) is combinatorial, and the feasible set of which is nonconvex. Therefore, solving the Monge problem is difficult. Kantorovitch form of OT is a convex relaxed version. Define $\Pi \in \mathcal{P}(\mathcal{X}_S \times \mathcal{X}_T)$ as a set of probabilistic coupling, whose marginal measures are p_X^s and p_X^s. The Kantorovitch problem requires finding the probabilistic coupling minimizing the following formula,

$$\gamma_0 = \arg\min_{\gamma \in \Pi} \int_{\mathcal{X}_s \times \mathcal{X}_t} c(\mathbf{x}_s, \mathbf{x}_t)d\gamma(\mathbf{x}_s, \mathbf{x}_t), \tag{10}$$

where γ can be regarded as the joint probability measure with marginal measure p_X^s and p_X^t, also known as transport map. The above formula has been proved to be applicable to define the distance between

distributions, which is called Wasserstein distance or Earth Mover distance. The Wasserstein distance of order n between p_X^s and p_X^t is defined as

$$
\begin{aligned}
d_{w,n}\left(p_X^s, p_X^t\right) &\equiv \left(\inf_{\gamma \in \Pi} \int_{\mathcal{X}_s \times \mathcal{X}_t} d(\mathbf{x}_s, \mathbf{x}_t)^n d\gamma(\mathbf{x}_s, \mathbf{x}_t)\right)^{\frac{1}{n}} \\
&= \left(\inf_{\gamma \in \Pi} \left(\mathbb{E}_{x_s \sim p_X^s, x_t \sim p_X^t} d(\mathbf{x}_s, \mathbf{x}_t)^n\right)\right)^{\frac{1}{n}}
\end{aligned}
\tag{11}
$$

Compared with MMD distance defined in formula (2), Wasserstein distance better describes the contour and detail differences of the two distributions.

As discrete samples are obtained in the actual situation, only the empirical distributions of p_X^s and p_X^s can be obtained, denoted as

$$
p_X^s = \sum_{i=1}^{n_s} p_i^s \delta_{\mathbf{x}_i^s}, p_X^t = \sum_{i=1}^{n_t} p_i^t \delta_{\mathbf{x}_i^t},
\tag{12}
$$

where $\delta_{\mathbf{x}_i}$ and p_i represent the Dirac function and the probabilistic mass at the sample x_i, respectively. p_i belongs to the probabilistic simplex, namely, $\sum_i p_i = 1$. Then, define the probability coupling matrix as follows,

$$
\mathcal{B} = \left\{ \gamma \in \left(\mathbb{R}^+\right)^{n_s \times n_t} | \gamma \mathbf{1}_{n_t} = p_X^s, \gamma^T \mathbf{1}_{n_s} = p_X^t \right\},
\tag{13}
$$

where $\mathbf{1}_d$ is the d-dimensional full 1 column vector. The OT of the discrete version of Kantorovitch form can be expressed as

$$
\gamma_0 = \arg\min_{\gamma \in \mathcal{B}} \langle \gamma, C \rangle_F,
\tag{14}
$$

where $\langle \cdot, \cdot \rangle_F$ represents the Frobenius product, $C \geq 0$ represents the cost function matrix, and $C(i,j) = c(\mathbf{x}_i^s, \mathbf{x}_j^t)$ represents the cost required to transport the probability mass from $\mathbf{x}_i^s$ to $\mathbf{x}_j^t$. For simplicity, the Euclid square distance is used as the cost in this paper, that is, $c(\mathbf{x}_i^s, \mathbf{x}_j^t) = \left\| \mathbf{x}_i^s - \mathbf{x}_j^t \right\|_2^2$.

In general, γ_0 is a sparse matrix containing $n_s + n_t - 1$ nonzero elements at most. (14) is a linear programming problem, which can be solved by simplex method [33]; however, the complexity is high. The OT regularized with entropy adds an entropy regularized term $\Omega s(\gamma)$, namely,

$$
< \gamma_0 >_{\text{Sink.}} = \arg\min_{\gamma \in \mathcal{B}} \langle \gamma, \mathbf{C} \rangle_F + \lambda_e \Omega_e(\gamma),
\tag{15}
$$

where $\Omega_e(\gamma) = \sum_{i,j} \gamma(i,j) \log \gamma(i,j)$ is the negative entropy of γ, and λ_e represents the corresponding regularization coefficient. By adding the entropy of γ, a smoother transport diagram can be obtained, and a more efficient algorithm has been derived, called Sinkhorn-knopp [34].

After γ_0 is solved, barycentric matching [35] can obtain the mapping of all source domain samples in the target domain:

$$
\hat{\mathbf{X}}_s = n_s \gamma_0 \mathbf{X}_t.
\tag{16}
$$

The purpose of the OT transfer learning is to correctly recover the transport graph from the data distribution in the source domain to the data distribution in the target domain, and what kind of transformation it can recover has not been proved theoretically. However, it has been proved that the affine transformation of discrete distributions can be recovered lossless [30].

4.2.2. Laplacian Regularization

In FL, fingerprints that are intuitively "close" in the source domain should also be "close" when transported to the target domain, and vice versa. Let $\hat{\mathbf{x}}_i^s$ represent the value after the source domain sample $\mathbf{x}_i^s$ is mapped to the target domain, and the Laplacian regularization term is introduced:

$$\Omega_l(\gamma) = \alpha \cdot \frac{1}{n_s^2} \sum_{i,j} \mathbf{S}_s(i,j) \|\hat{\mathbf{x}}_i^s - \hat{\mathbf{x}}_j^s\|_2^2 + (1 - \alpha) \cdot \frac{1}{n_t^2} \sum_{i,j} \mathbf{S}_t(i,j) \|\hat{\mathbf{x}}_i^t - \hat{\mathbf{x}}_j^t\|_2^2, \tag{17}$$

where $\mathbf{S}_s(i,j) \geq 0$ is the element of sample similarity matrix $\mathbf{S}_s$ in the source domain, and $\mathbf{S}_t$ is similar. α is a super parameter, which represents the importance factor of Laplacian regularization in the source domain. When the marginal distribution is uniform, the above formula can be further simplified by formula (16):

$$\Omega_l(\gamma) = \alpha Tr\left(\mathbf{X}_t^T \gamma^T \mathbf{L}_s \gamma \mathbf{X}_t\right) + (1 - \alpha) Tr(\mathbf{X}_s^T \gamma \mathbf{L}_t \gamma^T \mathbf{X}_s), \tag{18}$$

where $\mathbf{L}_s = diag(\mathbf{S}_s \mathbf{1}) - \mathbf{S}_s$ is the Laplacian matrix associated with the graph $\mathbf{S}_s$, similarly, $\mathbf{L}_t = diag(\mathbf{S}_t \mathbf{1}) - \mathbf{S}_t$. When using Laplacian regularization, we solve the following problem,

$$< \gamma_0 >_{\text{EMDL}} = \underset{\gamma \in \mathcal{B}}{\arg\min} \langle \gamma, \mathbf{C} \rangle_F + \lambda_l \Omega_l(\gamma), \tag{19}$$

where λ_l represents the coefficient of the Laplacian regularization. Then, the subsequent matching process goes the same as the Sinkhorn algorithm.

4.2.3. Joint Estimation of Transport Map and Transformation Function

The transport map is responsible for transporting the empirical probabilistic mass from the source domain to the target domain, or vice versa. The algorithm gets the transport map of probability density, not a transformation function. Jointly learning the transport map and transformation function makes the learner better extend to unknown samples, which is known as out-of-sample case [36]. The cost function of joint transport and transformation estimation is

$$< \gamma_0, T_0 >_{\text{M. OT}} = \underset{\gamma \in \mathcal{B}, T \in \mathcal{H}}{\arg\min} \frac{1}{n_s m} \|T(\mathbf{X}_s) - n_s \gamma \mathbf{X}_t\|_{\mathcal{F}}^2 + \frac{\lambda_\gamma}{\max(\mathbf{C})} \langle \gamma, \mathbf{C} \rangle_F + \frac{\lambda_T}{m^2} R(T), \tag{20}$$

where $R(\cdot)$ is the regularization term related to the transformation T; λ_γ and λ_T are regularization coefficients. The hypothesis space of the transformation T, $\mathcal{H}$, can be either a linear or nonlinear function space, and we adopted a linear function in this paper.

4.2.4. Data Preprocessing and Optimization Algorithm

In our experiments, the source domain and target domain data are both normalized by subtracting the mean value and dividing by the variance of the data. Data preprocessing is rarely mentioned in former transfer learning studies, but it significantly impacts performance. This is because data preprocessing can reduce the difference of mean value and variance between two domains to a certain extent, which is similar to the effect of "transfer".

For the optimization of the Laplacian regularization version in Formula (19), we adopt the Generalized Conditional Gradient (GCG) algorithm [37] to solve the optimization of the OT problem in this paper. The regularization in (19) is strictly convex, so the objective function can reach the minimum on $\mathcal{B}$. Specifically, using $f(\gamma)$ represents the objective function in formula (19), the method iterates the steps below until convergence:

$$\tilde{\gamma}^{l+1} = \arg \min_{\gamma \in \mathcal{B}} \left\langle \gamma, \nabla f(\gamma^l) \right\rangle,$$
$$\gamma^{l+1} = \gamma^l + \tau^l(\tilde{\gamma}^{l+1} - \gamma^l), \tag{21}$$

where τ^l is obtained by linear search. For the optimization of the joint matching algorithm version in Formula (20), the block-coordinate descent (BCD) [38] method is used. The idea is to alternatively optimize λ and T.

These optimization algorithms are available on a public website [39] that readers can refer to.

5. Wireless Fingerprint Channel Model

In this paper, RSS fingerprint characteristics are taken as an example, assuming that there are m APs in the interesting region and the region is a 2-dimensional Euclidean space. The coordinate of the i-th AP is $\mathbf{a}_i$, $i = 1, \ldots, m$. The APs transmit wireless signals at a certain power. The fingerprint at the coordinate $\mathbf{u}$ in the source domain is represented by $\mathbf{x}_s(\mathbf{u})$, which is composed of the received power from all APs, namely, $\mathbf{x}_s(\mathbf{u}) = \left[x_1^s(\mathbf{u}), x_2^s(\mathbf{u}), \ldots, x_m^s(\mathbf{u}) \right]$. In the target domain, the channel link parameters change, and the fingerprint at coordinate $\mathbf{u}$ is represented by $\mathbf{x}_t(\mathbf{u})$. The commonly used model of RSS fingerprints are free-space model, multi-wall model, and ray tracing [40]. In this paper, the free space model and multi-wall model are used to simulate the distribution changes in fingerprint transfer.

5.1. Free Space Loss Model

In the free space loss model, it is assumed that the received power is calculated over a long period of time without considering the small-scale fading of the channel. Assuming that the receiving power (unit mW) follows a lognormal distribution, the fingerprint component of the mobile device located at $\mathbf{u}$ from the i-th AP is [41]

$$x_i(\mathbf{u}) = P_i - 10 \cdot \alpha_i \log_{10} \left(\|\mathbf{a}_i - \mathbf{u}\| / d \right) + N_i, \tag{22}$$

where P_i represents the received power of the i-th AP at the reference distance d (usually 1 m), with unit dBm; α_i is the path loss exponent of the i-th AP; and N_i is the link noise of the i-th AP, which obeys a Gaussian white noise with variance σ_i^2. Assuming that AP is independent of each other, the conditional probability density function of the location fingerprint is

$$P(\mathbf{x} \mid \mathbf{u}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\left(x_i - \left(P_i - 10 \cdot \alpha_i \log_{10} \left(\|\mathbf{a}_i - \mathbf{u}\| / d \right) \right) \right)^2}{2\sigma_i^2} \right]. \tag{23}$$

The free space loss model is a signal propagation model in an ideal environment. In this paper, the free space model is used to model the fingerprint changes when the link parameters of the channel change. Suppose the link noise of all AP is equal, that is, $N_1 = N_2 = \cdots = N_m = N$, and remains unchanged. When the link parameters change from $\boldsymbol{\theta}_s = \left[P_1^s, P_2^s, \ldots, P_m^s, \alpha_1^s, \alpha_2^s, \ldots, \alpha_m^s \right]^T$ to $\boldsymbol{\theta}_t = \left[P_1^t, P_2^t, \ldots, P_m^t, \alpha_1^t, \alpha_2^t, \ldots, \alpha_m^t \right]^T$, the conditional probability of location fingerprint will change from $P_s(\mathbf{x} \mid \mathbf{u}; \boldsymbol{\theta}_s)$ to $P_t(\mathbf{x} \mid \mathbf{u}; \boldsymbol{\theta}_t)$.

5.2. Multi-Wall Model

The multi-wall model is an extension of the single-slope loss model, including an additional attenuation term, which is caused by the loss of direct path between transmitter and receiver encounters the wall and door [42]. In the multi-wall model, Formula (22) is rewritten as

$$x_i(\mathbf{u}) = P_i - 10 \cdot \alpha_i \log_{10} \left(\|\mathbf{a}_i - \mathbf{u}\| / d \right) - M_w + N_i, \tag{24}$$

where the additional attenuation term can be expressed as

$$\mathrm{M}_w = l_c + \sum_{i=1}^{I} k_{wi} l_i + \sum_{n=1}^{N_d} \chi_n l_d + \sum_{n=1}^{N_{fd}} \lambda_n l_{fd}, \tag{25}$$

where k_{wi} represents the number of penetrable wall of type i, l_i is the corresponding loss of signals passing through it; N_d and N_{fd} are the number of ordinary doors and fire doors passing through the direct path, respectively; l_d and l_{fd} are losses corresponding to signals passing through ordinary doors and fire doors, respectively; and χ_n / λ_n is a binary variable, indicating the state of the n-th ordinary door/fire door.

The multi-wall model considers the attenuation of the signal after passing through the wall, which can be used to simulate the RSS fingerprint under different indoor structures. In the experiment part, the multi-wall model is used to model the transfer learning when the local area's channel parameters change. Compared with the source domain, some wall structures of the target domain are changed.

6. Experiments

To verify the performance of OT-based transfer learning in FL, two models described in Section 5 are used for numerical simulation, and the performance of the algorithm is also verified with the public data set. First, we use the free space channel model to simulate the transfer scenario of RSS fingerprint when the radio link parameters change. Second, the indoor multi-wall model was used to simulate RSS fingerprint transfer's learning scene when the environmental parameters of a local area changed. Finally, the performance of the algorithm is further verified by using the publicly measured data set. One of the performance evaluation indicators used in this paper is the average positioning error (AE), and its calculation formula is

$$AverageError = \frac{1}{n_t} \sum_{j=1}^{n_t} \sqrt{\left\| \hat{\mathbf{u}}_j - \mathbf{u}_j \right\|_2}, \tag{26}$$

where $\hat{\mathbf{u}}_j$ and $\mathbf{u}_j$, respectively, represent the estimated and real coordinates of the j-th sample in the target domain test set, and n_t represents the total number of the test samples. The cumulative error value of 50% and 80% was used as additional evaluation indicators.

6.1. Free Space Channel Model RSS FL Transfer Learning Simulation

The simulation was set as a 1-d positioning scene, with range $[0, 10]$ and 2 APs located at -1 and 11, respectively. Using the model described in Section 5.1, the channel parameters, P_i and α_i, are changed to simulating the change in the radio link parameters. The parameters of the source domain and target domain are shown in Table 1. In the source domain, the positioning area is divided into 10 grid points, and the center of each grid serves as the real label of the position. Ten samples are randomly generated in each grid as the training set. In the target domain, 1000 samples were randomly collected as test sets. The standard deviation of the noise was set at 2 dBm.

Table 1. Channel model parameters change in the source domain and the target domain.

AP	AP1		AP2	
	P_1/dBm	α_1	P_2/dBm	α_2
source domain	50	1	80	3
target domain	40	1	100	4

The TCA method and OT linear map [43] were used to transfer the samples of the above simulation scenarios. The left side of Figure 2 shows the normalized fingerprint samples in the source domain and the target domain. The two axes, respectively, represent the two features, and the size of the sample point represents the relative value of the real coordinate values. Accordingly, the right side of Figure 2 shows the results after the source domain and the target domain are mapped to the latent domain through TCA transformation. Figure 3 shows the fitted distributions of the two features on the left and the fitted distributions after mapping to the latent domain through TCA transformation on the right.
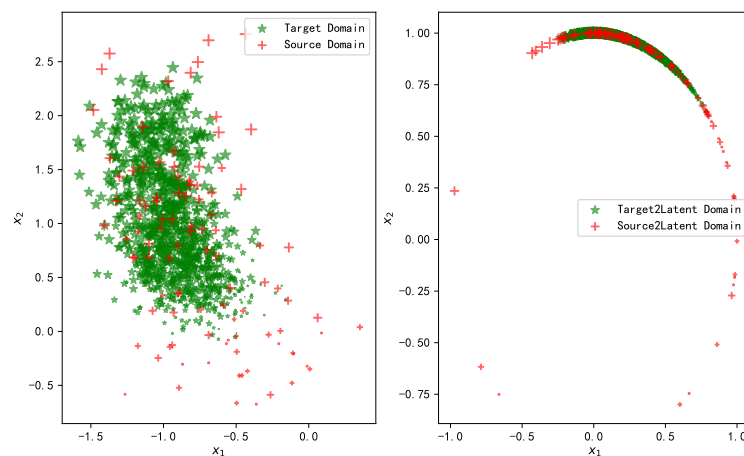
Figure 2. Fingerprint samples in source and target domains (**left**) after TCA transformation to the latent space (**right**).
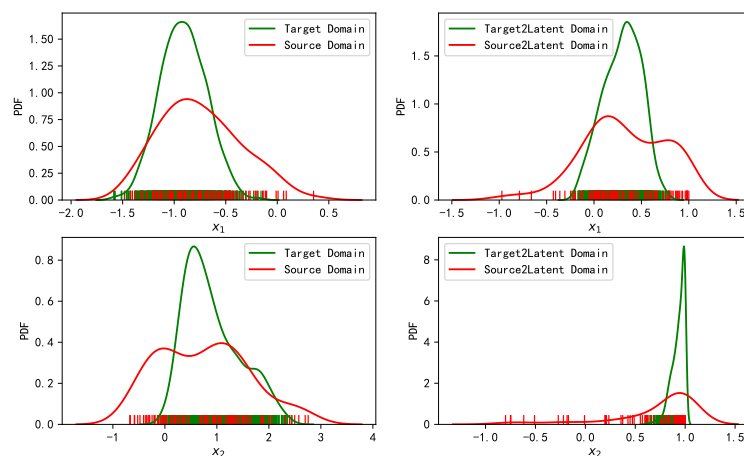

Figure 3. Fitted distributions of features in the source domain and target domain (**left**), and the fitted distributions of features after TCA transformation to the latent space (**right**).

It can be observed from Figures 2 and 3 that TCA transformation fails to match the distributions of the two domains. The samples' variance in the source domain is large, while it is small in the target domain. This is still the case after mapping to the latent domain. We know that the objective of TCA is to minimize the mean difference of the mapped samples. As the samples have been normalized, the mean difference between the two domains is relatively small. Therefore, TCA does not significantly improve it. As the cost function does not constrain the variance, the variance between the two distributions is still massive after TCA transfer.

Figure 4 shows the changes in the fingerprint samples in the data domain before and after the OT-based transfer learning. The left part of Figure 4 shows the samples in the fingerprint space from the source domain, the target domain, and the target domain mapped from the source domain through OT. The middle and the right part are the fitted distributions of the two features in different domains.

It can be observed from Figure 4 that the samples mapping from the source domain to the target domain with OT have a higher matching degree with the sample distribution in the target domain, whatever the mean, the variance, or the contour. We know that OT is to transport the source domain's distribution to the distribution of the target domain under the principle of minimum cost, so the two distributions are matched better.
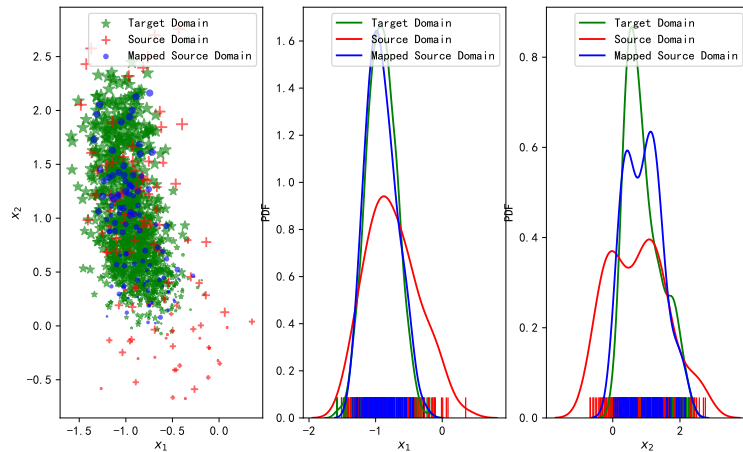

Figure 4. Fingerprint samples in the source domain, in the target domain, and after mapping from the source domain to the target domain with OT (**left**), the corresponding fitted distributions of features (**center** and **right**).

According to the Transfer Learning theory, the generalization error bound of transfer learning is related to the distance between the distributions of the two domains, and the smaller the distance is, the lower bound will be reached [44]. It is observed from the above simulation that the distributions obtained by OT match better than what by TCA. Moreover, Figure 5 shows the relationship between the average test error with changing the noise level without transfer learning, TCA learning, and OT learning. It can be observed from Figure 5 that the transfer performance using OT is the best, and the advantage gradually decreases with the increase of noise level. Besides, TCA does not improve the target domain location performance but rather weakens. From the fitted distribution of Figures 3 and 4, it is obvious that the positioning performance has a great correlation with the degree of distribution coincidence.
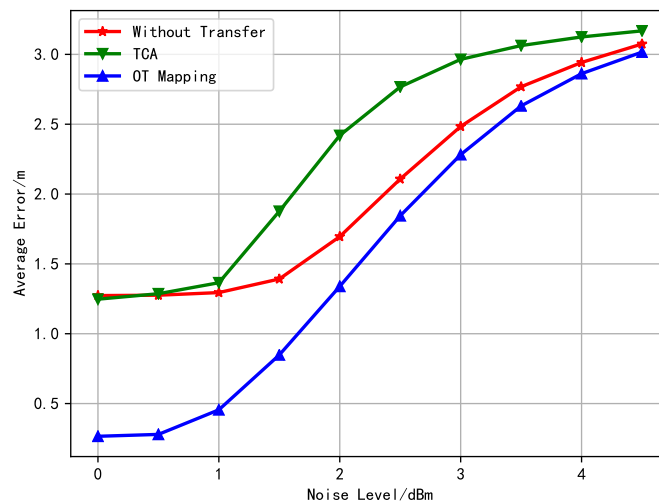

Figure 5. Comparison of the average positioning error of the transfer learning algorithms (using free space loss model).

6.2. Multi-Wall Model of RSS FL Transfer Learning Simulation

We use the indoor multi-wall model to simulate the transfer learning algorithm's performance when the local propagation environment changes. Figure 6 shows the heat maps of the power obtained by simulating two environments with 6 APs placed at the same relative locations. Roughly, the two environments have the same layout, and the area is about $50 * 20$ m, but there are fewer walls in Environment 1. The simulation data of Environment 1 are taken as samples in the source domain. Some walls are added in Environment 2 to simulate the local changes compared with Environment 1. The simulation data of Environment 2 is taken as samples in the target domain. In both environments, 7200 fingerprint samples were collected at uniform and the same locations, without noise. It can be observed from Figure 6 that there are some local differences in the power energy due to local layout changes.
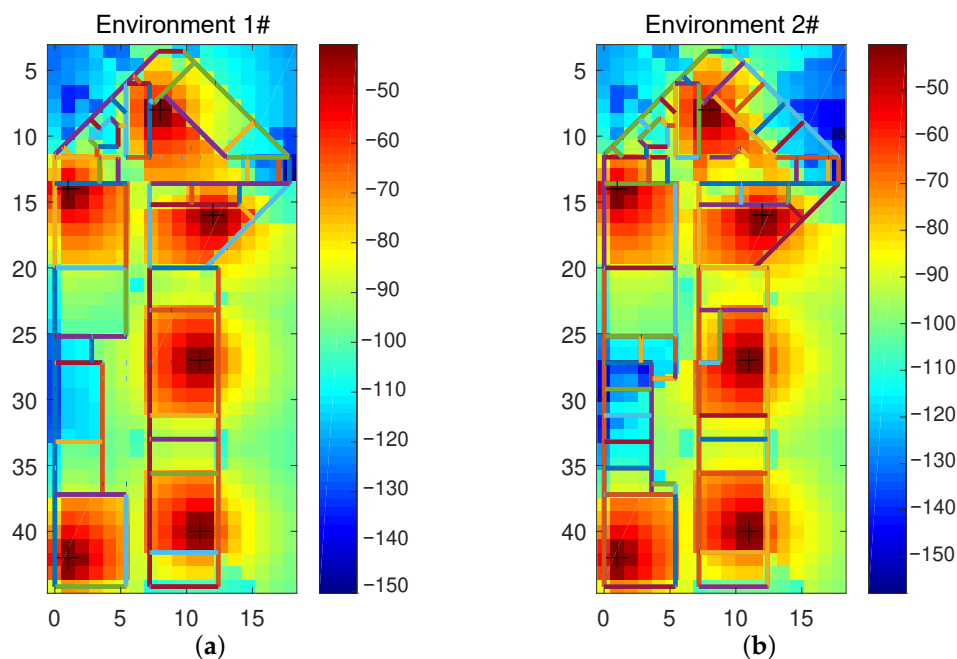

Figure 6. The layout (unit: meter) and the power heat maps (unit: dBm) of simulated indoor multi-wall model: (**a**) environment 1/the source domain; (**b**) Environment 2/the target domain.

Figure 7 shows a heat map of the power mapped from source to latent domain by TCA transformation on the left and from target to latent domain on the right. The data are preprocessed by normalization before the transformation. It can be observed that the heat map on the right is still locally different from that on the left.

Figure 8 shows the power heat map of the source domain on the left, the target domain in the middle, and the source domain after mapping to the target domain using OT on the right. It can be seen that OT makes the heat map mapped from the source domain more similar to that in the target domain. However, note that the heat map is the superposition of all APs, so it is not possible to ultimately determine whether the two are similar from this figure alone.

Finally, the relationship between the algorithms' average test error and the noise-level is shown in Figure 9. The curve in Figure 9 has a similar trend with that in Figure 5. This indicates that OT plays a positive transfer role in both models, while the TCA method appears to negative transfer [40] in both cases.
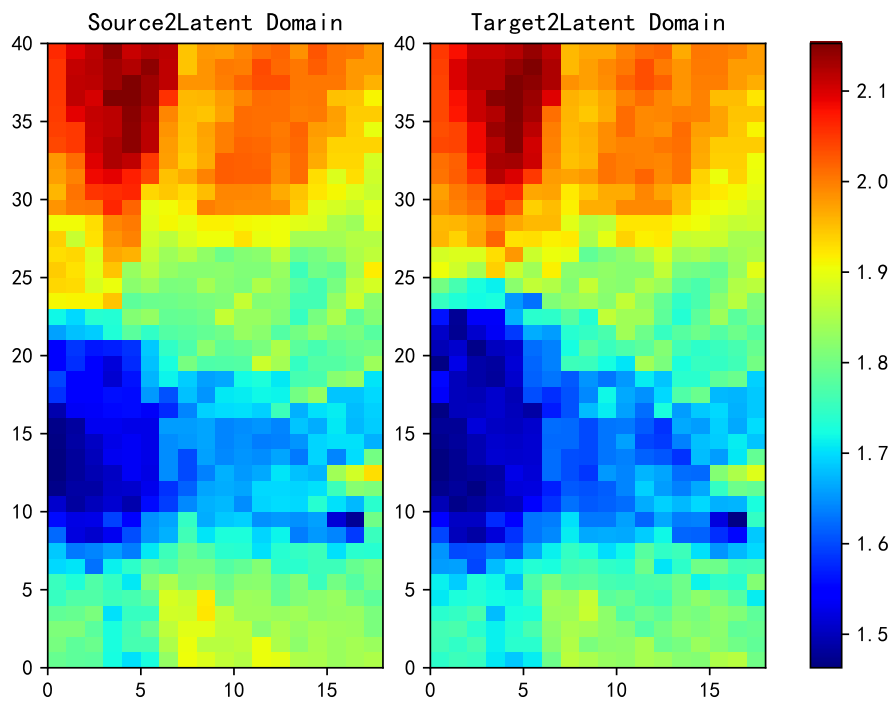
Figure 7. Power and heat graphs from the source domain (**left**) and the target domain (**right**) to latent domain obtained by TCA transformation.
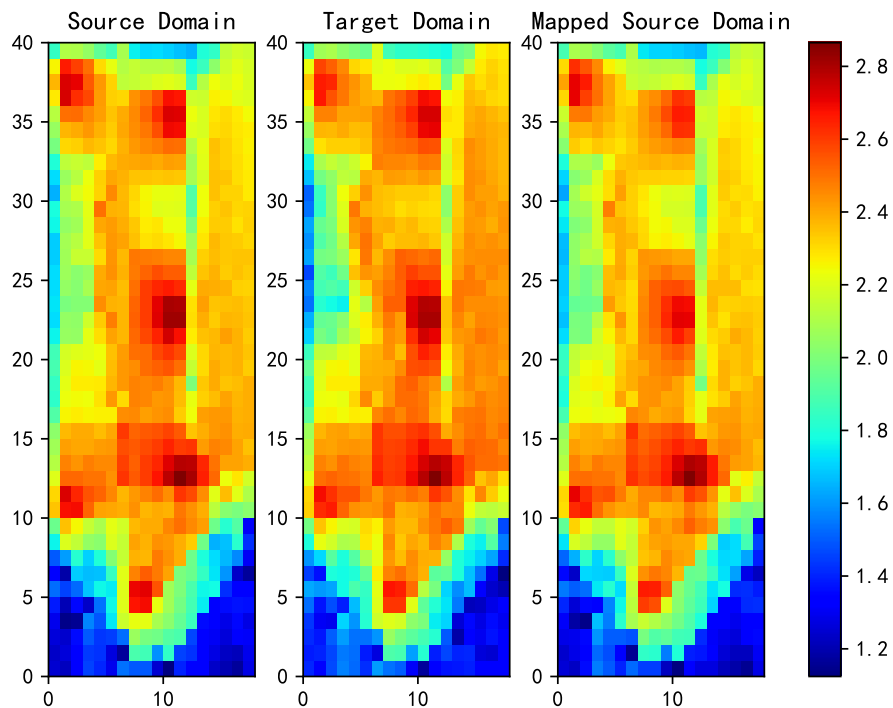

Figure 8. Power heat map in the source domain (**left**), power heat map in the target domain (**middle**), and power heat map after the source domain is mapped to the target domain by OT method.
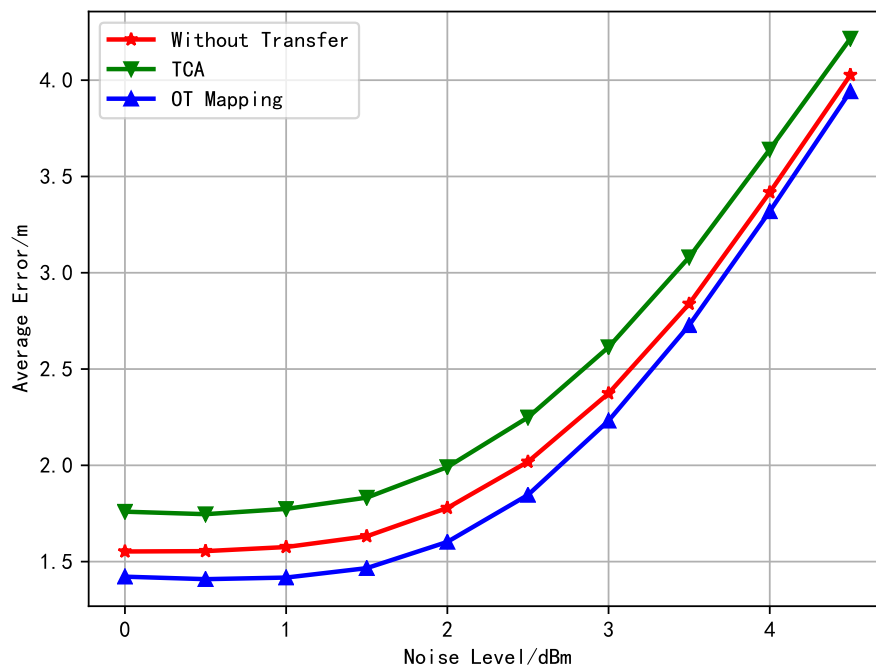
Figure 9. Comparison of the average positioning error of the transfer learning algorithms (using multi-wall model).

6.3. Measured Data Experiment

In this section, we use the recent publicly available measured data set [45] to test the performance of transfer learning based on OT and typical feature-based transfer learning algorithms in FL. The data set was collected at the same locations in a library over 15 natural months, using the same device each month. A total of 448 APs were detected in the whole data set. However, due to the long period of sample collection, the number of APs detected every month was different, especially since the 12th month, which had a significant change. In this experiment, the first month's data were taken as the sample set of the source domain, with a total of 8640 samples. The mean value of 6 samples in a continuous period and at the same location was taken, and 1440 training samples were finally obtained. The remaining data of each month is taken as the target domain sample set, which contains 6 data sets each month: one of them is taken as the validation set, containing 576 samples; 3 of them are taken as the test set 1, containing 1728 samples; the remaining two are taken as the test set 2, containing 1392 samples. This experiment focuses on the time transfer of location fingerprints, namely, training the model with the first month's data and then using the transfer learning algorithm to transfer the model to the unlabeled samples of the remaining months.

The validation set is used for choosing the super parameters. For each transfer learning algorithm, the grid searching method is used to select the super parameters that make the validation set perform best. In the transductive setting, the test set 1 is used as the target domain samples and test samples simultaneously. In the out-of-sample setting, the transfer learning algorithm uses the test set 1 as the target domain samples, but the test set 2 is used for testing. In this paper, the classical TCA transfer learning algorithm and its extension method BDA are selected as the comparison algorithms. In addition, the performance without transfer learning was also included in the comparison (represented by Without). The other three methods are based on the OT method: EMDL is the OT method with Laplacian regularization (described in Section 4.2.2), Sinkhorn is the OT method with entropy regularization (described in Section 4.2.1), and Map. OT is the joint estimation method (described in Section 4.2.3). The average positioning error of month 2 to 15 with different algorithms is shown in Figure 10, under the transductive setting on the left and out-of-sample setting on the right. It can be observed from the figure that traditional transfer learning methods (TCA and BDA) showed

no significant performance improvement during months 2 to 10 under both settings. While under the transductive setting the average error of transfer learning based on OT was reduced by about 10% compared with not using any transfer algorithms. All transfer learning methods improved in accuracy after month 10, and the OT-based method performs even better under both settings. When the cumulative error distribution function values are at 50% and 80% for each algorithm, the corresponding error values (represented by C.5 and C.8, respectively) are shown in Tables 2 and 3. We can draw similar conclusions from these results.
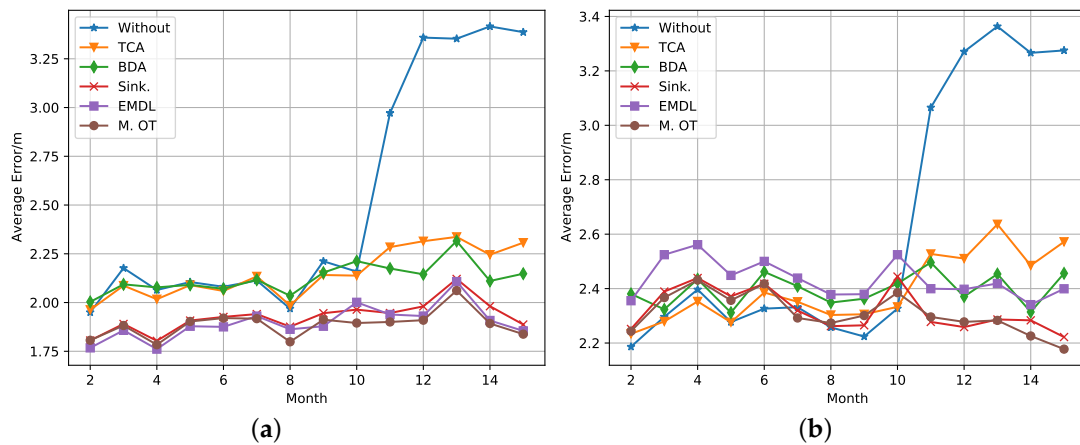

Figure 10. The performance of the transfer learning algorithms on the measured FL data set: (**a**) transductive setting and (**b**) out-of-sample setting.

Table 2. The performance of the transfer learning algorithm on the measured FL data set (transductive setting).

	Month	02	03	04	05	06	07	08	09	10	11	12	13	14	15
	Without	1.59	1.67	1.72	1.75	1.76	1.76	1.61	1.79	1.76	2.76	3.29	3.23	3.36	3.32
	TCA	1.63	1.63	1.63	1.75	1.75	1.76	1.61	1.79	1.76	2.01	1.99	1.97	1.91	1.99
C.5	BDA	1.69	1.68	1.76	1.76	1.73	1.78	1.67	1.79	1.88	1.89	1.76	1.95	1.76	1.79
	EMDL	**1.42**	1.45	1.42	1.45	1.47	1.48	1.45	1.54	1.48	**1.6**	1.49	1.75	1.47	1.48
	Sink.	**1.34**	**1.34**	**1.34**	1.43	1.44	1.46	1.36	1.44	1.47	1.49	1.45	1.54	1.34	1.43
	M. OT	1.44	1.44	1.44	1.45	1.5	1.47	1.42	1.47	1.45	1.5	1.45	1.61	1.42	1.44
	Without	2.93	3.24	3.08	3.2	3.07	3.2	**2.78**	3.2	3.19	4.3	4.6	4.67	4.64	4.57
	TCA	2.98	3.18	3.04	3.19	3.1	3.21	2.96	3.19	3.15	3.39	3.39	3.51	3.24	3.39
C.8	BDA	3.08	3.2	3.08	3.19	3.09	3.12	2.95	3.19	3.2	3.19	3.23	3.51	3.15	3.16
	EMDL	**2.76**	2.82	2.72	2.9	2.89	2.85	2.78	2.86	2.96	2.9	3.08	3.39	3.04	2.78
	Sink.	**2.7**	2.81	2.72	2.78	2.76	2.95	2.78	2.78	2.98	3	2.98	3.46	2.86	2.76
	M. OT	2.73	2.85	2.64	2.81	2.96	2.81	2.64	2.78	2.89	2.81	2.92	3.24	2.82	2.73

Table 3. The performance of the transfer learning algorithm on the measured FL data set (out of sample setting).

	Month	02	03	04	05	06	07	08	09	10	11	12	13	14	15
	Without	**1.99**	**2.06**	2.2	**2.06**	**2.19**	**2.19**	**1.99**	**1.97**	**2.15**	2.83	3.29	3.29	3.25	3.25
	TCA	**2.02**	1.97	2.19	2.08	**2.22**	**2.19**	2.06	2.16	**2.15**	2.31	2.22	2.37	2.22	2.26
C.5	BDA	2.06	**2.06**	**2.15**	2.08	**2.22**	2.2	2.16	2.19	**2.16**	2.31	**2.19**	**2.16**	**2.15**	**2.19**
	EMDL	**2.02**	2.19	**2.16**	2.19	**2.2**	**2.15**	**2.02**	**2.02**	2.2	**2.08**	1.9	1.99	**2.02**	1.97
	Sink.	2.2	2.26	2.26	2.22	2.26	2.22	2.19	2.19	2.22	**2.2**	**2.19**	2.2	2.19	**2.19**
	M. OT	2.08	2.19	**2.16**	2.19	**2.22**	2.06	1.97	2.15	2.2	**2.15**	1.97	2.06	2.02	1.9
	Without	**3.12**	**3.33**	**3.47**	**3.31**	**3.21**	**3.36**	**3.22**	3.2	**3.25**	4.35	4.49	4.73	4.66	4.64
	TCA	**3.21**	**3.32**	**3.51**	**3.32**	**3.4**	3.51	3.4	**3.36**	3.4	3.62	3.65	3.95	3.61	3.78
C.8	BDA	3.58	**3.48**	3.67	3.43	3.55	3.52	3.48	3.52	3.56	3.62	**3.51**	3.67	**3.36**	3.65
	EMDL	3.25	3.51	**3.6**	**3.34**	3.48	**3.47**	**3.21**	**3.25**	3.6	**3.22**	3.39	3.39	3.48	**3.2**
	Sink.	3.43	3.67	3.82	3.51	3.55	3.52	3.47	3.51	3.75	**3.47**	3.55	**3.58**	3.47	3.47
	M. OT	**3.2**	3.51	**3.6**	3.4	**3.47**	3.25	3.32	3.36	3.55	3.25	3.51	3.4	3.22	3.09

The results from the real data set have a slight difference from the simulated data. Nevertheless, they all show the superior performance of the OT-based methods.

6.4. Super Parameters

In this section, we explain how the super parameters in the algorithm are selected and how they affect the average positioning error. In the Sinkhorn algorithm, there is a super parameter λ_e, namely, the entropy regularization coefficient. We empirically select λ_e that minimizes the validation error, as shown in Figure 11. For space reasons, we only show the result in the 12th month, still using the 1st month data as the source domain data. It can be seen from the figure that when λ_e is set at 0, the error is large, and when it is set at 20, the error reaches the minimum value, indicating that moderate entropy regularization plays a role in improving the performance.
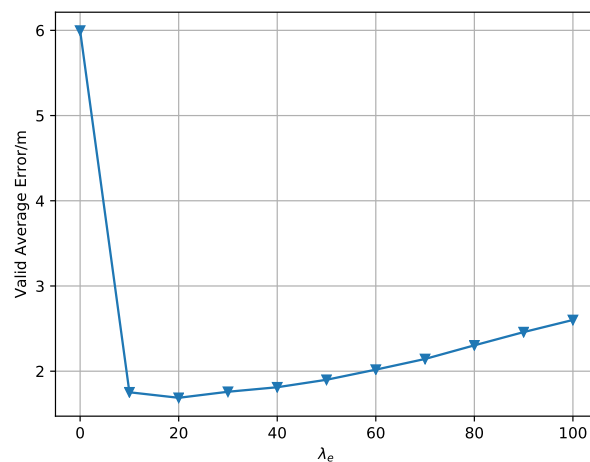
Figure 11. Validation error when taking different super parameters in the 12th month (Sinkhorn).

There are two super parameters in the EMDL algorithm, λ_l and α, which represent the regular coefficient of Laplacian and the proportion of the importance of the source domain data. The M. OT algorithm also has two super parameters, namely, λ_γ and λ_T, which represent the regularization coefficient of transport cost and the regularization coefficient of the matching function. Similarly, in Figures 12 and 13, we, respectively, show the variation of the validation error when taking different super parameters in the 12th month.
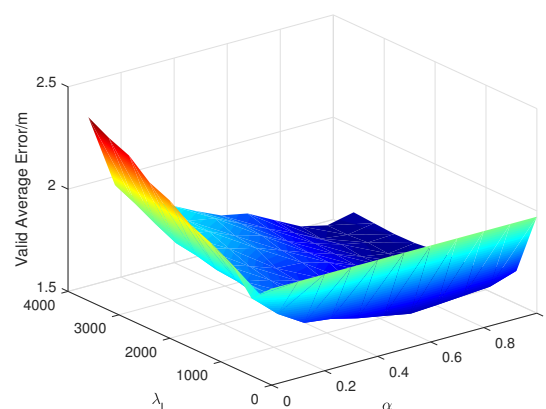

Figure 12. Validation error when taking different super-parameters in the 12th month (EMDL).
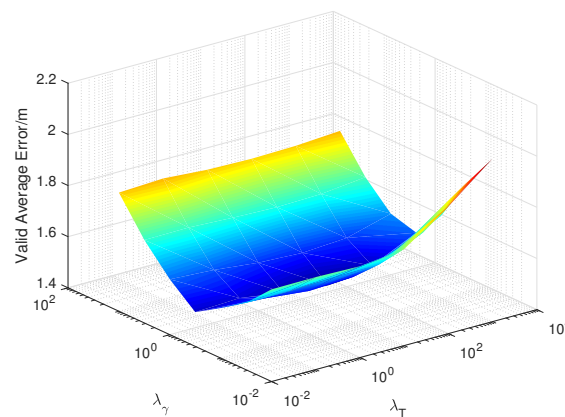
Figure 13. Validation error when taking different super-parameters in the 12th month (M. OT).

As you can see from Figure 12, in the EMDL algorithm, λ_l and α both need to be set to a larger value to achieve better performance. This indicates that Laplacian regularization has improved performance and that the source domain data is of greater importance. As can be seen from Figure 13, the selection of λ_T in the M. OT algorithm is more important than λ_γ to the performance, but, in general, they have little impact on the positioning mean error.

For these three algorithms, we observe that the optimal super parameter value of each month is almost unchanged, which is not shown in the paper due to the limited space. This shows that our algorithm has good robustness in super parameter selection.

7. Discussion

The problem of transfer learning is closely related to the distance metric of data distribution, and OT is an important tool to study the distance of data distribution. The distance derived from OT has many good properties. In this paper, the OT method is applied to the transfer learning of FL. The simulation and measured fingerprint data of different fingerprint models show that the transfer learning method based on the OT method has better transfer performance than the traditional TCA method. We find that this is related to the way they define the distance measure of the empirical distribution. The latter's MMD distance is the most commonly used distance metric in transfer learning, but it only describes the mean value of the distribution, while the Wasserstein distance used by the former gives a good description of the details of the distribution. The latest transfer learning review article does not cover the methods based on OT, which we believe should receive wider attention. This paper analyzes and conducts experiments under two different channel models for the transfer learning problem in FL. We find that the traditional method has a negative transfer effect in the experiment, while the OT method can achieve positive transfer, indicating that the occurrence of negative transfer is related to the algorithm. The experimental study in this paper causes us to think about the following theoretical questions.

1. What conditions the location fingerprints can be positively transferred under?
2. How good the generalization bound can be reached in the transfer learning of FL?
3. What causes the difference between the simulation model and the real data?

These questions will be further studied in our following work.

Author Contributions: Conceptualization, S.B. and Y.L.; methodology, S.B.; software, S.B.; validation, S.B.; formal analysis, S.B.; investigation, Y.L.; resources, Q.W.; data curation, Y.L.; writing—original draft preparation, S.B.; writing—review and editing, S.B. and Y.L.; visualization, S.B.; funding acquisition, Q.W. All authors have read and agreed to the published version of the manuscript.

References

1. Wang, B.; Xu, Q.; Chen, C.; Zhang, F.; Liu, K.J.R. The Promise of Radio Analytics: A Future Paradigm of Wireless Positioning, Tracking, and Sensing. *IEEE Signal Process. Mag.* **2018**, *35*, 59–80. [CrossRef]

2. Kokkinis, A.; Kanaris, L.; Liotta, A.; Stavrou, S. RSS Indoor Localization Based on a Single Access Point. *Sensors* **2019**, *19*, 3711. [CrossRef] [PubMed]

3. Hao, Z.; Yan, Y.; Dang, X.; Shao, C. Endpoints-Clipping CSI Amplitude for SVM-Based Indoor Localization. *Sensors* **2019**, *19*, 3689. [CrossRef] [PubMed]

4. Zhang, C.; Qin, N.; Xue, Y.; Yang, L. Received Signal Strength-Based Indoor Localization Using Hierarchical Classification. *Sensors* **2020**, *20*, 1067. [CrossRef] [PubMed]

5. Yuan, Z.; Zha, X.; Zhang, X. Adaptive Multi-Type Fingerprint Indoor Positioning and Localization Method Based on Multi-Task Learning and Weight Coefficients K-Nearest Neighbor. *Sensors* **2020**, *20*, 5416. [CrossRef] [PubMed]

6. Shrey, S.B.; Hakim, L.; Kavitha, M.; Kim, H.W.; Kurita, T. Transfer Learning by Cascaded Network to identify and classify lung nodules for cancer detection. In *International Workshop on Frontiers of Computer Vision*; Springer: Singapore, 2020.

7. Hudson, A.; Gong, S. Transfer Learning for Protein Structure Classification at Low Resolution. *arXiv* **2020**, arXiv:2008.04757.

8. Ye, H.; Tan, Q.; He, R.; Li, J.; Ng, H.T.; Bing, L. Feature Adaptation of Pre-Trained Language Models across Languages and Domains with Robust Self-Training. *arXiv* **2020**, arXiv:2009.11538.

9. Zhu, Q.; Xu, Y.; Wang, H.; Zhang, C.; Han, J.; Yang, C. Transfer Learning of Graph Neural Networks with Ego-graph Information Maximization. *arXiv* **2020**, arXiv:2009.05204.

10. Zhu, Y.; Xi, D.; Song, B.; Zhuang, F.; Chen, S.; Gu, X.; He, Q. Modeling Users' Behavior Sequences with Hierarchical Explainable Network for Cross-domain Fraud Detection. *Proc. Web Conf.* **2020**, 928–938. [CrossRef]

11. Safari, M.S.; Pourahmadi, V.; Sodagari, S. Deep UL2DL: Channel Knowledge Transfer from Uplink to Downlink. *arXiv* **2018**, arXiv:1812.07518.

12. Yin, J.; Yang, Q.; Ni, L.M. Learning Adaptive Temporal Radio Maps for Signal-Strength-Based Location Estimation. *IEEE Trans. Mob. Comput.* **2008**, *7*, 869–883. [CrossRef]

13. Pan, S.J.; Kwok, J.T.; Yang, Q.; Pan, J.J. Adaptive Localization in a Dynamic WiFi Environment through Multi-View Learning. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), Vancouver, BC, Canada, 22–26 July 2007.

14. Zheng, V.; Xiang, E.; Yang, Q.; Shen, D. Transferring Localization Models over Time. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), Chicago, IL, USA, 13–17 July 2008.

15. Sun, Z.; Chen, Y.; Qi, J.; Liu, J. Adaptive Localization through Transfer Learning in Indoor Wi-Fi Environment. In Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications, San Diego, CA, USA, 11–13 December 2008.

16. Zheng, V.; Pan, S.; Yang, Q.; Pan, J. Transferring MultiDevice Localization Models Using Latent Multi-Task Learning. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), Chicago, IL, USA, 13–17 July 2008.

17. Pan, S.J.; Shen, D.; Yang, Q.; Kwok, J.T. Transferring Localization Models Across Space. In Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI), Chicago, IL, USA, 13–17 July 2008.

18. Guo, X.; Wang, L.; Li, L.; Ansari, N. Transferred Knowledge Aided Positioning via Global and Local Structural Consistency Constraints. *IEEE Access* **2019**, *7*, 32102–32117. [CrossRef]

19. Peyré, G.; Cuturi, M. Computational Optimal Transport: With Applications to Data Science. *Found. Trends® Mach. Learn.* **2019**, *11*, 355–607. [CrossRef]

20. Bahl, P.; Padmanabhan, V.N. RADAR: An in-building RF-based user location and tracking system. In Proceedings of the IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064), Tel Aviv, Israel, 26–30 March 2000.

21. Lin, T.; Fang, S.; Tseng, W.; Lee, C.; Hsieh, J. A Group-Discrimination-Based Access Point Selection for WLAN Fingerprinting Localization. *IEEE Trans. Veh. Technol.* **2014**, *63*, 3967–3976. [CrossRef]

22. Yim, J. Introducing a decision tree-based indoor positioning technique. *Expert Syst. Appl.* **2008**, *34*, 1296–1302. [CrossRef]

23. Yu, L.; Laaraiedh, M.; Avrillon, S.; Uguen, B. Fingerprinting localization based on neural networks and ultra-wideband signals. In Proceedings of the 2007 International Symposium on Signals, Systems and Electronics, Montreal, QC, Canada, 30 July–2 August 2007.

24. Youssef, M.; Agrawala, A. The Horus WLAN Location Determination System. In Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services, Seattle, WA, USA, 6–8 June 2005; Association for Computing Machinery: New York, NY, USA, 2005. [CrossRef]

25. Mirowski, P.; Whiting, P.; Steck, H.; Palaniappan, R.; MacDonald, M.; Hartmann, D.; Ho, T.K. Probability kernel regression for WiFi localisation. *J. Locat. Based Serv.* **2012**, *6*, 81–100. [CrossRef]

26. Xiao, Z.; Wen, H.; Markham, A.; Trigoni, N. Lightweight map matching for indoor localisation using conditional random fields. In Proceedings of the 13th International Symposium on Information Processing in Sensor Networks IPSN 2014, Berlin, Germany, 15–17 April 2014.

27. Pan, S.J.; Tsang, I.W.; Kwok, J.T.; Yang, Q. Domain Adaptation via Transfer Component Analysis. *IEEE Trans. Neural Netw.* **2011**, *22*, 199–210. [CrossRef]

28. Long, M.; Wang, J.; Ding, G.; Sun, J.; Yu, P.S. Transfer Feature Learning with Joint Distribution Adaptation. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013.

29. Wang, J.; Chen, Y.; Feng, W.; Yu, H.; Huang, M.; Yang, Q. Transfer Learning with Dynamic Distribution Adaptation. *ACM Trans. Intell. Syst. Technol.* **2020**, *11*. [CrossRef]

30. Courty, N.; Flamary, R.; Tuia, D.; Rakotomamonjy, A. Optimal Transport for Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1853–1865. [CrossRef]

31. Ben-David, S.; Blitzer, J.; Crammer, K.; Pereira, F. Analysis of Representations for Domain Adaptation. In Proceedings of the 19th International Conference on Neural Information Processing Systems, Cambridge, MA, USA, 4–9 December 2006. [CrossRef]

32. Borgwardt, K.M.; Gretton, A.; Rasch, M.J.; Kriegel, H.P.; Schölkopf, B.; Smola, A.J. Integrating structured biological data by Kernel Maximum Mean Discrepancy. *Bioinformatics* **2006**, *22*, e49–e57. [CrossRef]

33. Dantzig, G.B. Application of the simplex method to a transportation problem. *Act. Anal. Prod. Alloc.* **1951**, *13*, 359–373.

34. Knight, P.A. The Sinkhorn–Knopp Algorithm: Convergence and Applications. *Siam J. Matrix Anal. Appl.* **2008**, *30*, 261–275. [CrossRef]

35. Reich, S. A non-parametric ensemble transform method for Bayesian inference. *arXiv* **2012**, arXiv:1210.0375.

36. Perrot, M.; Courty, N.; Flamary, R.; Habrard, A. Mapping Estimation for Discrete Optimal Transport. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Red Hook, NY, USA, December 2006. [CrossRef]

37. Bredies, K.; Lorenz, D.A.; Maass, P. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Comput. Optim. Appl.* **2009**, *42*, 173–193. [CrossRef]

38. Tseng, P. Convergence of a Block Coordinate Descent Method for Nondifferentiable Minimization. *J. Optim. Theory Appl.* **2001**, *109*, 475–494. [CrossRef]

39. Flamary, R.; Courty, N. POT Python Optimal Transport Library. Available online: https://buildmedia. readthedocs.org/media/pdf/pot/autonb/pot.pdf (accessed on 6 December 2020).

40. Wang, Z.; Dai, Z.; Póczos, B.; Carbonell, J. Characterizing and Avoiding Negative Transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.

41. Patwari, N.; Hero, A.O.; Perkins, M.; Correal, N.S.; O'Dea, R.J. Relative location estimation in wireless sensor networks. *IEEE Trans. Signal Process.* **2003**, *51*, 2137–2148. [CrossRef]

42. De Luca, D.; Mazzenga, F.; Monti, C.; Vari, M. Performance Evaluation of Indoor Localization Techniques Based on RF Power Measurements from Active or Passive Devices. *Eurasip J. Appl. Signal Process.* **2006**, *2006*, 074796. [CrossRef]

43. Knott, M.; Smith, C.S. On the optimal mapping of distributions. *J. Optim. Theory Appl.* **1984**, *43*, 39–49. [CrossRef]

44. Mansour, Y.; Mohri, M.; Rostamizadeh, A. Domain Adaptation: Learning Bounds and Algorithms. *arXiv* **2009**, arXiv:0902.3430.

45. Mendoza-Silva, G.; Richter, P.; Torres-Sospedra, J.; Lohan, E.; Huerta, J. Long-Term WiFi Fingerprinting Dataset for Research on Robust Indoor Positioning. *Data* **2018**, *3*, 3. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.