IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 June 2022, accepted 24 June 2022, date of publication 1 July 2022, date of current version 7 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3187837

RESEARCH ARTICLE

DNN-Based Indoor Localization Under Limited Dataset Using GANs and Semi-Supervised Learning

WAFA NJIMA[1], (Member, IEEE), AHMAD BAZZI[2], AND MARWA CHAFII[2,3], (Member, IEEE)
[1]ISEP, Institut Supérieur d'Electronique de Paris, 75006 Paris, France
[2]Engineering Division, New York University (NYU), Abu Dhabi, United Arab Emirates
[3]NYU WIRELESS, NYU Tandon School of Engineering, Brooklyn, NY 11201, USA

Corresponding author: Wafa Njima (wafa.njima@isep.fr)

ABSTRACT Indoor localization techniques based on supervised learning deliver great performance accuracy while maintaining low online complexity. However, such systems require massive amounts of data for offline training, which necessitates costly measurements. The essence of this paper is twofold with the purpose of providing solutions to missing data of different nature: available unlabeled data and missing unlabeled data. In both cases, we rely on a few labeled available data, which is costly yet insufficient to achieve a high localization accuracy. To address the problem of available unlabeled data, a weighted semi-supervised DNN-based indoor localization approach leveraging pseudo-labeling methods in combination with real labeled samples and inexpensive pseudo-labeled samples is proposed in order to boost localization accuracy, while overcoming the high cost of collecting additional labeled data. As for the extreme case of unavailable unlabeled data, we propose an alternative localization system generating fake fingerprints based on generative adversarial networks (GANs) named 'Weighted GAN based indoor localization'. Furthermore, a deep neural network is trained on a mixed dataset containing both real collected and fake produced data samples using a similar weighting technique in order to improve location prediction performance and avoids overfitting. In terms of localization accuracy, our proposed localization approaches outperform conventional supervised localization schemes utilizing the same collection of real labeled samples. We have tested our proposed methods on both simulated data and experimental data from the publicly available UJIIndoorLoc database, which is built to test indoor positioning systems relying on Wi-Fi fingerprints. Results based on experimental data provide the localization accuracy increase compared to the classical supervised learning method using the same set of labeled collected data when using the weighted semi-supervised and the weighted-GAN approaches by 10.11 % and 8.53 %, respectively.

INDEX TERMS Deep neural network (DNN), generative adversarial network (GAN), indoor localization, received signal strength (RSS), supervised learning, semi-supervised learning.

I. INTRODUCTION

Several localization based applications have been proposed in mobile communication systems [1], [2] such as enhanced emergency localization, personal navigation, and social networking. Different localization methods have been developed for 5G and, more specifically, for internet of things (IoT) applications [3], where it is imperative for users to receive autonomous and accurate navigation services in challenging surroundings. In the near future, larger bandwidths and higher frequencies will be provided by beyond 5G systems, offering enhanced opportunities for achieving accurate location information [4], [5]. Traditional indoor localization systems [6] are mainly based on geometric and fingerprinting-based methods. Using such methods, the localization accuracy is heavily affected by geometric constraints introduced by multipath propagation and can be a high consuming task in terms of time and energy. As an alternative to traditional

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

schemes, machine learning (ML) based-techniques shift the online complexity to an offline phase [7]–[10]. The ML-based localization model trained offline is stored and used online to predict the location information with low computational complexity.

In particular, deep learning (DL) methods, already provide a variety of advanced localization systems with high accuracy [11], [12]. However, they are data hungry requiring large labeled training databases. To overcome this problem, recent approaches have been developed based on semi-supervised learning, which leverages a small number of expensive labeled data combined with a large amount of inexpensive unlabeled data to ensure an improvement and a refinement of a totally supervised solution without an additional expensive labeled data collection cost [13], [14]. Pseudo-labels are predicted and associated to unlabeled data in order to provide additional training information and enlarge the training dataset. However, unlabeled data may not be always available. For this scenario, where only a small amount of labeled data is available, data augmentation can be used to extend the training database. Accordingly, new fake data are generated to supplement real collected data in order to enhance the model training and boost the localization accuracy while reducing measurement time and human effort.

A. RELATED WORK

A generic class of fingerprinting methods [15], [16] rely on the spatial-temporal characteristics of the various fingerprints, thus exploiting delay and angular profiles of multipaths at given positions. Although such an approach maximizes discrimination at different positions, they require proper antenna array calibration, as well as accurate timing synchronization, in order to claim a healthy database for online use. The work in [17], [18] propose RSSI fingerprints as data to train an ML model. Even though the work relies on RSSI fingerprints, the work done is fully-supervised (ex. [17] uses kNN and Random Forests). In addition, the main essence of [17] and [18] is the extraction of important features from RSSI data via Principle Component Analysis and Kernel Direct Discriminant Analysis, respectively. Furthermore, the work in [19] focuses on localization on a room-level instead of an accurate coordinate level, where 6 classifiers were used to predict the user in 4 different rooms. Cited methods did not address the problem of insufficient data available for localization. The work in [20] relies on a matrix completion method to construct complete training maps via available RSSI fingerprints. Although [20] fills missing data, the solution lacks the capability of leveraging unlabeled data that are easily available at our disposal. Even more, during runtime, the method in [20] should solve a convex optimization problem, which is deemed heavy for online applications.

As already mentioned, semi-supervised learning has been used to overcome the problem of the large amount of training data needed in the classical supervised learning. To adapt the semi-supervised context to the localization field, manifold-based models, such as manifold learning [21] and manifold alignment [22], are combined with graph-based methods [23]. Classical semi-supervised methods use a supervised model trained on a small amount of labeled data in order to predict unknown labels referred to as *pseudo-labels*. The resulting pseudo-labeled samples are used subsequently to enlarge the labeled data set, i.e., providing additional training information, to build a more general model. Pseudo-labels can be determined based on solving optimization problems as described in [24]–[27] or applying a supervised DL model on the labeled data as described in [28]. However, pseudo-labeled data can penalize the performance if they are not introduced to the training in an appropriate way. Also, the predicted labels may be noisy or may not really reflect the ground truth. Therefore, it is desirable to limit the reliance on pseudo-labeled data. Such principle has been applied to the classification problem of handwritten digit recognition in [29]. The unlabeled data collection task is less expensive than labeled data collection, but it is not always available. In such a case, data augmentation can be used to extend the training database.

Generative models have shown a good ability to generate additional realistic samples. In particular, generative adversarial networks (GANs) [30] aim at expanding and improving the diversity of a training database in different research fields including indoor localization. In [31], GANs use both labeled and unlabeled data when the former is insufficient in order to share weights with a localization classifier to benefit from useful information contained in the latter. In [32], [33], GANs are used to improve the diversity of the collected database generating fake received signal strength indicators (RSSIs) at known positions already used for data collection. Regarding the difficulty in collecting signal measurements under indoor space constraints, the authors in [34] propose to generate artificial data for the constrained space based on measured data collected in the free space. Thus, GANs aim to enhance the richness of a collected database to cover some regions where data collection is strenuous.

B. CONTRIBUTIONS

In this paper, we propose two solutions in order to deal with the problem of collected labeled data insufficiency to optimally train a localization model. The first solution is based on weighted semi-supervised learning combining labeled and unlabeled data, and the second one explores data augmentation using GANs based on the collected labeled data only. In Table 1, we mention existing works that deal with the problem of overcoming limited data used for indoor localization compared to our proposed schemes. Our contributions are summarised as follows:

- We propose a localization method called 'weighted semi-supervised' that combines labeled and unlabeled collected data in order to boost the localization performance. First, deep neural network (DNN)-based pseudo-labeling is used to generate pseudo-labels for unlabeled data using a supervised scenario with labeled

TABLE 1. Comparison of existing approaches.

	Nature of available data		Overcoming data insufficiency method						
	Labeled Data	Unlabeled Data	Data completion	Semi-supervised learning		Data augmentation			
				Manifold model	Pseudo labeling	Diversity of data	Richness of data	Fingerprints generation	RSSI vectors generation to be pseudo-labeled
[20]	X		X						
[21 - 23]	X	X		X					
[24 - 28]	X	X			X				
[31], [35]	X	X				X		X	
[32 - 33]	X					X		X	
[34]	X						X	X	
[36]	X					X	X		X
Weighted semi-supervised	X	X			X				
W-GAN	X					X	X	X	

training data. Then, real labeled and pseudo-labeled data are mixed together to train a generic model used for localization according to a coefficient weight in order to up-weight the most confident samples to increase their impact on the training of the generic model.

- We propose a localization system 'W-GAN' generating fake supplementary data based on labeled data only. Unlike previous works, we do not assume the use of unlabeled data to enhance the training of the model as in [31], [35], and we do not assume having sufficient collected data in some regions as in [34]. Moreover, we do not use GANs to further diversify signal measurements at known positions, as considered in [32], [33]. Instead, our approach generates RSSI measurements and its corresponding new positions to cover new areas. Furthermore, we do not use GANs to generate fake RSSI vectors to be pseudo-labeled later as conducted in our previous work [36]. Such labeling process increases the computational complexity of the whole system, and the error on pseudo-labels prediction can lead to localization accuracy loss. In this paper, a GAN is used to produce artificial labeled RSSI vectors, i.e. both RSSI vectors and their corresponding coordinates. Then, the real collected and fake generated data samples are mixed to train a localization-based DNN employing coefficient weights to limit the impact of the least confident data samples, which are the GAN-generated samples, especially in the early stages of the training process.

The rest of this paper is organized as follows: The studied problem is defined and described in Section II. The proposed weighted semi-supervised based localization system is detailed in Section III. The weighted GAN-based localization system is then provided in Section IV. The obtained results are presented in Section V and Section VI based on both simulated and experimental data from the UJIIndoorLoc database [37]. We discuss obtained performances in Section VII. Finally, conclusions are drawn in Section VIII.

II. PROBLEM DESCRIPTION

In this section, we briefly detail the classical indoor localization system based on RSSI fingerprints. This system consists of two main parts: a central unit (CU) connected to mobile users (MUs) through a network. The mobile users are equipped with different mobile devices to ensure the diversity of the database, given the heterogeneity of devices that can cause signal diversity. They perform a site survey task to collect RSSI data from different access points (APs) in the indoor environment. Then, RSSI fingerprints, that are composed of collected data associated with the coordinates of the corresponding mobile user, are transferred to the CU for the RSSI-fingerprint database construction and storage. It also determines the localization of a user node (UN) given a received RSSI vector, and sends back the estimated coordinates to the user. The localization can be performed by solving a linear equation or an optimization problem [38], or using DL techniques as considered in this work.

A classical supervised DNN system is composed of an offline phase and an online phase. In the offline phase, a trained model is constructed and validated based on an exhaustive set of data. In our case, the measured data are divided into training data and validation data to train and validate the DNN model. The collected RSSI vectors present the inputs of our DNN network, which takes as outputs the associated users location information (a room ID, a floor ID, a zone identifier, 2-D / 3-D coordinates, etc.) for training. Thus, the collected data is used by the CU for localization without any required pre-processing task. Once the DNN model is trained and well optimized,

the system is able to localize a given UN based on the collected data taking as input an RSSI vector and given the estimated coordinates as outputs. In order to achieve a good localization accuracy, a large amount of labeled data samples is required to construct an efficiently trained localization model. However, the acquisition of labeled RSSI vectors is a time and cost consuming repetitive task. To solve this issue, a weighted semi-supervised indoor localization framework is first proposed in this work, which involves location estimation based on labeled and unlabeled data. This system treats mixed data to reduce the reliance on labeled data often costly to collect, unlike unlabeled data.

It is true that the acquisition cost of unlabeled data is cheaper than that of labeled ones, but the data collection cost can still be expensive and unlabeled data may not always be available. To address this issue, data augmentation based on GANs is proposed as a second approach in this work, in order to generate fake data which compliment real labeled collected data. In our previous work [36], a system combining selective GANs and semi-supervised learning is proposed to perform location prediction based on real collected data and fake selected-generated pseudo-labeled data. This system generates RSSI vectors to be pseudo-labeled from which we select the most realistic-fake pseudo-labeled positions. In this paper, the second proposed localization system is based on weighted-augmented process with GANs. It takes advantage of generating both RSSI vectors and their corresponding coordinates to be mixed with real collected measurements for localization. During this combination, a coefficient weight is associated to each measurement in order to reduce the reliance on the least realistic samples and increase the effect of the most realistic ones. This procedure is simpler and less complex than our previously proposed algorithm [36], as depicted in Figure 1, by preventing errors that occur due to the pseudo-labeling process.

In this paper, we consider M APs placed in an indoor environment and mobile sensor nodes placed at known training positions. These nodes collect T RSSI measurements with respect to each reachable AP to alleviate time-varying RSSI fluctuations. Collected fingerprints, composed of RSSI vectors associated with the corresponding coordinates, are sent to a CU to be stored and used for localization. A DNN model, trained on the training database, is applied online in order to predict the user coordinates.

III. PROPOSED WEIGHTED SEMI-SUPERVISED DNN-BASED LOCALIZATION

As mentioned above, we propose in this paper two localization systems to address the insufficiency of collected labeled data needed to optimally train a localization model. In this section, we describe the first proposed system which exploits labeled and unlabeled collected data samples as depicted in Figure 2. The classical semi-supervised indoor localization [28] is first presented, then, a detailed description of the proposed weighted semi-supervised system is provided.

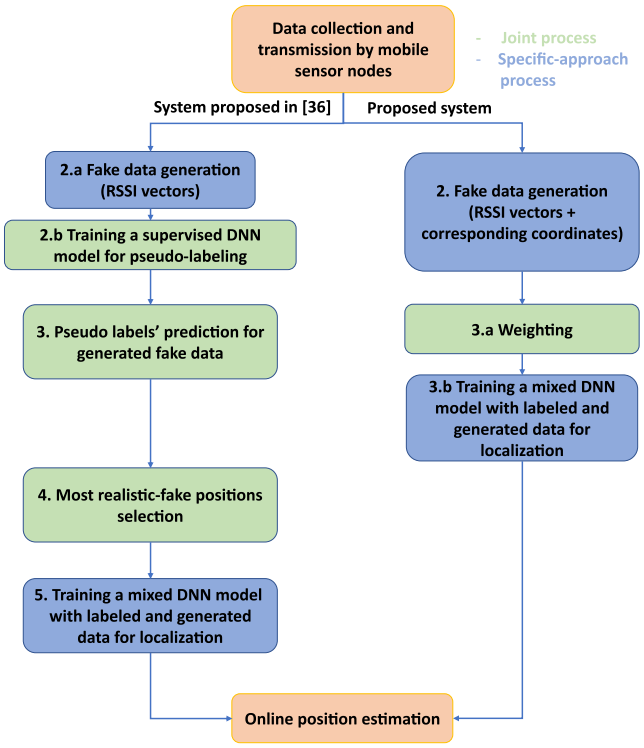


FIGURE 1. The proposed weighted GAN based localization method vs. the system proposed in [36].

A. CLASSICAL SEMI-SUPERVISED INDOOR LOCALIZATION SYSTEM

In this part, we introduce semi-supervised learning to deal with collected unlabeled RSSI vectors so as to improve the localization accuracy.

1) DATA COLLECTION

RSSI measurements are collected during the offline phase. Collected RSSI vectors are considered as labeled fingerprints when associated to the corresponding location identifier. In this paper, we consider the user coordinates as location identifier called *label*. To reduce the effort of data acquisition and environment labeling, a massive collection of unlabeled data can be performed by the acquisition of RSSI measurements when mobiles are moving in the covered indoor environment. In fact, they collect labeled data when placed at known positions. And, when moving from one position to another, they still collect unlabeled data. Thus, we consider an MU collecting RSSI data received from M APs at reference positions known by their labels and collecting data when moving to construct unlabeled database.

Let $P^t = [p_1^t, \ldots, p_m^t, \ldots, p_M^t]^T$ be the RSSI vector received by an MU at position t, where p_m^t is the RSSI measurement received from the m^{th} AP with $m \in \{1, 2, \ldots, M\}$. The coordinates or labels associated to the vector P^t are $C_r^t = [x_r^t, y_r^t]^T$ when collecting real labeled data, whereas no label is assigned to the RSSIs collected on-the-fly from massive measurements. For system evaluation purposes, we consider that $C_p^t = [x_p^t, y_p^t]^T$ are the exact values of pseudo-labels.

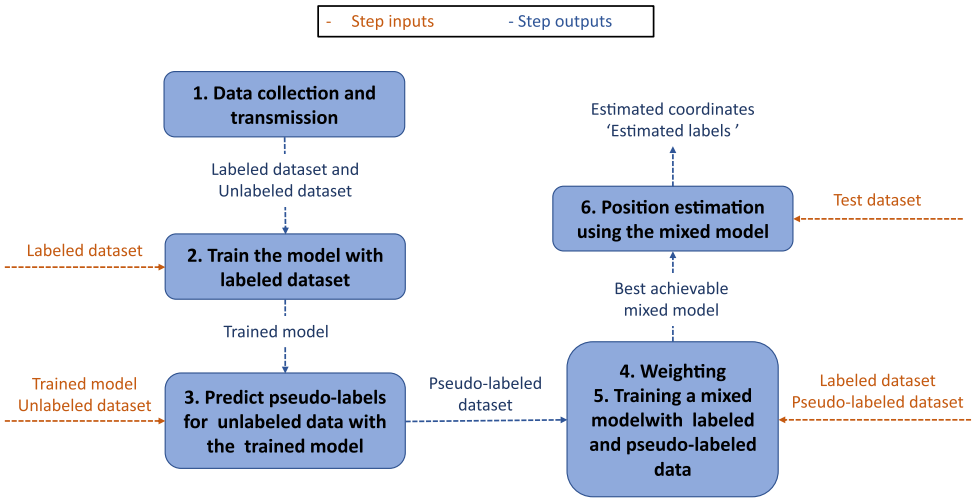


FIGURE 2. The pipeline of the proposed weighted semi-supervised based localization method during the training phase.

2) PSEUDO-LABELING FOR INDOOR LOCALIZATION

Localization based on semi-supervised learning aims to address the costs and complexity of labeled data measurements as well as to avoid the over-fitting problem for DL based localization, which can be caused by a limited number of labeled training data. The pseudo-labeling technique is a semi-supervised learning method widely used in the fields of text and image classification and recognition. Such techniques exploit both labeled and unlabeled data to improve supervised learning performance. The pseudo-labeling steps are summarized as follows:

- Step 1: Train the model on labeled RSSI fingerprints only, in a supervised way during E_1 epochs. For each input vector $P^t \in \mathbb{R}^{M\times 1}$, the associated output vector is $C_r^t \in \mathbb{R}^{2\times 1}$.
- Step 2: Using the trained model, predicted labels i.e. 'pseudo-labels', are generated for unlabeled vectors.
- Step 3: A general model is then trained mixing labeled and pseudo-labeled vectors. This model is different from the model trained in step 1 and used in step 2. To train such a general model, we use as inputs the entire collected measurements and the outputs are the labels (real labels and predicted pseudo-labels). During the training process, mixing both labeled and pseudo-labeled data is crucial in order to achieve a good localization accuracy.

The pseudo-labeling technique is simple and easy in terms of implementation with promising results. However, it may result in gradual drifts and poorly perform if the prediction accuracy of pseudo-labels is low. In fact, we cannot give the same level of trust to pseudo-labeled data, as we do with labeled ones. Therefore, limiting the reliance on pseudo-labeled data may be efficient. Thus, a process which automatically weights labels in order to down-weight less promising ones and high-weight the more promising ones is proposed.

B. PROPOSED WEIGHTED SEMI-SUPERVISED SCHEMES

To limit the reliance on artificial pseudo-labels, two types of weights have been used to train a mixed model: a static weight and a variable weight.

1) STATIC WEIGHT

We assume that all pseudo-labeled data have the same fixed weight during the whole procedure of model fitting. This weight should not exceed the labeled weight since real labels are more confident. Thus, we propose to update the losses calculated at different training steps using refined weights. For epoch e, the loss function $L_{\text{sw}}(e)$ which is a variant of a refined root mean square error (RMSE) [39] can be expressed as follows:

$$L_{\text{sw}}(e) = \sqrt{\omega_r(e)\sum_{t=1}^{T_r}\frac{\|C_r^t-\hat{C}_r^t\|^2}{T_r}+\omega_p(e)\sum_{t=1}^{T_p}\frac{\|C_p^t-\hat{C}_p^t\|^2}{T_p}}, \quad (1)$$

where C_r^t, C_p^t, $\hat{C}_r^t$ and $\hat{C}_p^t$ are the available and pseudo coordinates: real and estimated, respectively. Note that T_r is the number of labeled samples and T_p is the number of unlabeled samples. The total number of available samples is $T_1 = T_r + T_p$. Moreover, $\omega_r(e)$ and $\omega_p(e)$ stand for static weights associated with the labeled and unlabeled samples, respectively. The weights are adjusted experimentally through an exhaustive experimental process.

2) VARIABLE WEIGHT

Here, the used loss function L_{vw} has the same form as L_{sw} described in (1). However, we consider that the labeled weight $\omega_r(e)$, is normalized to 1 and we use a variable weight $\omega_p(e)$ associated with the pseudo-labels for an epoch e. A proper calibration of $\omega_p(e)$ is required to benefit from unlabeled data without disturbing the training for labeled data

to ensure a reliable training performance. Furthermore, $\omega_p(e)$, which is related to the epoch e, is slowly increased passing from 0 to a final value ω. It is described as follows:

$$\omega_p(e) = \begin{cases} \dfrac{e}{E_2} \times \omega & \text{if } e < E_2 \\ \omega & \text{if } e \geq E_2 \end{cases}, \tag{2}$$

where E_2 refers to a specific number of epochs. ω and E_2 are hyperparameters to be tuned experimentally. Moreover, we introduce the pseudo-labels in the training and we gradually increase their weights through the epochs. The value of E_2 can be greater than the total number of training epochs, provided that $w_p(e)$ does not exceed $w_r(e)$ which is equal to 1.

IV. PROPOSED WEIGHTED GAN BASED INDOOR LOCALIZATION

To enhance the richness of collected labeled training data, GANs are used to generate fake RSSI vectors and its corresponding labels (i.e. 2D coordinates) due to the fact that collecting a large amount of real location samples is costly. Thus, based on a small amount of real labeled data samples, the size and the diversity of the training dataset are increased by generating supplementary fake samples. Real collected labeled data and fake generated data are mixed in order to train a DNN model used for localization. However, generated data can penalize the accuracy if it is not properly considered. Thus, we propose to limit the dependency on generated data by associating a coefficient weight to fake samples during the training phase, as discussed in the next section.

A. INTRODUCTION TO DNN ARCHITECTURE

Let $i_{(0)} \in \mathbb{R}^{N_0 \times 1}$ be the input vector to the DNN model and $o \in \mathbb{R}^{N_{H+1} \times 1}$ its associated output [40]. Let H be the number of hidden layers where $0 \leq h \leq H+1$ and N_h is the number of neurons in the h^{th} layer. $\boldsymbol{b}_h \in \mathbb{R}^{N_h \times 1}$ and $W_h \in \mathbb{R}^{N_h \times N_{h-1}}$ denote the biases and the weights matrices, respectively. The output vector of the h^{th} layer can be expressed as:

$$o_{(h)} = g_h(b_{(h)} + W_{(h)} i_{(h-1)}), \; i_{(h)} = o_{(h)}, \tag{3}$$

where the vector $i_{(h)}$ undergoes a linear transformation represented by $W_{(h)}$, a bias vector $b_{(h)}$, and then a nonlinear activation function $g_{(h)}$ is applied. During DNN training, the loss function is calculated in order to iteratively update its parameters $\theta = (W, b)$.

B. TRAINING GANs FOR DATA AUGMENTATION

GANs have achieved promising performance across a multitude of fields. In this paper, GANs are used for data augmentation to increase the training dataset size and diversity. Such models are composed of two DNNs: the generator G and the discriminator D [41], [42]. The generator model G learns how to produce a realistic representation similar to the real data, and the discriminator model D learns how to distinguish between fake and real samples, as shown in Fig. 3. These DNN models are trained together until G

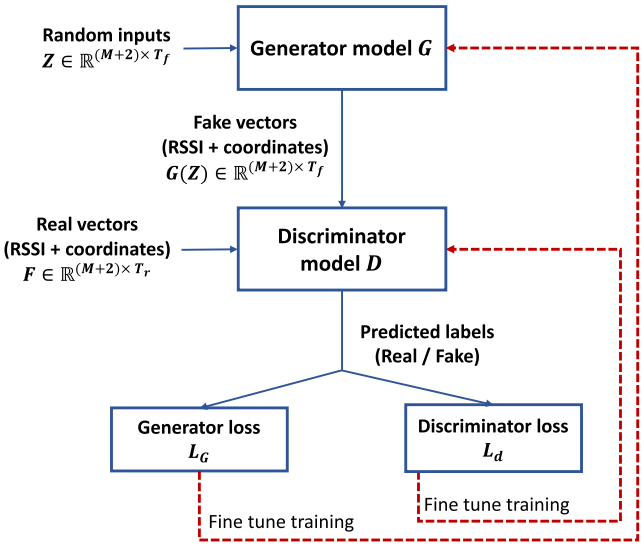


FIGURE 3. GAN network for data generation during training.

is able to generate fake samples that can be seen as real by D. Let T_f be the number of generated fake samples, and let $z^{(i)} \in \mathbb{R}^{(M+2)\times 1}$, $i = 1, \cdots, T_f$ be an input noise vector fitted to the generator, whose samples are uniformly distributed over $[-1, 1[$. The size of such vector is $(M+2, 1)$ since the goal is to generate a vector composed of M RSSIs received from M APs and the corresponding position identified by the 2D coordinates. At the output of the generator, each input noise vector $z^{(i)}$ produces a fake vector $G(z^{(i)}) \in \mathbb{R}^{(M+2)\times 1}$ following (3), where $i_{(0)} = z^{(i)}$, $N_0 = M+2$, $o_{(H+1)} = G(z^{(i)})$ and $N_{H+1} = M+2$. Then, $G(z^{(i)})$ is passed to the discriminator which predicts the reliability of $G(z^{(i)})$. The generator loss is calculated by:

$$\mathcal{L}_G = \nabla\theta_g \Big[-\frac{1}{T_f} \sum_{i=1}^{T_f} \log\Big(D\big(G(z^{(i)})\big)\Big)\Big], \tag{4}$$

where $D(G(z^{(i)}))$ refers to the probability of an example being fake. Once the loss function is minimized, the generator parameters $\theta_g = (W_g, b_g)$ are saved.

During the training, the discriminator computes the probability $D(G(z^{(i)}))$, following (3), giving both fake generated samples $G(Z)$, and the real dataset $F \in \mathbb{R}^{(M+2)\times T_r}$. Each real vector is denoted as $f^{(i)} = \text{vect}(P^{(i)}, C_r^{(i)})$, $i = 1, \ldots, T_r$ where $P^{(i)} \in \mathbb{R}^{M\times 1}$ is the i^{th} RSSI vector and $C_r^{(i)} = (x_r^{(i)}, y_r^{(i)})$ are the corresponding 2D coordinates. This probability is given to the generator to improve its performance as expressed in (4). The training of the discriminator is performed by minimizing the loss function between real and fake data:

$$\mathcal{L}_d = \nabla\theta_d \Big[-\frac{1}{T_r} \sum_{i=1}^{T_r} \log\big(D(f^{(i)})\big) - \frac{1}{T_f} \sum_{i=1}^{T_f} \log\Big(1 - D\big(G(z^{(i)})\big)\Big)\Big], \tag{5}$$

where $\theta_d = (W_d, b_d)$ are the parameters of the discriminator model. Maximizing $\log(D(f^{(i)}))$ refers to the fact that the discriminator is correctly classifying the real examples while maximizing $\log\left(1 - D\left(G(z^{(i)})\right)\right)$ would help the discriminator to correctly classify the fake samples that come from the generator.

C. TRAINING A MIXED WEIGHTED DNN FOR LOCALIZATION

Localization is conducted applying a mixed DNN model combining collected and generated data samples. The losses are updated at different training steps using refined variable coefficient weights to limit the reliance on fake generated positions. For epoch e, the loss function $L_{\text{vw}}(e)$ used for model fitting takes the same format as (1) and can be expressed as follows:

$$L_{\text{vw}}(e) = \sqrt{\omega_r(e)\sum_{t=1}^{T_r}\frac{\|C_r^t - \hat{C}_r^t\|^2}{T_r} + \omega_f(e)\sum_{t=1}^{T_f}\frac{\|C_f^t - \hat{C}_f^t\|^2}{T_f}}, \quad (6)$$

where C_r^t, C_f^t are the collected and fake coordinates respectively, while $\hat{C}_r^t$ and $\hat{C}_f^t$ refers to the predicted ones. Note that T_r is the number of real collected samples and T_f is the number of generated fake samples. The total number of available samples is $T_2 = T_r + T_f$. Moreover, $\omega_r(e)$ and $\omega_f(e)$ stand for variable weights associated with the collected and generated samples, respectively. We consider that the real weight ω_r, is normalized to 1 for all epochs, while the weighting function $\omega_f(e)$ is a piece-wise linear function of epoch e, and takes values from 0 to ω as expressed in (2).

V. SIMULATION RESULTS

In this section, we provide an evaluation of the two proposed localization schemes by specifying a common simulation environment as well as different used DNN architectures and adjusted parameters followed by the corresponding localization accuracy.

A. ENVIRONMENTAL SETUP

We consider a noisy indoor environment covering 400 m^2 with an existing WiFi infrastructure and $M = 10$ APs. We use a propagation model with realistic parameters based on real measurements conducted in an indoor environment. In this model, we consider the degradation of signals, combining path loss and shadowing effects, and the signals blockage. Let p_m^t be the RSSI measured by an MU at position t of the signal transmitted by the m^{th} AP. It can be expressed as:

$$p_m^t = p_e - p_{L_{mt}} + B_{\sigma_{mt}} \ [\text{dBm}], \quad (7)$$

where p_e is the transmitted power, which is considered constant. $B_{\sigma_{mt}}$ is a Gaussian random variable representing the shadowing effects, and $p_{L_{mt}}$ is the path loss calculated as follows:

$$p_{L_{mt}} = p_{L_0} + 20\log_{10}(f) + 10\mu\log_{10}(\frac{d_{mt}}{d_0}), \quad (8)$$

where p_{L_0} denotes the pathloss value at a reference distance d_0, f is the frequency, μ is the pathloss exponent and d_{mt} is the distance between the position t and m^{th} AP. These experiments are performed using $p_e = 20$ dBm, $d_0 = 1$ m, $f = 2.4$ GHz, $\mu = 3.23$ and $B_{\sigma_{mt}} \sim \mathcal{N}(0, 4)$. As already mentioned, we consider signal blockage when modeling the environment to reflect more realistic propagation conditions, where at each measurement position, the signal transmitted by each AP is not detected due to the limitation of the communication range and other signal propagation constraints. Thus, we assume that the weakest 40% of RSSIs are unknown. The choice of these values is based on several experiments conducted in our indoor environment. At each position, 10 RSSI measurements are collected in order to minimize temporal RSSI fluctuations caused by indoor signal propagation limitations (shadowing and fading effects).

B. SIMULATION RESULTS FOR THE WEIGHTED SEMI-SUPERVISED DNN-BASED LOCALIZATION

T_r labeled RSSI training measurements taken at M_r labeled positions have been collected from different APs composing a labeled database of size $(T_r \times M + 2)$. The unlabeled dataset of size $(T_p \times M)$ contains T_p unlabeled RSSI vectors. We consider different indoor scenarios with different numbers of labeled and unlabeled samples: $S_1 = \{T_r = 1000, M_r = 100, T_p = 3000\}$, $S_2 = \{T_r = 1000, M_r = 100, T_p = 5000\}$ and $S_3 = \{T_r = 3000, M_r = 300, T_p = 3000\}$. The test database contains 6000 RSSI vectors collected at 600 test positions placed randomly in the studied area. These test positions are different from the training positions, while using the same propagation model.

1) DNN ARCHITECTURES AND ADJUSTED HYPERPARAMETERS

The architectures have been identified to satisfy a trade-off between localization accuracy and online complexity based on several experiments. Therefore, different architectures have been used to train the data using different values of the hyperparameters: number of layers, number of neurons in each layer, number of epochs, mini batch size and learning rate. As DNN inputs, we have an input RSSI vector of size (10×1) and the corresponding (2×1) coordinates as outputs. Unknown RSSI values have been replaced by -110 dBm which is a value chosen experimentally. We use adaptative moment estimation (ADAM) as the optimization algorithm [43] and a learning rate equal to 0.01 has been selected. The intensive experiments have led to a DNN architecture of 2-hidden layers model with 30 neurons in the first layer and 15 neurons in the second one, concerning the first model used for pseudo-labeling using $E = 200$ and a mini batch size equal to 50.

TABLE 2. Mixed DNN model's parameters considering a classical semi-supervised scenario and weighted semi-supervised scenarios.

Indoor scenario	Parameters for the classical semi-supervised scenario	Parameters for the (static) weighted semi-supervised	Parameters for the (variable) weighted semi-supervised
S_1	$E = 300$, $B = 100$	$E = 300$, $B = 100$, $w_r = 1$ and $w_p = 0.2$	$E = 150$, $B = 100$, $\omega = 3$ and $E_2 = 550$
S_2	$E = 300$, $B = 100$	$E = 300$, $B = 100$, $w_r = 2$ and $w_p = 0.4$	$E = 50$, $B = 100$, $\omega = 3$ and $E_2 = 450$
S_3	$E = 300$, $B = 50$	$E = 300$, $B = 50$, $w_r = 1$ and $w_p = 0.6$	$E = 270$, $B = 50$, $\omega = 2$ and $E_2 = 570$

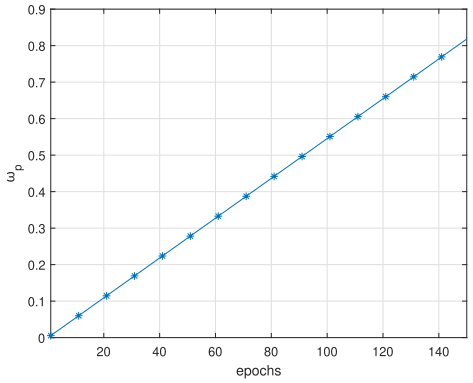


FIGURE 4. The variation of ω_p with 150 epochs considering the first scenario (1000 labeled samples and 3000 unlabeled samples) during the DNN mixed model training.

For each scenario (i.e. S_1, S_2 and S_3), different DNN architectures and parameters have been tested to find a good model for both classical and weighted semi-supervised learning methods. We achieved a good performance, using a simple DNN architecture with two hidden layers having 30 neurons and 15 neurons, respectively. Also, different parameters have been tested to select the appropriate static weight, we have tried $\omega_r \in \{1, 2, 3, 4\}$, and ω_p has been varied from 0.1 to 1. Different optimal parameters are summarized in Table 2. where E is the number of epochs and B is the mini batch size considered for the DNN training.

For the weighted method using variable weights, we have conducted intensive experiments to select the hyperparameters. For example, we set $\omega = 3$ and $E_2 = 550$ for 1000 labeled and 3000 unlabeled samples as mentioned in Table 2. As depicted in Figure 4, these variable weights increase gradually from 0 to 0.81, considering 150 epochs for system training. Thus, labeled data is combined with unlabeled data weighted related to the epoch until reaching 0.81. By exceeding this variable weight, the localization accuracy starts to decrease. Thus, it is defined as the maximum value to reach.

2) LOCALIZATION ACCURACY

To evaluate our proposed weighted semi-supervised methods (denoted as fixed-weight semi-supervised method and variable-weight semi-supervised method), we compare it with the state of the art supervised (denoted as supervised method) and classical semi-supervised DNN methods (denoted as classical semi-supervised method). Thus, we compare the mean error in the user coordinates estimation, of the following methods:

- Supervised based localization method considering only labeled data.
- Classical semi-supervised using a pseudo-labeling process to determine pseudo-labels and combine them with real labels to construct a generic localization method.
- Fixed-weight semi-supervised using a static weight.
- Variable-weight semi-supervised by integrating a variable weighting process to the classical method.

Table 3 and Figure 5 present the localization performance obtained by the cited methods considering 1000 labeled samples and 5000 unlabeled samples. We notice that the use of 1000 labeled data in a supervised way gives the worst results. The combination of 1000 labeled data with 5000 unlabeled data improves the localization accuracy by 37 cm minimizing the localization error by 26.42%. Adding a weight coefficient to the classical semi-supervised method is always beneficial, improving the localization accuracy by 3.88% and 12.62% for static and variable weight, respectively. Thus, our two proposed methods improve the localization performance of the state of the art methods. We mention that the accuracy 0.9 m obtained when using the variable weight for 1000 labeled data and 5000 unlabeled data requires 2000 labeled data samples using supervised model as mentioned in Table 4. This shows that we can reduce the cost of collecting labeled data by half while achieving the same accuracy. From Table 4, we can observe that the use of unlabeled data boosts the localization accuracy and minimizes the cost involved in collecting labeled data. In particular, mixing 3000 labeled data samples and 3000 unlabeled data samples in variable weighted semi-supervised framework provides the same localization accuracy, which can require 6000 labeled data in a classical supervised system.

C. SIMULATION RESULTS FOR THE WEIGHTED GAN

In this part, we consider a test database (T_t,M_t) containing T_t test RSSIs vectors taken at M_t test positions and a training database (T_r,M_r) where M_r training real positions have been used to collect T_r RSSI vectors for training. In Table 5, we present the number of different types of simulation data.

1) DNN ARCHITECTURES AND PARAMETERS USED FOR DATA AUGMENTATION AND LOCALIZATION

GANs are used to generate T_f fake RSSI data samples along with their M_f fake coordinates. In Figure 6, we consider $T_f = 2000$ and $T_r = 1000$. The GANs introduced in this

TABLE 3. Obtained localization performance considering 1000 labeled data samples and 5000 unlabeled data.

	Localization error (m)	Accuracy increase vs supervised	Accuracy increase vs classical semi-supervised
Supervised	1.4	–	–
Classical semi-supervised	1.03	37cm | 26.42%	–
Fixed-weight semi-supervised	0.99	41 cm | 29.28%	4 cm | 3.88%
Variable-weight semi-supervised	0.9	50 cm | 35.7%	13 cm | 12.62%

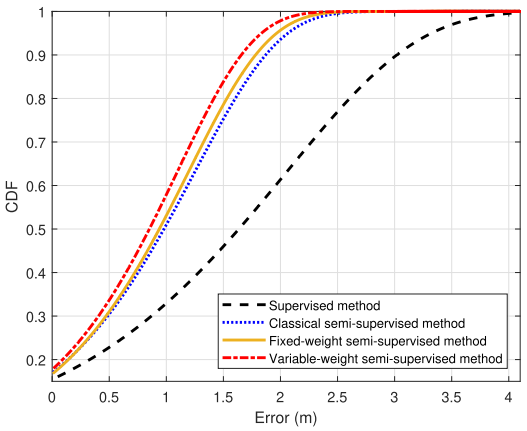


FIGURE 5. Localization performance comparison using 1000 labeled data samples and 5000 unlabeled data samples.

TABLE 4. Localization error in meters using the supervised learning and the variable weight semi-supervised learning [m].

Number of samples	Localization error (m)
$T_r = 1000$	1.4
$T_r = 2000$	**0.9**
$T_r = 3000$	**0.8**
$T_r = 6000$	**0.71**
$T_r = 7000$	0.70
$T_r = 1000$ - $T_p = 3000$	0.94
$T_r = 1000$ - $T_p = 5000$	**0.9**
$T_r = 2000$ - $T_p = 4000$	**0.8**
$T_r = 3000$ - $T_p = 3000$	**0.71**

part are based on a DNN model optimized with ADAM using 0.01 as the learning rate during 200 epochs. The activation function used by G is the rectified linear unit (ReLU) [44] used in one hidden layer having 10 neurons. The activation function of D is the ReLU function, while the last layer uses the sigmoid function. A one-hidden layer discriminator with 10 neurons is used. For localization, the DNN models are trained on real labeled data samples and weighted fake

TABLE 5. The number of data used during simulations.

Parameter	Definition	Value
T_r	Number of real labeled samples	1000
$M_f = T_f$	Number of fake positions	$[100 - 5000]$
T_t	Number of test measurements	8000
M_t	Number of test positions	800
M_r	Number of real positions	100

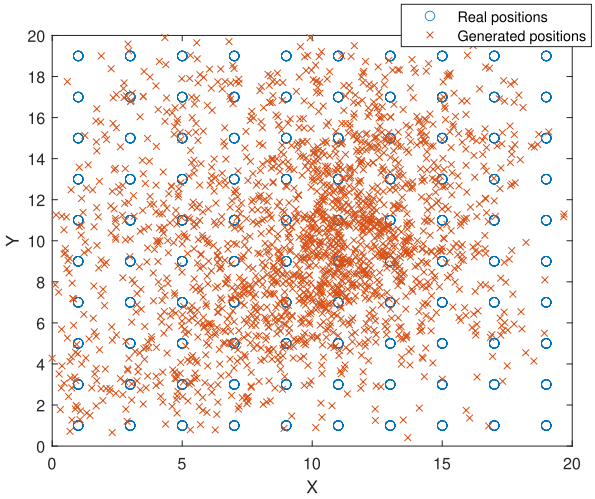


FIGURE 6. 2000 Generated fake positions based on 100 real training positions.

labeled data samples. These models take the RSSI vectors as inputs and give the corresponding coordinates as outputs. We use ADAM optimization algorithm and a learning rate equal to 0.01. The DNN architectures and parameters are summarized in Table 6 where $N_i(\cdot)$ refers to the number of neurons in the i^{th} hidden layer. Note that our model converges before reaching E_2 with a final weight value between 0.6 and 0.96.

2) LOCALIZATION ACCURACY

In this section, we compare the following algorithms for different parameters values:

- Supervised(T_r,M_r) is a supervised localization method based on T_r real data collected at M_r different positions.
- W-GAN(T_r,M_r,T_f) is the localization method, where we combine T_r real samples collected at M_r different positions, and T_f fake weighted samples.

In Table 6, we present the localization accuracy for W-GAN(T_r,M_r,T_f) trained over $T_r = 1000$ real data samples in addition to different numbers of fake data samples vs. supervised learning model trained only on $T_r = 1000$ real samples. We generate different numbers of fake data samples $[100 - 5000]$. We notice that for all augmented training datasets, the localization accuracy is improved compared to a dataset limited only to real data. The localization improvement varies between 17.92% and 23.58%. The best localization accuracy is obtained with $T_f = 2000$ weighted

TABLE 6. Obtained localization performance considering 1000 labeled data samples.

	Localization accuracy	Accuracy increase vs supervised(1000,100)	DNN used architecture	DNN training parameters
Supervised(1000,100)	1.06 m	–	$N_1(30)$ and $N_2(20)$	E = 150 and B= 50
Supervised(2000,1100)	0.81 m	25 cm | 23.58%	$N_1(30)$ and $N_2(20)$	E = 150 and B= 50
Supervised(3000,2100)	0.77 m	29 cm | 27.35%	$N_1(50)$, $N_2(20)$ and $N_3(10)$	E = 200 and B = 50
W-GAN(1000,100,100)	0.87 m	19 cm | 17.92%	$N_1(30)$ and $N_2(15)$	E = 200, B = 50, ω= 3 and E_2= 590
W-GAN(1000,100,250)	0.84 m	22 cm | 20.75%	$N_1(30)$ and $N_2(15)$	E = 200, B = 50, ω= 3 and E_2= 590
W-GAN(1000,100,1000)	0.83 m	23 cm | 21.96%	$N_1(60)$, $N_2(40)$ and $N_3(20)$	E = 200, B= 50, ω= 3 and E_2= 500
W-GAN(1000,100,2000)	**0.81** m	25 cm | 23.58%	$N_1(50)$, $N_2(30)$ and $N_3(10)$	E = 200, B = 100, ω= 3 and E_2= 500
W-GAN(1000,100,3000)	0.84 m	22 cm | 20.75%	$N_1(60)$, $N_2(40)$ and $N_3(20)$	E = 200, B= 100, ω= 3 and E_2= 500
W-GAN(1000,100,4000)	0.84 m	22 cm | 20.75%	$N_1(60)$, $N_2(40)$ and $N_3(20)$	E = 200, B = 100, ω= 3 and E_2= 500
W-GAN(1000,100,5000)	0.85 m	21 cm | 19.81%	$N_1(60)$, $N_2(40)$ and $N_3(20)$	E = 200, B = 100, ω= 3 and E_2= 500

generated fake data samples, where we achieve 23.58% localization accuracy increase vs. the conventional supervised algorithm without any additional cost in collecting additional data. This improvement is explained by the consideration that the DNN has been trained over a larger dataset which contains new positions that are not included in the limited dataset constructed from only measured data. Starting from 3000 weighted generated samples, the performance saturates and no improvement can be achieved by generating additional fake samples. This can be explained by the fact that based on 1000 real vectors collected at 100 positions, we cannot provide a higher measurement diversity to the GAN.

We can easily notice that the supervised indoor localization system based on 1000 real samples collected at 100 known positions corresponds to the worst localization accuracy. For fair comparison, we use the same dataset of real labeled positions to which we add (i) 1000 real measurements collected at 1000 different real positions placed randomly in the considered area i.e. Supervised(2000,1100), (ii) 2000 real measurements collected at 200 different positions placed randomly in the considered area i.e. Supervised(3000,2100) and (iii) based on these data samples, we generate 2000 fake positions i.e. W-GAN(1000,100,2000). We consider that 1000 real measurements collected at 100 positions construct the initial dataset. Instead of collecting 1000 extra real measurements at 1000 positions i.e. Supervised(2000,1100), we can achieve the same performance 0.81 by artificially generating 2000 fake data samples i.e. W-GAN(1000,100,2000) based on the initial real data samples. Thus, the proposed data generation process provides an improvement of localization accuracy without additional data collection cost. We notice that if we assume having 3000 real data samples collected uniformly at 300 positions, we can only improve the proposed localization scheme by 4 cm while the required number of collected data is multiplied by 3.

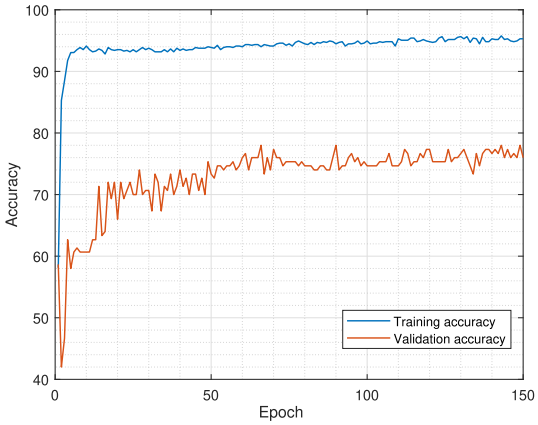
To get more insight into the presented results, we show in Figure 7a and Figure 7b the training and validation accuracy of the DNN model for Supervised(1000,100) and W-GAN(1000,100,2000). As we can see, relying on a small set of data leads to overfitting i.e. 95% training accuracy vs. 78% validation accuracy, while this issue is eliminated when using additional fake data i.e. 98% training accuracy vs. 94% validation accuracy. Figure 7b shows that starting from $E = 100$, which is the epoch where we start progressively introducing the fake data, the training accuracy gets better which means that the model is able to learn better, while the validation accuracy improves in a rapid way, which means that the model gets more generalized and does not overfit anymore.

VI. PERFORMANCE EVALUATION OF THE PROPOSED SYSTEMS USING EXPERIMENTAL DATA

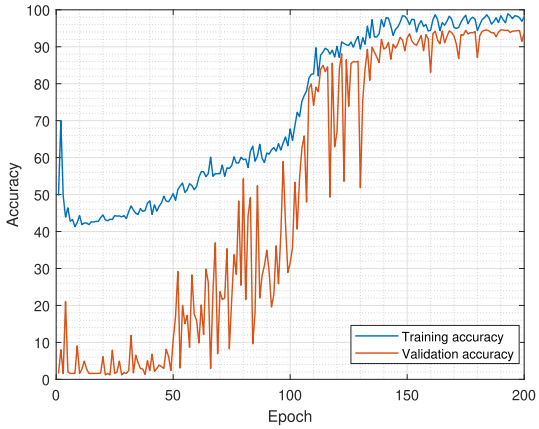
In order to support the simulation results, we evaluate the performance of the proposed systems based on experimental data from the UJIIndoorLoc database corresponding to Building1-Floor2 collected through measurements. We have 1395 training fingerprints taken at 80 training positions and 40 validation positions, collected four months later than training ones, taken at different validation positions received from 520 deployed APs.

A. UJIIndoorLoc DATABASE DESCRIPTION

UJIIndoorLoc is a publicly available WiFi fingerprinting database. It was created at the University of Jaume I, Spain in 2013. It contains three multifloor buildings (4/5 floors per building) covering 110 000 m2, and is composed of 19937 training fingerprints and 1111 validation fingerprints.



(a) Localization model accuracy for Supervised(1000,100) based on real data samples.



(b) Localization model accuracy for W-GAN(1000,100,2000) based on real and weighted generated data samples.

FIGURE 7. Localization performance during training phase and validation phase of the DNN model.

Each fingerprint contains 520 RSSI values corresponding to each AP received at a reference position given by its longitude and latitude. In this paper, we ignore the floor/building related information due to the fact that we only estimate the location (latitude and longitude), regardless the building and floor.

B. OBTAINED LOCALIZATION PERFORMANCE

Limited by the number of collected measurements corresponding to Building1-Floor2, we consider $T_r = 500$ labeled data samples, $T_p = 500$ unlabeled data samples and $T_t = 435$ data samples for the test. We randomly select labeled, unlabeled and test samples from the whole database. The presented results correspond to the average of several random draws, where the number of labeled positions and the number of test positions change. Consequently, in this part, we did not mention the number of labeled positions M_r and unlabeled positions M_p. We note that the number of fingerprints is not the same as the number of positions, since at one position, users take one or more fingerprint measurements. At first, we only keep APs detected at least once, which is equal to 190 from 520 deployed to construct the database. Then, after data pre-processing, we are able to apply our algorithms.

- **Weighted semi-supervised localization system:** In Table 7, we mention the parameters for each used DNN when applying the weighted semi-supervised localization system. The DNNs used in this part are trained during 200/250 epochs with 50/100 as mini-batch size and 0.01 as learning rate. $\mathrm{N}_i(.)$ refers to the number of neurons in the i^{th} hidden layer. When considering the semi-supervised algorithm based on a fixed weight, we use $w_r = 1$ and $w_p = 0.3$. When using a dynamic weight, we have $w = 3$ and $E_2 = 570$.
- **Weighted-GAN localization algorithm:** For data generation, we use GANs based on one hidden layer DNN generator with 200 neurons and one hidden layer DNN discriminator with 200 neurons. This generation is trained during 200 epochs using 0.01 as learning rate. The DNNs used for localization are composed of three hidden layers with 200 neurons, 150 neurons and 100 neurons, respectively. For weighting process, we consider $\omega = 3$ and $E_2 = 500$.

TABLE 7. DNN architectures when using 500 labeled data and 500 unlabeled data corresponding to building1-floor2 from the UJIIndoorLoc database.

Localization algorithm	DNN used for pseudo-labeling	DNN used for localization
Supervised	–	$\mathrm{N}_1(200)$, $\mathrm{N}_2(100)$ and $\mathrm{N}_3(30)$
Classical semi-supervised	$\mathrm{N}_1(200)$ and $\mathrm{N}_2(100)$	$\mathrm{N}_1(200)$ and $\mathrm{N}_2(100)$
Fixed-weight semi-supervised	$\mathrm{N}_1(200)$ and $\mathrm{N}_2(100)$	$\mathrm{N}_1(200)$, $\mathrm{N}_2(100)$ and $\mathrm{N}_3(50)$
Variable-weight semi-supervised	$\mathrm{N}_1(200)$, $\mathrm{N}_2(120)$ and $\mathrm{N}_3(50)$	$\mathrm{N}_1(200)$ and $\mathrm{N}_2(100)$

DNN architectures and different parameters are chosen based on an exhaustive empirical process to determine the best achievable ones.

We compare the localization accuracy of a supervised algorithm and three variants of semi-supervised algorithms (without weight, with a fixed weight and with a variable weight) using 500 labeled data samples and 500 unlabeled data samples and GAN-weighted localization systems using 500 labeled data and generating $T_f = 2000$. We note that we generated different T_f based on 500 labeled data from 250 to 4000 and the best obtained parameters correspond to $T_f = 2000$. Table 8 gives the mean localization error and the localization improvement compared with the supervised using the same set of data samples. We notice that the obtained results confirm the simulation results. Thus, the worst localization accuracy is obtained by the supervised algorithm and the use of weight enhances the performance

TABLE 8. Obtained localization performance considering 500 labeled data samples and 500 unlabeled data from the UJIIndoorLoc database building1-floor2.

	Localization error (m)	Accuracy increase vs supervised
Supervised	4.45	–
Classical semi-supervised	4.35	10 cm | 2.24%
Fixed-weight semi-supervised	4.2	25 cm | 5.61%
Variable-weight semi-supervised	4	45 cm | 10.11%
W-GAN(500,2000)	4.07	38 cm | 8.53%

TABLE 9. Localization performance using 1000 real labeled data from Floor2-Building1 of the UJIIndoorLoc database.

Algorithm	Mean localization error	Mean localization error decrease vs supervised (1000)
Supervised(1000)	4.1 m	–
W-GAN(1000,500)	3.55 m	13.41 %
W-GAN(1000,700)	3.52 m	14.14 %
W-GAN(1000,1000)	**3.39 m**	**17.31 %**
W-GAN(1000,1500)	3.58 m	12.68 %
W-GAN(1000,2000)	3.59 m	12.43 %
Selective SS-GAN (1000,1000) [36]	3.47 m	15.36 %

of the classical semi-supervised algorithm by 3.44% and 8.04% using a fixed and a variable weight, respectively, which justifies the combination of labeled data samples and weighted pseudo-labeled data samples. If we consider that having only labeled data, we note that the localization accuracy improvement is not promising compared to variable-weight semi-supervised since generated data are not good enough due to the small amount of labeled data.

To generate more realistic data samples, we choose randomly 1000 fingerprints from the training set. We test the localization accuracy combining real and generated data i.e. W-GAN(T_r,T_f). Table 9 presents the localization error, which shows that the error decreases using W-GAN compared to the conventional supervised system using only labeled data i.e. Supervised($T_r = 1000$). We generate different supplement fake positions [500 − 2000], where we notice that the best localization accuracy is obtained with 1000 generated positions achieving 17.31% localization improvement vs. the conventional supervised algorithm. Additionally, the performance is saturated, and thus, we cannot provide better accuracy from 1500 generated positions. As a result, even when working in a realistic environment with high dynamic and heterogeneous devices, our proposed system achieves good localization accuracy. We mention in Table 9 the performance of our previous work [36] corresponding to the algorithm Selective SS-GAN(1000,1000) using 1000 collected real data and 1000 selected-generated data. Such algorithm is based on (i) fake data generation, (ii) fake data pseudo-labeling and (iii) fake data selection. We can notice that we succeed to minimize the computational complexity without sacrificing the localization accuracy. We are even able to achieve almost 2% of localization accuracy improvement due to overcoming the pseudo-labels estimation error.

VII. SYSTEM PERFORMANCE DISCUSSION

As proved above, based on simulations and real data, the proposed methods both succeed in improving the localization accuracy compared to the supervised and classical-semi supervised learning. It is true that integrating such a data augmentation process increases the calculation complexity, but, this concerns only the offline phase when training and preparing our localization system. Consequently, once the localization model is trained, it is applied directly without any increase in online complexity.

Fixed-weight and variable-weight semi-supervised system: such system guarantees a localization accuracy improvement compared to the conventional semi-supervised and supervised schemes, especially when using a variable weight introducing gradually the pseudo-labels in the localization system training which allows to not disturb the training process and to ensure an optimized model fitting. The limitation of such method is that unlabeled data is not always available. Consequently, its application depends on whether we have access to unlabeled data or not.

W-GAN: This scheme can be applied once a small set of labeled data is available without the need to have extra unlabeled data. We notice that the number of real labeled data for data generation directly influences the obtained localization performances. As we can see in Section VI, when using the W-GAN proposed system based on 500 labeled data, we improve the localization accuracy by 8.53% compared to the supervised scheme. However, when we rely on 1000 labeled data and generate the same number of fake generated data 2000, the localization improvement is almost 17.31%. Thus, even if the data generation is a promising method for localization performance improvement, it is essential to have sufficient data, which can enable effective data generation. As mentioned in Section VI, compared to a previous work which generates RSSI vectors to be pseudo-labeled, we attain the same localization accuracy with lower computational complexity. Even more, we achieve almost 2% of localization accuracy improvement due to overcoming the pseudo-labels estimation error.

However, even when generating extra fake data so as to improve the training process of the localization model, such a model should be retrained periodically based on new collected data in order to cope with the indoor propagation conditions variations. Such operation needs additional human effort and heavy calculation resources. To deal with this problem, (1) federated learning is explored recently in order

to distribute the training process and maintain activity on some unities only and (2) transfer learning which transforms a model from a known environment to another variant environment. In future work, our research will be oriented towards deeper study of the aforementioned challenges.

VIII. CONCLUSION

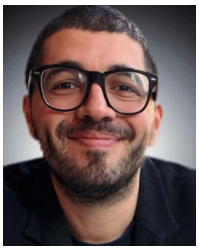
Machine learning-based indoor localization systems provide good localization accuracy with low online complexity. However, a proper training of a deep neural network (DNN) based localization model requires a large amount of collected labeled data which makes data collection an expensive task. To address this problem, in this paper, we propose two localization schemes. The first scheme, which is a semi-supervised system based on DNN for indoor localization, explores both real labeled data and pseudo labeled data in order to boost localization accuracy. This solution has been validated showing that the integration of a fixed or variable weight is beneficial in terms of localization performance compared to the supervised scheme and the classical semi-supervised scheme. When unlabeled data are not available and only a small set of real labeled data samples are collected, we propose a second localization scheme to deal with this scenario. We generate fake fingerprints using generative adversarial networks (GANs). RSSI samples and their labels are both directly provided by the GAN so that pseudo-labeling error is minimized. In order to enhance location prediction performance and avoid overfitting, a DNN model is trained on mixed dataset both comprising of real collected and fake generated data samples. During the training stage of the DNN-based localization model, a variable weighting coefficient is appropriately associated to the generated data samples to limit their reliance on fake data especially in the early training epochs. The proposed weighted data augmentation process leads to a localization improvement of 17.31% using the UJIIndoorLoc database. In future work, we will explore the transfer learning technique to overcome the challenge of collecting costly measurements [14]. Therefore, we will transfer a model obtained from a rich-data environment to a poor-data environment with limited measurements.

REFERENCES

[1] J. Schiller and A. Voisard, *Location-Based Services*. Amsterdam, The Netherlands: Elsevier, 2004.
[2] D. Dardari, P. Closas, and D. M. Djuric, "Indoor tracking: Theory, methods, and technologies," *IEEE Trans. Veh. Technol.*, vol. 64, no. 4, pp. 1263–1278, Apr. 2015.
[3] T. Pedersen and B. H. Fleury, "White paper on new localization methods for 5G wireless systems and the Internet-of-Things," Tech. Rep., 2018.
[4] A. Bourdoux, A. N. Barreto, B. van Liempd, C. de Lima, D. Dardari, D. Belot, E.-S. Lohan, G. Seco-Granados, H. Sarieddeen, H. Wymeersch, J. Suutala, J. Saloranta, M. Guillaud, M. Isomursu, M. Valkama, M. R. K. Aziz, R. Berkvens, T. Sanguanpuak, T. Svensson, and Y. Miao, "6G white paper on localization and sensing," 2020, *arXiv:2006.01779*.
[5] C. Chaccour, W. Saad, O. Semiari, M. Bennis, and P. Popovski, "Joint sensing and communication for situational awareness in wireless THz systems," 2021, *arXiv:2111.14044*.
[6] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2568–2599, 3rd Quart., 2017.
[7] W. Njima, I. Ahriz, R. Zayani, M. Terre, and R. Bouallegue, "Deep CNN for indoor localization in IoT-sensor systems," *Sensors*, vol. 19, no. 14, p. 3127, Jul. 2019.
[8] N. Singh, S. Choe, and R. Punmiya, "Machine learning based indoor localization using Wi-Fi RSSI fingerprints: An overview," *IEEE Access*, vol. 9, pp. 127150–127174, 2021.
[9] A. Poulose and D. S. Han, "Hybrid deep learning model based indoor positioning using Wi-Fi RSSI heat maps for autonomous applications," *Electronics*, vol. 10, no. 1, p. 2, Dec. 2020.
[10] M. Chafii, F. Bader, and J. Palicot, "Enhancing coverage in narrow band-IoT using machine learning," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2018, pp. 1–6.
[11] W. Njima, M. Chafii, A. Nimr, and G. Fettweis, "Convolutional neural networks based denoising for indoor localization," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Apr. 2021, pp. 1–6.
[12] I. B. F. de Almeida, M. Chafii, A. Nimr, and G. Fettweis, "Blind transmitter localization in wireless sensor networks: A deep learning approach," in *Proc. IEEE 32nd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2021, pp. 1241–1247.
[13] Y. Gu, Y. Chen, J. Liu, and X. Jiang, "Semi-supervised deep extreme learning machine for Wi-Fi based localization," *Neurocomputing*, vol. 166, pp. 282–293, Oct. 2015.
[14] M. I. AlHajri, R. M. Shubair, and M. Chafii, "Indoor localization under limited measurements: A cross-environment joint semi-supervised and transfer learning approach," in *Proc. IEEE 22nd Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2021, pp. 266–270.
[15] E. Kupershtein, M. Wax, and I. Cohen, "Single-site emitter localization via multipath fingerprinting," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 10–21, Jan. 2013.
[16] A. Jaffe and M. Wax, "Single-site localization via maximum discrimination multipath fingerprinting," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1718–1728, Apr. 2014.
[17] A. H. Salamah, M. Tamazin, M. A. Sharkas, and M. Khedr, "An enhanced WiFi indoor localization system based on machine learning," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Oct. 2016, pp. 1–8.
[18] J. Zhao and J. Wang, "WiFi indoor positioning algorithm based on machine learning," in *Proc. 7th IEEE Int. Conf. Electron. Inf. Emergency Commun. (ICEIEC)*, Jul. 2017, pp. 279–283.
[19] K. Sabanci, E. Yigit, D. Ustun, A. Toktas, and M. F. Aslan, "WiFi based indoor localization: Application and comparison of machine learning algorithms," in *Proc. 23rd Int. Seminar/Workshop Direct Inverse Problems Electromagn. Acoustic Wave Theory (DIPED)*, Sep. 2018, pp. 246–251.
[20] D. Milioris, M. Bradonjic, and P. Muhlethaler, "Building complete training maps for indoor location estimation," in *Proc. IEEE Conf. Comput. Commun. Workshops (INFOCOM WKSHPS)*, Apr. 2015, pp. 75–76.
[21] Y. Xia, L. Ma, Z. Zhang, and Y. Wang, "Semi-supervised positioning algorithm in indoor WLAN environment," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, May 2015, pp. 1–5.
[22] M. Zhou, Y. Tang, W. Nie, L. Xie, and X. Yang, "GrassMA: Graph-based semi-supervised manifold alignment for indoor WLAN localization," *IEEE Sensors J.*, vol. 17, no. 21, pp. 7086–7095, Nov. 2017.
[23] M. Zhou, Y. Tang, Z. Tian, L. Xie, and W. Nie, "Robust neighborhood graphing for semi-supervised indoor localization with light-loaded location fingerprinting," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3378–3387, Oct. 2017.
[24] J. Yoo and J. Park, "Indoor localization based on Wi-Fi received signal strength indicators: Feature extraction, mobile fingerprinting, and trajectory learning," *Appl. Sci.*, vol. 9, no. 18, p. 3930, Sep. 2019.
[25] J. J. Pan, S. J. Pan, J. Yin, L. M. Ni, and Q. Yang, "Tracking mobile users in wireless networks via semi-supervised colocalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 587–600, Mar. 2012.
[26] L. Chen, I. W. Tsang, and D. Xu, "Laplacian embedded regression for scalable manifold regularization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 6, pp. 902–915, Jun. 2012.
[27] J. Yoo, "Time-series Laplacian semi-supervised learning for indoor localization," *Sensors*, vol. 19, no. 18, p. 3867, Sep. 2019.
[28] W. Li, C. Zhang, and Y. Tanaka, "Pseudo label-driven federated learning-based decentralized indoor localization via mobile crowdsourcing," *IEEE Sensors J.*, vol. 20, no. 19, pp. 11556–11565, Oct. 2020.
[29] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Workshop Challenges Represent. Learn. ICML*, 2013, vol. 3, no. 2, p. 896.

[30] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014, *arXiv:1406.2661*.
[31] K. M. Chen and R. Y. Chang, "Semi-supervised learning with GANs for device-free fingerprinting indoor localization," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.
[32] Q. Li, H. Qu, Z. Liu, N. Zhou, W. Sun, S. Sigg, and J. Li, "AF-DCGAN: Amplitude feature deep convolutional GAN for fingerprint construction in indoor localization systems," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 5, no. 3, pp. 468–480, Jun. 2019.
[33] Y. Lei, D. Li, H. Zhang, and X. Li, "Wavelet feature outdoor fingerprint localization based on ResNet and deep convolution GAN," *Symmetry*, vol. 12, no. 9, p. 1565, Sep. 2020.
[34] H. Zou, C.-L. Chen, M. Li, J. Yang, Y. Zhou, L. Xie, and C. J. Spanos, "Adversarial learning-enabled automatic WiFi indoor radio map construction and adaptation with mobile robot," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 6946–6954, Mar. 2020.
[35] W. Njima, M. Chafii, and R. M. Shubair, "GAN based data augmentation for indoor localization using labeled and unlabeled data," in *Proc. Int. Balkan Conf. Commun. Netw. (BalkanCom)*, Sep. 2021, pp. 36–39.
[36] W. Njima, M. Chafii, A. Chorti, R. M. Shubair, and H. V. Poor, "Indoor localization using data augmentation via selective generative adversarial networks," *IEEE Access*, vol. 9, pp. 98337–98347, 2021.
[37] J. Torres-Sospedra, R. Montoliu, A. Martinez-Uso, J. P. Avariento, T. J. Arnau, M. Benedito-Bordonau, and J. Huerta, "UJIIndoorLoc: A new multi-building and multi-floor database for WLAN fingerprint-based indoor localization problems," in *Proc. Int. Conf. Indoor Positioning Indoor Navigat. (IPIN)*, Oct. 2014, pp. 261–270.
[38] W. Njima, M. Chafii, A. Nimr, and G. Fettweis, "Deep learning based data recovery for localization," *IEEE Access*, vol. 8, pp. 175741–175752, 2020.
[39] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.
[40] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*.
[41] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Process.*, vol. 35, no. 1, pp. 53–65, Jan. 2017.
[42] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: Introduction and outlook," *IEEE/CAA J. Autom. Sinica*, vol. 4, no. 4, pp. 588–598, Sep. 2017.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[44] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.

WAFA NJIMA (Member, IEEE) received the Engineering degree from the Institut National des Sciences Appliquées et de Technologies de Tunis, in 2015, and the Ph.D. degree in the field of radio-communications from the Conservatoire National des Arts et Métiers in Paris in collaboration with the Ecole Supérieure des Communications de Tunis, in 2019. She joined the ETIS Laboratory at ENSEA Cergy as a Temporary Assistant and a Researcher also as a Postdoctoral Researcher. She is currently an Associate Professor at ISEP, Paris. Her publications span several research areas, and her research interests include related to several topics, including signal processing, wireless communications, sparse data and data augmentation, indoor localization, the IoT, and machine learning for communications. She served as a TPC member and a reviewer for many leading international conferences and journals.



AHMAD BAZZI was born in Abu Dhabi, United Arab Emirates. He received the M.Sc. degree *(summa cum laude)* in wireless communication systems (SAR) from the Centrale Supélec, in 2014, and the Ph.D. degree in electrical engineering from EURECOM, France, in 2017. He is currently a Research Associate with NYU Abu Dhabi working on 6G and joint sensing and communications and prior to that, he was the Algorithm and Signal Processing Team Leader at CEVA-DSP, Sophia Antipolis, leading the work on Wi-Fi (802.11ax) and Bluetooth (5.x BR/BLE/BTDM/LR) high performant (HP) PHY modems, OFDMA MAC schedulers, and RF-related issues. Since 2018, he has been devoting to publishing YouTube lectures, where his channel contains mathematical, algorithmic, and programming topics, with more than 120K subscribers and more than 10M views, as of April 2022. He was awarded a CIFRE Scholarship from ANRT France, in 2014. He was nominated for Best Student Paper Award at IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), in 2016. He received a Silver Plate Creator Award from YouTube, in 2022. He is a co-inventor in several patents involving intellectual property of Wi-Fi and Bluetooth products, all of which have been implemented and sold to key clients. He has served as a TPC member for some IEEE conferences and a reviewer at several top IEEE conferences and well-known IEEE journals. His research interests include signal processing, wireless communications, statistics, and optimization.

MARWA CHAFII (Member, IEEE) received the master's degree in the field of advanced wireless communication systems (SAR) and the Ph.D. degree in electrical engineering from the Centrale Supélec, France, in 2013 and 2016, respectively. From 2014 to 2016, she was a Visiting Researcher with the Poznan University of Technology, Poland, University of York, U.K., Yokohama National University, Japan, and the University of Oxford, U.K. She joined the Technical University of Dresden, Germany, in 2018, as a Research Group Leader, and ENSEA, France, in 2019, as an Associate Professor, where she was the Chair of Excellence on Artificial Intelligence from CY Initiative. Since September 2021, she has been an Associate Professor with New York University (NYU) Abu Dhabi, and NYU WIRELESS, NYU Tandon School of Engineering. Her research interests include advanced waveform design, machine learning for wireless communications, and indoor localization. She received the prize of the Best Ph.D. in France in the fields of signal, image & vision, and she has been nominated in the top ten Rising Stars in Computer Networking and Communications by N2Women, in 2020. She served as the Associate Editor for IEEE Communications Letters, from 2019 to 2021, where she received the Best Editor Award, in 2020. From 2018 to 2021, she was a Research Lead at the Women in AI Organization. She is currently an Associate Editor of IEEE Transactions on Communications. She serves as the Vice-Chair of the IEEE ComSoc ETI on Machine Learning for Communications and leading the Education working group of the ETI on Integrated Sensing and Communications.