GAN Based Data Augmentation for Indoor Localization Using Labeled and Unlabeled Data

Wafa Njima, Marwa Chafii, Raed M Shubair

GAN Based Data Augmentation for Indoor Localization Using Labeled and Unlabeled Data

Wafa Njima*, Marwa Chafii†, Raed M. Shubair†
*ETIS UMR 8051, CY Paris Université, ENSEA, CNRS, France
†New York University (NYU), Abu Dhabi 129188, UAE
Email: wafa.njima@ensea.fr, {marwa.chafii, raed.shubair}@nyu.edu

***Abstract*—Machine learning techniques allow accurate indoor localization with low online complexity. However, a large amount of collected data samples is needed to properly train a deep neural network (DNN) model used for localization. In this paper, we propose to generate fake fingerprints using generative adversarial networks (GANs) based on a small amount of collected data samples. We consider an indoor scenario where collected labeled data samples are rare and insufficient to generate fake samples of a good multitude and diversity in order to provide a good localization accuracy. Thus, both labeled and unlabeled fingerprints are provided to the GAN so that more realistic fake data samples are generated. Then, a DNN model is trained on mixed dataset comprising real collected labeled and pseudo-labeled fingerprints as well as fake generated pseudo-labeled fingerprints. The data augmentation based on real measurements leads to a mean localization accuracy improvement of 9.66% in comparison to the conventional semi-supervised localization algorithm.**

***Index Terms*—Indoor localization, deep neural network (DNN), generative adversarial network (GAN), received signal strength indicator (RSSI), semi-supervised learning.**

I. Introduction

Future wireless systems will support a wide range of innovative applications such as healthcare monitoring, autonomous vehicles and personal navigation [1] [2]. To this end, these systems have to provide accurate and efficient location-based services, resulting in an increasing demand for accurate location information. Classical localization techniques, comprising geometric and fingerprints-based methods use wireless signal parameters including angle of arrival (AoA), time of arrival (ToA), channel state information (CSI) and received signal strength indicator (RSSI) [3]. These techniques are severely impacted by multi-path effects and hinders real-time implementation and operation, especially in extended networks. To deal with this issue, machine learning (ML) tools, in particular deep learning (DL) techniques have been widely used in which an online localization model is employed, whereby training and optimization has been performed offline [4]–[6]. To optimally train a DL model, a large amount of data is needed. However, data collection is a highly consuming task in terms of time, hardware resources, and human effort.

To overcome these challenges, various techniques have been used to generate fake data that would compliment the real collected data for localization accuracy improvement. Recently, generative models have been widely adopted for data augmentation, in particular, generative adversarial networks (GANs) [7] [8]. Such networks are used to (i) increase the diversity of samples, by generating fake fingerprints at training positions already used during data collection and (ii) increase the size of the database by generating fake fingerprints at new positions. Unlike our previous work [9] where we assume that only labeled collected data are available for data generation, we consider in this work a scenario where a small amount of labeled collected data is insufficient to generate good and realistic fake data samples. Thus, we leverage both labeled and unlabeled collected data samples for data augmentation and then the whole data (collected and generated) are combined and used for localization. To the best of the authors knowledge, it is the first time that labeled and unlabeled data are explicitly used for data augmentation based on GANs in the localization context. Such combination was implicitly used in [10] to benefit from unlabeled data only in the step of optimizing the GAN model weights while in our work we also use it to build the localization model.

In this paper, collected RSSI fingerprints (labeled and unlabeled) are used for new fake fingerprints generation. Once fake fingerprints are generated, a pseudo-labeling process [11] is performed to predict pseudo-labels for both unlabeled and fake generated fingerprints. Then, the whole data are combined in order to build a deep neural network (DNN) model for localization. The remainder of this paper is organized as follows: In Section II, we describe and formulate the problem. Section III presents the proposed semi-supervised GAN for location information augmentation. The performance of the proposed approach in comparison to the conventional method is presented in Section IV for a real environment. Finally, we conclude our work in Section V.

II. Problem Description

Fingerprints-based localization methods have been widely studied and adopted for its localization accuracy and simplicity. Such technique consists of an offline phase and an online phase. During the offline phase, a site survey is conducted collecting RSSI measurements at each training position received from different access points (APs) when considering WiFi signals. Collected RSSI associated to the corresponding location coordinates construct training fingerprints contained in the training database to be transmitted and stored in a central unit (CU). For online localization, RSSI measurements are compared with the training fingerprints for location estimation. A prediction based on a DNN model is recommended since the online complexity is shifted to the offline phase. A good localization performance is achieved when the DNN model is trained and optimized based on a large set of collected data which makes the fingerprints collection time and cost consuming. Thus, data augmentation based on GANs for fake localization data generation is widely studied. However, RSSI vectors labeled by the (x, y) coordinates, collected during the offline phase, can be insufficient for accurate fake data generation. To solve this issue, we propose a data augmentation module using GANs based on collected labeled and unlabeled data. Then, the location prediction is conducted combining collected data (labeled and unlabeled) and fake generated data. This system reduces the reliance on expensive labeled collected data, mixing both labeled and unlabeled data with fake generated data.

III. Proposed Semi-Supervised GAN for Location Data Augmentation

We consider an indoor environment covering $(L \times W)$ m^2, where M APs are already deployed. Mobile sensor nodes, used during the offline phase, collect RSSI measurements at known training positions 'labeled data' and at unknown positions 'unlabeled data' for training database construction. Other mobile sensor nodes are requiring localization online periodically or when needed. The localization process and the data pre-processing are performed by the CU. Different steps of the system, depicted in Fig. 1, are explained below.

A. Using GANs for data generation

To increase the training dataset size and diversity, we use a special class of generative models: GANs. A GAN is composed by a generator model G which learns how to produce a representation similar to real data and a discriminator model D which learns how to distinguish between real and fake data. G and D, based on DNN, are trained together until fake generated data samples start looking as real samples. The goal is to generate F_g RSSI vectors composed of M RSSIs received from used APs starting from an input noise.

B. Training a supervised DNN model for pseudo-labeling

To be used for localization, collected unlabeled data and fake generated data need to be identified by predicted coordinates called 'pseudo-labels' using a DNN model. Based on labeled collected RSSI vectors, a DNN model can be built taking as input an RSSI vector and its corresponding coordinates as output. Once trained, this DNN model is used for pseudo-labeling. Thus, it takes each real unlabeled or fake generated RSSI vector to predict its pseudo-labels.

C. Using a mixed DNN model for localization

Localization is performed applying a mixed DNN model trained offline while combining labeled and pseudo-labeled data. To localize a mobile sensor node online, such model is applied to estimate its coordinates based on the corresponding collected RSSI vector.

IV. Performance evaluation

We test the proposed system on real measurements from the UJIndoorLoc database [12]. Since we work on one floor environment, we consider only collected data corresponding to Building1-Floor2. We consider $F_l = 500$ labeled fingerprints and $F_u = 500$ unlabeled fingerprints during training. For test, we use the rest of fingerprints which is equal to $F_t = 435$. The data collected is first pre-processed to eliminate redundant and useless data since at each position only 18 AP are detected from 520 existing. Thus, we keep only APs detected at least once which is equal to 190. After data pre-processing, the fake data samples are generated. This step is performed with a GAN based on one-hidden DNN layer as a discriminator with 200 neurons and a one-hidden DNN layer generator with 200 neurons. The GAN is trained during 200 epochs using 0.01 learning rate. For data pseudo-labeling, the considered DNN is trained during 200 epochs with 50 as mini-batch size and 0.01 as learning rate. For localization, 250 epochs are considered with a mini-batch size equal to 100 and a learning rate equal to 0.01. The DNN architectures used for pseudo-labeling and localization are mentioned in Table I where $L_i(.)$ refers to the number of neurons in the i^{th} hidden layer.

We compare the localization accuracy of different localization methods combining different types of data.

- **Supervised (F_l)**: when using a supervised method based on F_l labeled fingerprints.
- **SS (F_l, F_u)** is the localization method when using classical semi supervised learning. Where F_l indicates the number of labeled data samples and F_u indicates the number of unlabeled data samples.
- **SS-GAN (F_l, F_u, F_g)** is the localization method when we combine F_l labeled, F_u unlabeled and F_g fake generated data samples for localization.

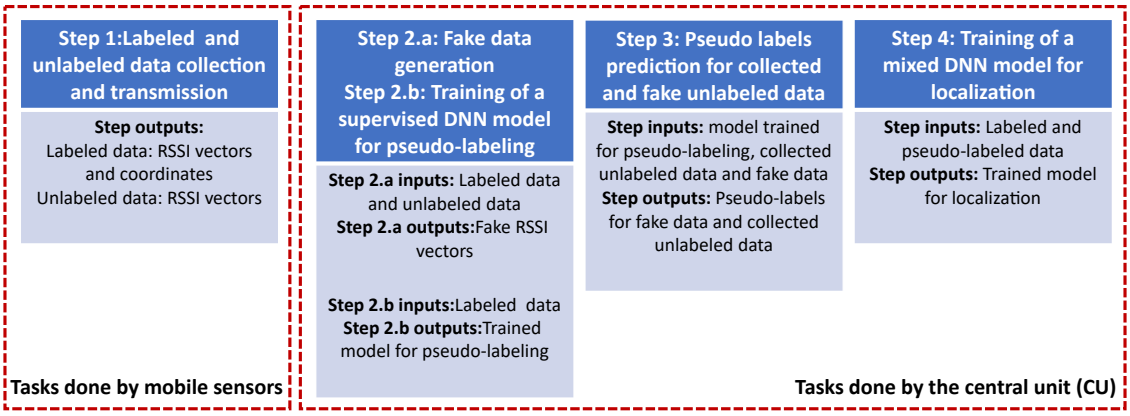

Fig. 1: Overview of the system model and the different training steps.

TABLE I: Used DNN architectures.

Localization method	DNN used for pseudo-labeling	DNN used for localization
SS (500, 500)	$L_1(200)$ and $L_2(100)$	$L_1(200)$ and $L_2(100)$
SS-GAN (500,500,100)	$L_1(200)$ and $L_2(100)$	$L_1(200)$, $L_2(100)$ and $L_3(50)$
SS-GAN (500,500,250)	$L_1(200)$, $L_2(100)$ and $L_3(100)$	$L_1(200)$, $L_2(100)$ and $L_3(50)$
SS-GAN (500,500,500)	$L_1(200)$, $L_2(100)$ and $L_3(100)$	$L_1(200)$, $L_2(100)$ and $L_3(50)$
SS-GAN (500,500,1000)	$L_1(200)$, $L_2(100)$ and $L_3(100)$	$L_1(200)$, $L_2(100)$ and $L_3(50)$
SS-GAN (500,500,1500)	$L_1(200)$, $L_2(100)$ and $L_3(100)$	$L_1(100)$, $L_2(100)$ and $L_3(50)$

Table II gives the localization errors (e.g., mean localization error, min localization error and max localization error) and the localization improvement compared with SS $(500, 500)$ i.e. using 500 labeled with 500 unlabeled data samples. We can first notice that the conventional semi-supervised localization algorithm improves the mean localization accuracy by 2.24% compared with the conventional supervised scheme. This shows that pseudo-labeled data can indeed improve the localization accuracy. For fake RSSI samples generations, we use both available labeled and unlabeled data samples in order to improve the quality of the fake generated samples, since the GAN will be trained over a larger dataset. $[100 - 1500]$ RSSI vectors are then generated. These RSSIs correspond to new fake positions covering new regions of the studied indoor environment, which gives more diversity and coverage to the dataset. We notice that for all augmented datasets, the localization accuracy is improved compared to the initial training dataset limited only to real collected data. The best localization accuracy is obtained when generating 1000 fake positions in terms of mean localization error 3.93 m and optimal localization error 3.91 m, improving the conventional semi-supervised system by 9.66% and 8.85%, respectively. This improvement is explained by the fact that the DNN, used for localization, is trained over a larger dataset which contains new positions that are not included in the limited dataset based only on collected data and covering a large area in the considered indoor environment as shown in Fig. 2. In this figure, we show a distribution of 1000 generated fake positions based on 1000 real positions (500 labeled positions and 500 pseudo-labeled positions). In addition, using additional fake data samples, the DNN model is able to learn better which means that it gets more generalized and avoids overfitting. Thus, even when working in a real environment with high dynamic and heterogeneous devices, our proposed system achieves good localization accuracy and improves the performance obtained by the conventional semi-supervised framework. Starting from a certain number of generated fake data samples which is equal to 1000, the performance is saturated and no further improvements can be provided which can be explained by two reasons: (i) Based on 500 labeled vectors and 500 pseudo-labeled vectors, we cannot provide a higher diversity to the GAN, which leads to generating less realistic samples, (ii) The prediction error resulting from pseudo-labels prediction.

V. Conclusion

In this paper, GANs are leveraged for augmenting data used for localization. Thus, a combination of real collected labeled and unlabeled data along with fake generated data is used to improve the localization accuracy. The data augmentation process leads to a mean localization accuracy improvement compared to the conventional localization system based only on collected labeled and unlabeled data considering real measurements taken from the UJIndoorLoc database, without any online complexity increase. In future works, we intend to generate both RSSI vectors and their corresponding

TABLE II: Obtained localization performance considering 500 labeled data samples and 500 unlabeled data samples with real measurements from the UJIndoorLoc Database: Building1-Floor2.

Localization method	**Optimal localization error (m)**	**Mean localization error (m)**	**Min - Max localization error (m)**	**Optimal localization accuracy improvement vs SS** (500, 500)	**Mean localization accuracy improvement vs SS** (500, 500)
SS (500, 500)	4.29	4.35	4.29 - 4.41	–	–
SS-GAN (500,500,100)	4.19	4.25	4.19 - 4.28	10 cm | 2.33%	10 cm | 2.29%
SS-GAN (500,500,250)	3.98	4.02	3.98 - 4.07	31 cm | 7.22%	33 cm | 7.58%
SS-GAN (500,500,500)	3.91	3.96	3.91 - 4.00	38 cm | 8.85%	39 cm | 8.96%
SS-GAN (500,500,1000)	**3.91**	**3.93**	**3.91 - 3.97**	**38 cm | 8.85%**	**42 cm | 9.66%**
SS-GAN (500,500,1500)	3.93	3.95	3.93 - 3.97	36 cm | 8.39%	40 cm | 9.19%
Supervised (500)	4.37	4.45	4.37 - 4.5	-8 cm | -1.83 %	-10 cm | -2.24%

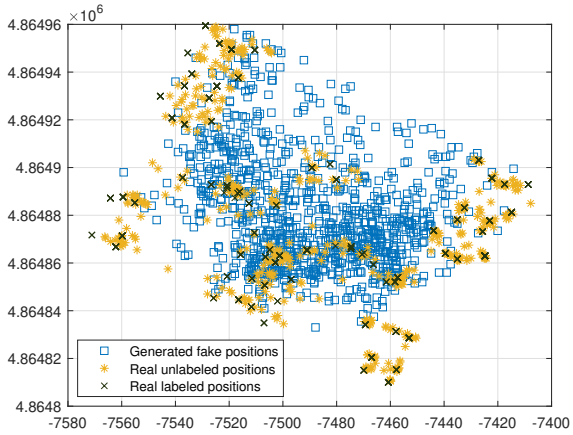

Fig. 2: An example of a distribution of 1000 fake positions generated based on 500 labeled positions and 500 pseudo-labeled positions.

coordinates to overcome the prediction error of pseudo-labels.

ACKNOWLEDGMENT

This work was supported in part by CY Initiative of Excellence (grant "Investissements d'Avenir" ANR-16-IDEX-0008) and by the project DELICATE funded by CNRS/INS2I.

REFERENCES

[1] Pedersen Troels and Fleury Bernard Henri. "White Paper on New Localization Methods for 5G Wireless Systems and the Internet-of-Things." 2018.

[2] Bourdoux Andre, Barreto Andre Noll, van Liempd Barend, de Lima Carlos, Dardari Davide, Belot Didier, Lohan Elana-Simona, Seco-Granados Gonzalo, Sarieddeen Hadi, Wymeersch Henk and others. "6G White Paper on Localization and Sensing." arXiv preprint arXiv:2006.01779 (2020).

[3] Zafari Faheem, Gkelias Athanasios and Leung Kin K. "A survey of indoor localization systems and technologies." IEEE Communications Surveys & Tutorials 21.3 (2019): 2568-2599.

[4] Njima Wafa, Ahriz Iness, Zayani Rafik, Terre Michel and Boullegue Ridha. "Deep CNN for Indoor Localization in IoT-Sensor Systems." Sensors 19.14 (2019).

[5] Njima Wafa, Chafii Marwa, Nimr Ahmad and Fettweis Gerhard. "Deep Learning Based Data Recovery for Localization." IEEE Access (2020): 175741-175752.

[6] Bizon Ivo, Chafii Marwa, Nimr Ahmad and Fettweis Gerhard. "Blind Transmitter Localization in Wireless Sensor Networks: A Deep Learning Approach." IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2021.

[7] Belmonte-Hernández Alberto, Hernandez-Penaloza Gustavo, Gutiérrez David Martín and Alvarez Federico. "Recurrent Model for Wireless Indoor Tracking and Positioning Recovering Using Generative Networks." IEEE Sensors Journal 20.6 (2019): 3356-3365.

[8] Bowles Christopher, Chen Liang, Guerrero Ricardo, Bentley Paul, Gunn Roger, Hammers Alexander, Dickie David Alexander and Hernández Maria Valdés, Wardlaw Joanna and Rueckert Daniel. "GAN Augmentation: Augmenting Training Data Using Generative Adversarial Networks." arXiv preprint arXiv:1810.10863 (2018).

[9] Njima Wafa, Chafii Marwa, Chorti Arsenia, Shubair Raed M and Poor H Vincent. "Indoor Localization using Data Augmentation via Selective Generative Adversarial Networks." IEEE Access (2021): 98337-98347.

[10] Odena Augustus. "Semi-Supervised Learning with Generative Adversarial Networks." arXiv preprint arXiv:1606.01583 (2016).

[11] Lee Dong-Hyun and others. "Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks." Workshop on challenges in representation learning (ICML). 2013.

[12] Torres-Sospedra Joaquín, Montoliu Raúl, Martínez-Usó Adolfo, Avariento Joan P, Arnau Tomás J, Benedito-Bordonau Mauri and Huerta Joaquín. "UJIIndoorLoc: A New Multi-Building and Multi-Floor Database for WLAN Fingerprint-Based Indoor Localization Problems." 2014 IEEE International Conference on Indoor Positioning and Indoor Navigation (IPIN). IEEE, 2014.