Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

A survey on data augmentation for WiFi fingerprinting indoor positioning

XU FENG[1], KHUONG AN NGUYEN[1], and ZHIYUAN LUO[1]

[1] Computer Science Department, Royal Holloway University of London, Surrey, TW20 0EX, United Kingdom

Corresponding author: Xu Feng (email: Xu.Feng@rhul.ac.uk).

ABSTRACT WiFi fingerprinting has been a prominent solution for indoor positioning, yet its dependence on labour-intensive data collection and susceptibility to environmental dynamics are on-going major challenges. Thus, this paper presents a comprehensive survey and analysis of the data augmentation techniques designed to enhance WiFi fingerprinting datasets, focusing on the efficiency in data construction and the robustness in positioning accuracy. We reviewed over 70 studies, and proposed a novel taxonomy that categorises existing methods into 6 groups: traditional (e.g., interpolation, perturbation), propagation models, machine learning, deep learning, hybrid approaches, and other emerging techniques. Our quantitative analysis correlates key metrics, such as input data size, synthetic data volume, and augmentation ratios, with positioning performance. We found that traditional methods achieved notable performance enhancements with minimal computational overhead. Surprisingly, deep learning models became less efficient when generating more data, particularly when the synthetic data exceeded an threefold ratio over the input samples. Our findings provide actionable guidance for selecting data augmentation strategies and bridge the gap between theoretical advancements and practical deployment for WiFi fingerprinting dataset enhancement.

INDEX TERMS Indoor positioning, WiFi fingerprinting, data augmentation, generative models.

I. INTRODUCTION

WHILE Global Positioning System (GPS) performs reliably in outdoor environments, its accuracy significantly degrades indoors due to signal attenuation and multipath effects. This limitation has motivated the exploration of alternative indoor positioning techniques. Amongst these, WiFi fingerprinting has emerged as one of the most popular techniques for indoor positioning systems leveraging existing infrastructure and enabling cost-effective deployment compared to dedicated hardware solutions like Ultrawide band (UWB), and Bluetooth Low Energy (BLE) [1]–[4].

WiFi fingerprinting involves constructing a dataset through systematic site surveys, during which the WiFi signal measurements such as Received Signal Strength (RSS), Channel State Information (CSI), or Round-Trip Time (RTT) are recorded at known locations [1]–[3], [5]. This pre-constructed fingerprinting dataset will subsequently be used to train a Machine Learning model to estimate the user's location based on newly observed WiFi signal measurements collected at an unknown position. Thus, the accuracy of WiFi fingerprinting-based systems rely heavily on the quality and comprehensiveness of this fingerprinting dataset.

Although dense and spatially uniform WiFi fingerprint data collection and construction are theoretically optimal, its practical implementation faces significant challenges, including intensive labour demands and high maintenance cost [5], [6]. To ensure the accuracy of the fingerprinting dataset, precise ground truth location acquisition must be carried out prior to the actual collection of WiFi signal measurements. Moreover, to capture the variability and stability of signals, each reference point (RP) typically requires data recording for over one minute. For example, conducting a systematic site survey in the open spaces of a university building floor measuring 92 × 15 metres at 1-metre intervals demands more than 40 hours of manual effort by a human tester [7]–[9]. Additionally, to address the temporal changes in the indoor environment and fluctuations in WiFi signal characteristics, ongoing maintenance and periodic recalibration are necessary [1], [10]–[13].

To tackle these challenges, data augmentation techniques have been widely adopted in the literature. Data augmentation was proposed as a part of a broad set of techniques designed to enhance the performance of machine learning and deep learning models by introducing additional information [14]–[17]. In WiFi fingerprinting data construction,

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

data augmentation is used either as a method to generate synthetic WiFi data samples in uncovered areas to reduce human labours, or to enrich the WiFi fingerprinting dataset by changing some characteristics of the training dataset itself for positioning performance enhancement.

However, **despite the growing number of data augmentation methods proposed in the literature, there remains a lack of systematic analysis of their strengths, limitations, and comparative performance**. To address this need, our paper presents a comprehensive review and analysis of over 70 recent research works on WiFi fingerprinting data augmentation. Specifically, we examine and compare key aspects such as testbed types and sizes, the total number of collected data, the total number of real-world data used as input to augmentation models, the total number of synthetic data generated, and the positioning performance before and after augmentation. Through this analysis, the paper aims to provide valuable insights into the most widely adopted augmentation techniques, the most detailed performance comparisons to date, the most effective data augmentation methods and most efficient ratios of synthetic to original input data for positioning performance improvement. Additionally, we explore current trends and challenges in WiFi fingerprinting data augmentation and highlight promising directions for future research on more efficient and robust data construction methods.

In summary, our paper makes the following contributions:

- We conducted an in-depth and extensive analysis of over 70 WiFi fingerprinting data augmentation research papers, from physics-driven propagation model to the latest generative models, providing a comprehensive overview of the current state-of-the-arts.
- We present a novel taxonomy of WiFi fingerprinting data augmentation approaches to represent the current landscape of research and practice. Specifically, we categorise all existing methods into 6 groups: traditional methods, propagation model, machine learning methods, deep learning methods, hybrid methods, and others.
- We systematically identify and analyse the key characteristics of the existing WiFi fingerprinting data augmentation methods, including those not explicitly stated in the original research paper. Specifically, it examines: (1) the total volume of data collected, (2) the quantity of real-world data employed as input for augmentation models, (3) the total amount of synthetic data generated, and (4) positioning performance metrics both before and after data augmentation for the included methods.
- We identify the most effective data augmentation methods and most efficient ratios of synthetic to original input data that enhance WiFi-based indoor positioning accuracy.

The remainder of this paper is structured as follows. Section II introduces the review focus and research methodology adopted by this paper. Section III provides detailed and comprehensive descriptions of all trending data augmentation methods in WiFi fingerprinting. Section IV reveals the positioning performance with data augmentation in relationship to numerous factors. Finally, Section V concludes the paper.

II. REVIEW SCOPE AND METHODOLOGY

To ensure the reproducibility of the literature selection process, this review adopts a systematic methodology inspired by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [18]. The PRISMA framework follows a structured four-phase process: (1) Identification (searching databases with predefined keywords), (2) Screening (filtering records based on titles/abstracts/contents), (3) Eligibility (assessing full-text articles against inclusion/exclusion criteria), and (4) Inclusion (finalizing selected studies). This section outlines the research scope in detail, provides the database, keywords and inclusion/exclusion criteria utilised, and describes the systematic approach used to search for, select, screen, include, exclude, and analyse existing WiFi fingerprinting data augmentation methods, thereby ensuring a rigorous and comprehensive examination of the topic.

A. REVIEW FOCUS

This paper is dedicated to providing a comprehensive review and in-depth analysis of current data augmentation methods currently used in WiFi fingerprinting, with the goal of offering valuable and unique insights into the trends and challenges in this field. Our primary focus is on research papers that explore **synthetic data generation** and **dataset enrichment** for WiFi fingerprinting, specifically in the dataset construction stage of the indoor positioning systems. Therefore, **studies that primarily focused on transfer learning, rapid adaptation to new indoor environments, and WiFi fingerprinting dataset update towards environment changes were excluded from this review.** Studies on indoor positioning systems using other wireless signals, such as BLE, UWB and 5G, are excluded from this review. To ensure a comprehensive review, this paper also includes research that, while not specifically focused on proposing novel WiFi fingerprinting data augmentation methods, employs widely-used models or approaches, or generates synthetic data related to enrich the fingerprinting dataset.

B. PAPERS SELECTION CRITERIA

To perform an extensive search and determine the highly relevant WiFi fingerprinting data augmentation research, the following keywords were used: *"WiFi," "WLAN," "indoor," "localization,", "localisation," "indoor positioning," "navigation," and "fingerprinting"*. To explore general data augmentation methods in the dataset construction stage of WiFi fingerprinting, we employed keywords including *"data", "radio map", "construction", "collection", "syn-*

thetic", "synthesis", "augmentation", "generation", and "generative" in our search. These keywords were searched in the title, keywords, and main body of the research papers on well-known websites and research platforms such as Google Scholar, Web of Science, IEEE Xplore, ACM Digital Library, ScienceDirect and SpringerLink. Additionally, research publications that include a comparison table covering a selected number of WiFi fingerprinting data augmentation methods were also incorporated into the scope of the literature search [19]–[22].

The inclusion criteria for studies in this review were as follows:

- Peer-reviewed articles published in English;
- Studies employing quantitative designs that specifically analyse positioning performance, particularly in the context of augmentation;
- Studies intended to either reduce the human effort in WiFi fingerprinting data collection and construction, or improve the performance of the indoor positioning models on existing testbeds;
- Studies that either generated synthetic WiFi fingerprinting data samples or enriched existing training datasets.

Exclusion criteria encompassed grey literature and non-empirical studies. Furthermore, in accordance with the criteria and our research scope, studies focusing on transfer learning methods, data augmentation techniques aimed at rapid deployment in new indoor environments, and WiFi fingerprinting dataset update methods towards environment changes were excluded from consideration. For research papers employing multiple distinct data augmentation methods, the positioning performance corresponding to each method will be recorded separately. To ensure relevance to the current trends in WiFi fingerprinting data augmentation, only research articles published post-2015 were included. A comprehensive literature search was conducted across major research platforms using the previously specified keywords, initially identifying 243 papers. After removing duplicates and excluding outdated or less relevant studies, titles, abstracts, and keywords were evaluated to eliminate works unrelated to WiFi fingerprinting and data augmentation (e.g., studies focusing on fingerprint data augmentation for BLE). Subsequently, through meticulous manual review and critical analysis of each paper's technical content, methodological approach, and experimental validation, the selection was refined to a final set of over 70 publications.

III. WIFI FINGERPRINTING DATA AUGMENTATION

Data augmentation was proposed as a data-space solution to addressing the challenges of limited data. The goal of data augmentation is to enhance the quantity and quality of the training data samples so that better machine learning and deep learning models can be built using them [14]–[17].

As WiFi fingerprinting heavily relies on the positioning models to match user-reported WiFi signal measurements to a pre-constructed fingerprint database, it is highly data-intensive. To reduce the human effort in the collection and construction of WiFi fingerprinting datasets or to enhance the performance of the positioning estimation model trained on the datasets, numerous data augmentation methods were adopted in the literature to enrich and expand the pre-collected WiFi fingerprinting datasets (see Figure 1).

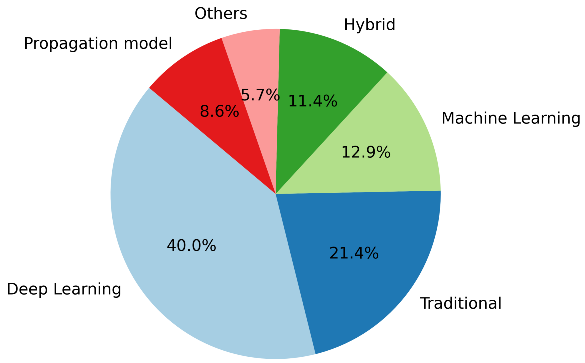


FIGURE 1: The overview of the WiFi fingerprinting data augmentation methods included in this review. It is observed that deep learning and traditional methods are most popular.

Based on the augmentation technique utilised, the included WiFi fingerprinting data augmentation methods are categorised into 6 groups: traditional methods, propagation model, machine learning methods, deep learning methods, hybrid methods, and other methods (see Figure 2). For the 70 included research papers, these six groups provide sufficient granularity for meaningful methodological comparisons while remaining cognitively manageable for readers. The taxonomy also allows for robust statistical analysis of methodological trends, cross-group comparisons, and identification of research gaps without creating excessive complexity that would obscure key patterns in the literature. Traditional data augmentation methods group are methodologies that enhance pre-collected WiFi fingerprint datasets by directly modifying the collected WiFi signal measurements, without additional modelling and training. Propagation model-based method group employ physics-based signal modelling to characterise how WiFi signals propagate through complex indoor environments, enabling the estimation of signal strength at unknown locations. Moving beyond purely physical models, machine learning methods group identify and model the inherent patterns within WiFi radio map using machine learning models and generate synthetic data at unsurveyed areas. In contrast, deep learning approaches, a subfield of machine learning, utilise neural networks with multiple layers to automatically learn and apply complex transformations directly to large amounts of WiFi signal measurements without any feature engineering. Hybrid method group contain methodologies that use more than one data augmentation methods or combine multiple augmentation techniques, typically integrating machine learning and deep learning approaches (e.g., utilising GPR+GAN [22], combining GAN and AE [23],etc.). Other method group categorise unique strategies that do not fall neatly into the aforementioned categories (e.g., Geography Weighted Regression [24], Singular Value Decomposition [25], Tensor Completion [26], etc.).

This section will present the fundamental principles of the six groups, supported by numerous examples from recent research studies.

A. TRADITIONAL METHODS

Traditional WiFi fingerprinting methods are a group of techniques that augment and enrich the pre-collected dataset by directly altering the recorded WiFi signal measurements, either in their preprocessed vector form or as transformed signal measurement images, to improve the performance of the positioning model. These preserve the original structure and existing hidden patterns of the fingerprinting data while adding controlled variations, introducing additional fluctuation or generating new data samples that closely follow the distribution of the original WiFi fingerprints. Since traditional methods require no extra model training, they are cost-effective and time-efficient in creating new training data samples. The traditional data augmentation methods employed in the literature include resampling, permutation, perturbation, and interpolation (see Figure 3). It is observed from Figure 3 that interpolation is the most popular traditional method owing to its capability to preserve the temporal and spatial continuity of WiFi signals while effectively filling gaps in sparse or noisy datasets without compromising structural integrity.

1) RESAMPLING AND PERMUTATION

Resampling involves feature-level recombination of WiFi signal measures, where new samples are synthesised by randomly selecting and combining the values from the WiFi readings from different Access Points (APs) at a reference point, while preserving their statistical distributions and spatial/temporal correlations. This was adopted by [27] where for each original RSS sample row, retain it as the first row and then generate multiple new synthetic samples by randomly selecting RSS values from the existing measurements for each AP at the same RP, as shown in Table 1. By generating 4819620 data samples from the 9620 input collected WiFi fingerprints, the proposed systems in [27] reduced positioning accuracy by 4.17 metres.

TABLE 1: In the resampling method utilised by [27], the first row “Aug 1” in the augmented dataset replicates the first row “1” of the original WiFi RSS data samples. For the second row “Aug 2” in the augmented dataset, the RSS values are randomly chosen from the collected measurements for each AP at the same RP as shown in dark grey colour.

# RSSI	AP1	AP2	AP3	AP4	AP5	AP6	AP7
1	34	37	36	54	26	38	26
2	41	37	0	56	34	38	21
3	42	37	36	54	0	38	0
4	42	38	36	56	33	36	22
5	39	36	44	50	37	38	18
6	42	0	39	0	0	0	0
7	41	37	0	57	0	35	27
8	44	38	39	0	30	38	0
9	44	41	41	54	21	40	25
10	40	38	39	0	26	38	0
11	45	44	0	0	0	44	0
12	30	38	41	48	0	37	0
13	32	42	45	50	0	41	0
14	33	45	45	50	0	43	0
15	32	45	45	50	33	45	0
16	33	44	0	49	0	45	21
17	36	45	40	51	39	45	0
18	37	43	40	51	0	43	19
19	37	45	46	47	32	41	17
20	37	45	44	0	0	45	0
Aug 1	34	37	36	54	26	38	26
Aug 2	42	38	36	0	39	38	17
				…			
				Augmented data samples			
Aug 501	36	44	0	50	32	41	0
Aug 502	41	37	0	56	34	38	21

Permutation in WiFi fingerprinting data augmentation is to randomly reorder the WiFi signal measure values in the fingerprints while maintaining the original statistical distribution. Synthetic data samples generated by permutation alters the sequence of WiFi signal measurements without changing their underlying characteristics. Given a WiFi data sample $\mathbf{R} = [R_1, R_2, R_3, \ldots, R_N]$ recorded at a RP from a total number of N APs, the data sample generated by permutation is defined as:

$$\mathbf{R}_{\text{perm}} = [R_{\sigma(1)}, R_{\sigma(2)}, \ldots, R_{\sigma(N)}] \tag{1}$$

where σ is a random permutation of indices [1, 2, …, N]. In [28], for each AP’s RSS measurements, the order of measurements is shuffled to generates new synthetic data matrices with the same statistical properties but different temporal arrangements of RSS values, claiming an improvement of 10%. However, no detailed empirical evaluation was provided in the study to support the results.

Resampling and permutation are basic feature-level data augmentation methods that **generate new WiFi data sam-**

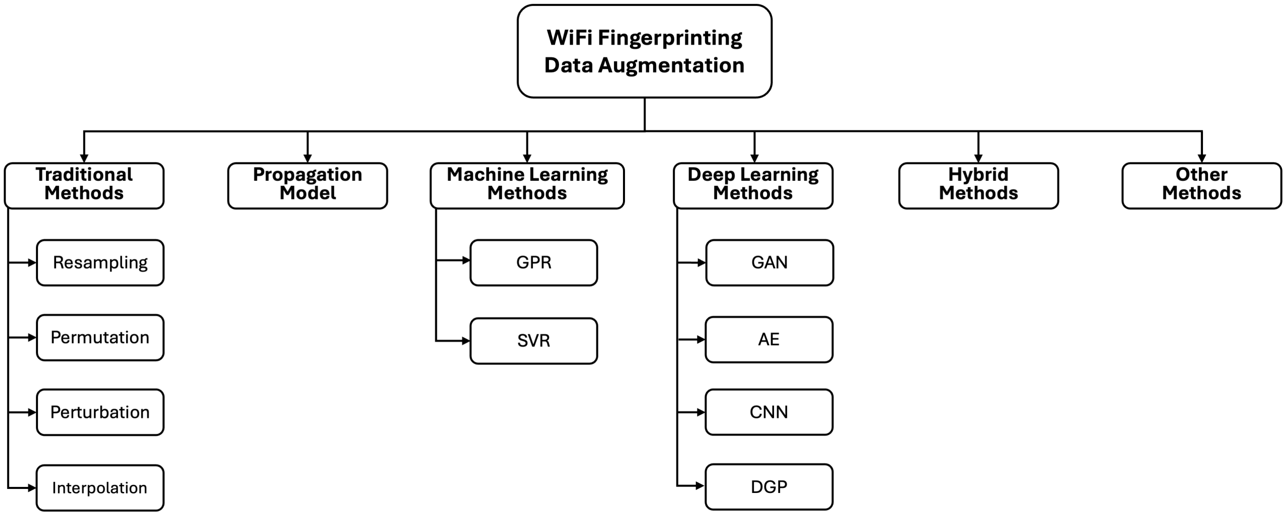


FIGURE 2: The novel taxonomy proposed to categorise the most popular data augmentation methods for WiFi fingerprinting utilised in the literature.

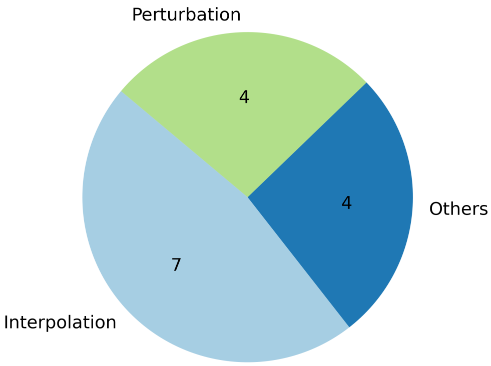


FIGURE 3: The distribution of the traditional methods employed in the research studies included in this review. It is observed that interpolation is the most popular traditional method because it effectively preserves the temporal and spatial continuity of WiFi signals and fills gaps in sparse or noisy datasets while maintaining structural integrity. .

ples by recombining and reordering of the collected WiFi signal measures, without introducing any unseen values.

2) PERTURBATION

Perturbation is a feature-level data augmentation technique that introduces variability or randomness into preconstructed WiFi fingerprinting datasets, aiming to enhance model generalisation, reduce overfitting, improve robustness, and explore alternative WiFi signal propagation behaviours.

In WiFi fingerprinting, perturbation method initially aims at adding deliberate, structured and controlled changes to the data to simulate realistic variations or domain-specific distortions. Subsequently, the variations from perturbation are often physics-guide or systematic-based. In [29], consistent phase shifts in CSI data were utilised to mimic imperfect transceiver clock synchronisation, while amplitude variations were used to mimic hardware gain instability or environmental attenuation (e.g., due to temperature). A simpler perturbation was proposed in [30] that for each RP's WiFi signal measurements, a fixed constant $C = -5$ is subtracted sequentially from each non-zero RSS value, creating N copies (where N= number of visible APs at the RP). The authors in [30] also employed a mean-constrained randomness to generates uniform random values between the original RSS and its mean (μ) for each RP. Another perturbation methods were proposed in [31] that introduce physical displacements (ΔL) to the original data's spatial coordinates. In other words, WiFi signal measurements were collected not only at the exact RP but also in their nearby spatial neighbourhoods, all of which were labelled with the same ground truth coordinates of the original RP.

Specifically, in a testbed consisting of M RPs, the objective of training a positioning model is to minimise the Mean Distance Error (MDE) (i.e., Original Loss) between predicted locations ($\hat{L}_m$) and true locations (L_m):

$$\text{Original Loss: } \mathcal{L} = \frac{1}{M}\sum_{m=1}^{M} \|\hat{L}_m - L_m\| \tag{2}$$

After the perturbation, the loss becomes:

$$\text{Augmented Loss: } \mathcal{L}_{\text{pertur}} = \frac{1}{M}\sum_{m=1}^{M} \|\hat{L}_m - (L_m + \Delta L)\| \quad (3)$$

where ΔL is small enough and satisfies $\|\Delta L\| \ll \|\hat{L}_m\|$. Compared to resampling and permutation, perturbation methods introduce unseen WiFi signal measurements to the positioning model by incorporating variability into the original fingerprinting datasets. Since WiFi signal measurements are highly sensitive to environmental changes and the multipath effect, particularly in non-line-of-sight (NLOS) conditions where fluctuations are more pronounced, introducing controlled or random disturbance to the fingerprinting data may enhance the quality and diversity of the datasets.

3) INTERPOLATION

As discussed in Section I, the collection and construction of a high quality WiFi fingerprinting dataset are labour-intensive. Therefore, researchers have proposed numerous methods to automatically construct a dense dataset based on sparse WiFi signal measurements.

Interpolation is a mathematical technique used to estimate unknown WiFi fingerprinting values such as RSS measurements at certain locations by leveraging known values at nearby reference points from pre-collected sparse datasets. Unlike resampling, permutation, and perturbation techniques discussed above, which primarily aim to expand the original complete dataset by generating additional data samples, interpolation is intended to reduce the effort involved in constructing dense fingerprint maps and to generate synthetic data for areas not covered during site surveys. As shown in Figure 3, 7 out of 15 included traditional method studies employed interpolation for WiFi fingerprinting data augmentation.

One common numerical model utilising polynomials for WiFi fingerprinting interpolation is defined as follows:

$$\hat{R}(dist) = a_i dist^3 + b_i dist^2 + c_i dist + d_i \quad (4)$$

where $\hat{R}(dist)$ is the interpolated WiFi signal measures at a specific distance $dist$, a_i, b_i, c_i, d_i are the coefficients calculated from measured WiFi fingerprints to ensure continuity and smoothness in fitting the WiFi signal measures. For instance, in [32] quadratic polynomial fitting was utilised to model the WiFi RSS-distance relationship for dense fingerprint maps generation, where a_i was set to zero, and $dist$ was the distance from the target point to the AP. In [33], dual-frequency bands (2.4 GHz and 5 GHz) and a cubic spline interpolation algorithm was leveraged to enhance localisation accuracy and efficiency, where $dist$ was the distance between the interpolated RP and the pre-set anchor RP used for data augmentation.

Another general formulation of WiFi fingerprinting interpolation using fingerprints from neighbour RPs is defined as follows:

$$\hat{R}(m) = \sum_{i=1}^{n_{sur}} \lambda_i R(m_i) \quad (5)$$

where $R(m_i)$ is the recorded WiFi signal measures at observation point RP_{m_i}, n_{sur} is the total number of surveyed RPs in the specified neighbourhood of the unsurveyed RP_m, λ_i is the weights solved via different interpolation methods. For instance, Kriging interpolation was utilised in [34] to densify fingerprint maps by modelling spatial drift and RSS variance via the weights λ_i generated by variograms. Variogram model is a function that describes the degree of spatial dependence in a dataset and defines how correlation decays with distance. By interpolating WiFi fingerprints for 84 RPs from a pre-collected set of 28 RPs (25% of the 112 manually collected RPs), the authors achieved a positioning accuracy that was only 0.093 metres lower, while requiring just 25% of the manual data collection effort. In [35], Linear Interpolation and Delaunay Triangulation were employed to create radio map with reduced calibration, where Linear Interpolation proved to have smaller WiFi RSS reconstruction error than Delaunay Triangulation. To achieve automated fingerprint database construction, an inverse distance weighting (IDW) based interpolation method with signal propagation model parameters named Signal-Propagated Modified Shepard's Method (SP-MSM) was proposed in [36] that outperforms IDW and Kriging at 52% of test points. Synthetic Minority Over-sampling Technique (SMOTE) is a statistical technique used to mitigate class imbalance in datasets. By leveraging interpolation, SMOTE generates synthetic data samples of the minority class between existing minority class instances, effectively increasing its representation in the fingerprinting dataset. This method was adopted by [37] to address the challenge of imbalanced datasets in Wi-Fi fingerprinting for indoor positioning systems. By generating synthetic fingerprints in areas that are hard to reach or not frequently visited, SMOTE helps to reduce the number of real data points required to construct a WiFi fingerprint dataset.

Since the interpolation methods either model the relationship between the distance and the signal measures or leverage fingerprints from surrounding RPs for synthetic data generation, they are easily affected by challenging conditions in complex indoor environment and by severe multipath effects in indoor spaces. At an unsurveyed RP, it is still difficult to accurately generate synthetic data samples to reflect the actual WiFi signal propagation properties with limited neighbouring RPs. Therefore, wireless signal propagation models grounded in physical principles have been employed in the literature to support data augmentation.

B. PROPAGATION MODEL

Although data augmentation techniques such as interpolation can densify WiFi fingerprinting datasets and generate synthetic samples to compensate for missing values, the inherent characteristics of WiFi signal propagation remain unknown. Under NLOS conditions in complex indoor environment as shown in Figure 4, wireless signals like WiFi suffer greatly from multipath effect, reflection, scattering, attenuation, fading and interference. Subsequently, the WiFi signal measures at cornered and unvisited locations where drastic interior changes happen remain a research challenge. While interpolation assumes smooth spatial variation of WiFi signal measurements in indoor space, propagation models take into account the physical behaviour of radio waves in complex indoor environment and the impact of obstacle and materials, thus producing more reliable synthetic data generation. Furthermore, in sparse but complicated fingerprinting scenarios, interpolation methods often produce poor performance due to their strong dependence on the availability of nearby RPs. In contrast, propagation models can estimate WiFi signal characteristics across broader areas by leveraging known AP locations and environmental parameters.

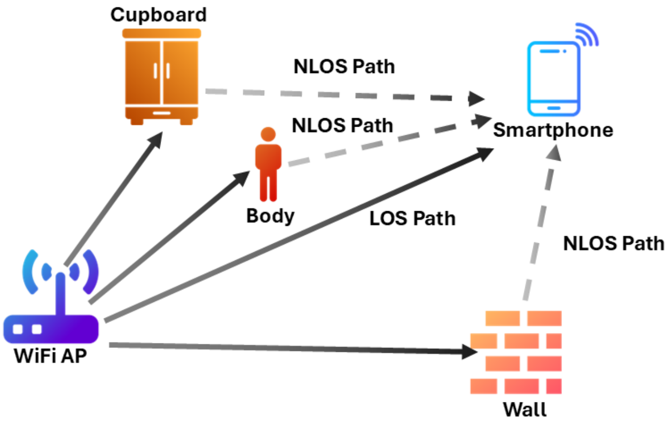


FIGURE 4: The WiFi signal suffers greatly from multipath effect, reflection, scattering, attenuation, fading and interference under NLOS conditions in complex indoor environment. .

The well-known Log-Distance Path Loss (LDPL) Model was utilised in [38] to account for indoor obstacles to estimate WiFi RSS values at unmeasured locations based on a small number of manually measured primary calibration points where Gaussian filtering was applied to smooth out RSS fluctuations. The LDPL model employed was defined as:

$$P(d) = P_0 - 10\alpha \log_{10}\left(\frac{d}{d_0}\right) + X_r \tag{6}$$

where P_0 is the reference power of WiFi signal at distance d_0, d_0 is the reference distance from the AP to the reference location where reference signal strength P_0 is measured, α is the path loss exponent determined by the actual indoor space characteristics, d is distance between AP and targeted location, and X_r is Gaussian random variable to mimic shadow fading.

Instead of applying a single propagation model across the entire indoor environment, the proposed method in [10] divided the indoor area into architectural zones, such as rooms or corridor segments, and independently estimates path-loss parameters for each zone using measurements from a small subset of reference points. The Zone-Based Remedy Algorithm proposed in [12] also adopted the zone-based idea and limited the search space to the most probable zone and bounded the path loss exponent within a realistic range. Specifically, the path loss exponent in Equation 6 was constrained within a range [α_{avg} - 1, α_{avg} + 1] to reduce error propagation due to extreme values, where α_{avg} is the average value of α among each direction and antenna set. Similar solution was employed in [39], where the authors proposed using a log-distance path loss model, whose parameters are estimated via a least squares approach, to generate RSS values at unmeasured locations at a distance d from the AP, defined as:

$$P(d) = \begin{cases} P_A - 10\alpha_n \log_{10}(d) + X_\sigma, & \text{if RSS} \geq \text{minRSS.} \\ minRSS, & \text{otherwise.} \end{cases} \tag{7}$$

where P_A is the WiFi RSS measure at 1-metre distance, α_n is the path loss exponent, X_σ is the Gaussian noise to model RSS fluctuation, $minRSS$ is the minimum detectable RSS threshold. Different from the basic Log-Distance Path Loss Model, the modified model estimated the parameters P_A and α_n for each small local area using the least squares method, based on a set of nearby sparse RPs. It also avoided physically implausible values by applying the $minRSS$ threshold.

To tackle indoor spaces with multiple walls and floors, the Multi-Wall (MW) propagation model based on the COST 231 indoor path loss framework was employed in [40]. The MW model accounts for signal attenuation due to both distance and environmental obstacles such as walls and floors, requiring site-specific information and AP locations, as defined below:

$$L(d) = L_0 + 10\alpha_{MW} \log_{10}(d) + \sum_w \beta_w W_w + \sum_f \beta_f F_f \tag{8}$$

where d is the distance between the AP location and the targeted location, L_0 is the reference loss at d_0, α_{MW} is the MW model path loss exponent, β_w and β_f are the attenuation per wall/floor, and W_w and F_f are the numbers of walls/floors in the path. Likewise, a Robust, cost-effective and scalable localization in large indoor areas (REAL) was proposed in [41], employing a propagation model enhanced by intelligent calibration techniques and taking into account the wall attenuation. The modified path-loss model leveraged is defined as:

$$P(d) = P_0 + \alpha_{REAL} \log_{10}(d) + \sigma_{REAL} N_{\text{ob}} + X_\epsilon \tag{9}$$

where P_0 is the WiFi RSS value at 1 metre from the AP, α_{REAL} is the attenuation factor due to distance, σ_{REAL} is the attenuation due to obstacles (e.g., walls), N_{ob} is the number of obstacles (walls) between AP and the targeted location, X_ϵ is the Gaussian modelling error. The REAL adapted based on available training data: for small training sets it assumed homogeneous APs (with the same P_0 and ϵ), while for larger training sets it modelled them separately for each AP (heterogeneous). Moreover, REAL always included wall count (N_{ob}) in the model, which significantly improved RSS prediction accuracy, particularly in complex indoor layouts.

It was observed that most propagation models used in the literature for WiFi fingerprinting data augmentation are based on the LDPL model. The key differences among them lie in how they account for signal attenuation caused by obstacles, walls, and floors, as well as in their use of zone division and refinement techniques to mitigate unrealistic WiFi signal measure estimates. Though propagation model is good for providing a basic understanding of signal attenuation over distance, it cannot capture complex, non-linear environmental factors. In contrast, machine learning and deep learning-based data augmentation methods can learn from real or simulated data, adapt to changing conditions, and generate realistic, and diverse datasets.

C. MACHINE LEARNING METHODS

Building on the limitations of numerical data augmentation approaches that leveraging traditional methods and propagation model, researchers have increasingly turned to machine learning-based data augmentation methods for WiFi fingerprinting. These approaches utilise the ability of machine learning algorithms to learn complex, non-linear patterns directly from data, without the need for manually defined rules or assumptions about signal behaviour. By training on real-world or simulated WiFI signal measurements, these models can better capture the intricate spatial and environmental dependencies that affect WiFi signal distributions in complex indoor environment.

Gaussian Process Regression (GPR) is one of the most popular machine learning-based data augmentation techniques for WiFi fingerprinting. GPR is a supervised machine learning technique used for regression tasks, where the goal is to predict continuous values [42], [43]. It is a non-parametric, Bayesian approach that models the underlying function as a distribution over possible functions, rather than assuming a fixed functional form. One of its key strengths is that it provides not only predictions but also a measure of uncertainty for each prediction, making it particularly useful in applications like WiFi fingerprinting where data may be sparse or noisy.

In GPR, a standard modelling of the joint distribution of the observed outputs $\mathbf{R}$ and the predicted outputs $\mathbf{R}^*$ as a multivariate Gaussian distribution is expressed as:

$$\begin{bmatrix}\mathbf{R}\\ \mathbf{R}^*\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{m}(\mathbf{X})\\ \mathbf{m}(\mathbf{X}^*)\end{bmatrix}, \begin{bmatrix}\mathbf{K}(\mathbf{X},\mathbf{X}) & \mathbf{K}(\mathbf{X},\mathbf{X}^*)\\ \mathbf{K}(\mathbf{X}^*,\mathbf{X}) & \mathbf{K}(\mathbf{X}^*,\mathbf{X}^*)\end{bmatrix}\right) \tag{10}$$

where $\mathbf{R}$ is the observed WiFi signal measures at RPs, $\mathbf{R}^*$ is the predicted signal measurement at unsurveyed RPs, $\mathbf{X}$ is the ground truth location coordinates of known RPs, $\mathbf{X}^*$ is the unsurveyed RPs where WiFi signal measures are to be generated, $\mathbf{m}(\cdot)$ is the mean function that defines the expected average value of the WiFi signal measure at a given location, $\mathbf{K}(\cdot)$ is the covariance function defining how similar or correlated the WiFi signal measures are between two locations.

This joint distribution was leverage in [44], [45] to predict RSS values at unmeasured locations based on a limited set of labeled reference points RPs. To improve the prediction accuracy, compound kernels such as a combination of the Matern kernel and the Rational Quadratic (RQ) kernel were adopted to capture both smooth trends and local variations. This allows GPR to model more diverse and nuanced relationships in the data by combining the strengths of different kernel types. The Matern kernel provides flexibility in modeling spatial correlations with controlled smoothness, while the RQ kernel acts as a scale mixture of squared exponential kernels, which allows it to model both short- and long-range signal correlations simultaneously. These compound kernels were also used in [46] to generate synthetic RSS values at virtual RPs generated by globally and locally self-adaptive approach in unsurveyed areas. The log-distance path loss model was also incorporated with GPR in [47] to reflect differences in signal propagation due to environmental factors like obstacles and multipath effects. To capture non-uniform WiFi RSS distributions more effectively in complex indoor environments, a second-degree polynomial surface fitting was utilised in [48] to determine the mean RSS distribution, defined as:

$$R(\mathbf{x}) = \beta_0 + \beta_1 lon + \beta_2 lat + \beta_3 lon^2 + \beta_4 lat^2 + \beta_5 lon \times lat \tag{11}$$

where $R(\mathbf{x})$ is the estimated mean WiFi RSS at location x, lon, lat are the coordinates of x, β_0 to β_5 are the coefficients learned from WiFi RSS measurements at known locations. Moreover, two simpler mean functions, Const-Linear (CL) and Quadratic Polynomial (QP), were proposed in [49] to better model the spatial distribution of RSS based on empirical signal characteristics observed in indoor environments, defined as:

$$R(\mathbf{x})_{CL} = \beta_0 + \beta_1 lon + \beta_2 lat \tag{12}$$

$$R(\mathbf{x})_{QP} = \beta_1 lon + \beta_2 lat + \beta_3 lon^2 + \beta_4 lat^2 \tag{13}$$

To jointly model the outputs of multiple APs in a multi-building, multi-floor environment, a Multi-Output Gaussian Process Regression (MOGP) model with Linear Model of Coregionalisation (LMC) was employed in [50], [51], defined as:

$$\mathbf{f}(\mathbf{X}) \sim \mathcal{MOGP}(\mathbf{m}(\mathbf{X}), \mathbf{K}(\mathbf{X}, \mathbf{X}^*)) \tag{14}$$

where $\mathbf{f}(\mathbf{X})$ represents the output vector of WiFi values from multiple APs. The joint distribution is then given by:

$$\begin{bmatrix} \mathbf{f}(\mathbf{X}) \\ \mathbf{f}(\mathbf{X}^*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{Au}(\mathbf{X}) \\ \mathbf{Au}(\mathbf{X}^*) \end{bmatrix}, \begin{bmatrix} \mathbf{AK}(\mathbf{X},\mathbf{X})\mathbf{A}^\top & \mathbf{AK}(\mathbf{X},\mathbf{X}^*)\mathbf{A}^\top \\ \mathbf{AK}(\mathbf{X}^*,\mathbf{X})\mathbf{A}^\top & \mathbf{AK}(\mathbf{X}^*,\mathbf{X}^*)\mathbf{A}^\top \end{bmatrix} \right) \tag{15}$$

where $\mathbf{A}$ is the coregionalisation matrix that linearly maps latent functions to outputs, $\mathbf{u}(\cdot)$ is a set of latent functions, each independently modelled as a Gaussian Process (GP), capturing underlying spatial characteristics, such as WiFi signal propagation behaviour that influence multiple APs. The MOGP leveraged captured inter-AP correlations that are particularly significant when APs are spatially close (e.g., located on the same floor), and therefore enhance the synthetic data generation.

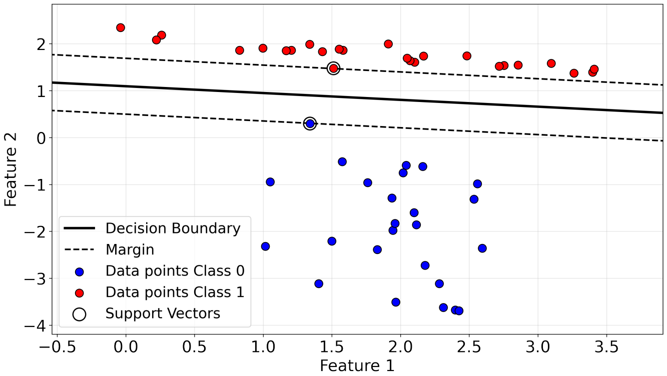


FIGURE 5: The basic SVM algorithm, which finds the optimal decision boundary (solid black line) that best separates two classes of data (red and blue points). The dashed lines represent the margin, and the support vectors (circled points) are the critical data points that lie on the edge of this margin and influence the position of the boundary.

Other machine learning-based data augmentation method was reported in [52], where a Support Vector Regression (SVR) model with linear kernel function was proposed leveraging spatial and environmental features. Support Vector Machines (SVM) are supervised learning algorithms that identify an optimal hyperplane for separating data into distinct classes while maximizing the margin between them, ash shown in Figure 5. As an extension of SVM, SVR aims to find an optimal hyperplane for predicting continuous target values, maximizing the margin while allowing deviations within a specified tolerance threshold. The regression function in Linear SVR is defined as:

$$f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b \tag{16}$$

where $f(\mathbf{x})$ is the predicted RSS values, $\mathbf{w}$ is the weight vector learned from the data to find the best-fitting regression line while minimising prediction errors, b is the bias term, $\mathbf{x}$ is the input variables. Specifically, the input variables the proposed method used to model WiFi signal behaviour were the positions of the RPs and APs, their in between distance, and obstacle information, derived from a floor plan using the Bresenham algorithm [52]. This allows for accurate RSS estimation, especially for strong signals and in environments where obstacles significantly affect signal propagation.

D. DEEP LEARNING METHODS

While machine learning methods like GPR have shown promising results in modelling complex signal distributions for WiFi fingerprinting, they often rely on carefully engineered features and may struggle with scalability in high-dimensional environments. To address these limitations, researchers have begun exploring deep learning-based data augmentation techniques for WiFi fingerprinting, which can automatically learn hierarchical representations of data and generate realistic synthetic samples. Models such as Generative Adversarial Networks (GANs), Autoencoders (AEs), Super-resolution Convolutional Neural Network (CNN) and Deep Gaussian Process (DGP) have emerged as powerful tools in the literature, as shown in Figure 6. These deep generative models are capable of capturing intricate spatial and temporal patterns in WiFi signals automatically, enabling the generation of large, diverse datasets that enhance the robustness and accuracy of indoor positioning systems.

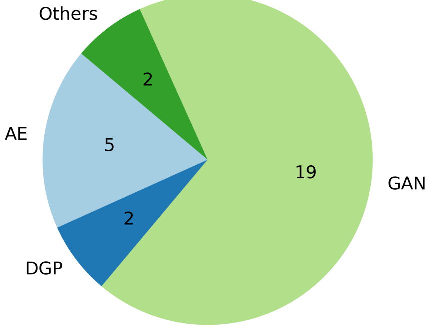


FIGURE 6: Overview of the deep learning methods for WiFi fingerprinting data augmentation in this review. GANs are the most widely used deep learning approaches in the literature because they generate highly realistic synthetic WiFi data samples by adversarially learning complex signal distributions and intricate patterns.

1) GENERATIVE ADVERSARIAL NETWORK

Generative Adversarial Networks (GANs) are a powerful class of deep learning models for synthetic data generation to enrich and improve the WiFi fingerprinting datasets [53]–[56]. A standard GAN is composed of two competing neural networks: a generator that creates synthetic data sample and a discriminator that attempts to distinguish between real and generated data samples (see Figure 7).

In a standard GAN architecture, the generator (G) is trained to produce synthetic WiFi fingerprinting data samples, by transforming a random noise vector (z) into a synthetic (fake) fingerprint. The discriminator (D) receives two types of inputs: real fingerprints from the training dataset and synthetic fingerprints generated by G, and it learns to classify them as REAL or FAKE. The generator is trained to produce synthetic fingerprints that are indistinguishable from real ones, while the discriminator is trained to correctly identify whether a fingerprint is real or generated. This adversarial training process continues until the discriminator can no longer reliably distinguish between real and fake samples. The objective function of GAN is defined as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))] \quad (17)$$

where $\mathbf{x}$ is the real data samples from the training fingerprint dataset, $\mathbf{z}$ is the random noise vector, $D(\mathbf{x})$ is the discriminator's estimate of the probability that $\mathbf{x}$ is real, $G(\mathbf{z})$ is the synthetic WiFi data samples generated from noise $\mathbf{z}$, $\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}(\mathbf{x})}$ is the expected or average value of the real WiFi data samples, $\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}$ is the expected or average value over all possible noise vectors $\mathbf{z}$ drawn from the distribution $p_{\mathbf{z}}(\mathbf{z})$, $V(D,G)$ is the objective function of the min-max game between the discriminator D and generator G, where D tries to maximise the value (i.e., become better at distinguishing REAL from FAKE), G tries to minimise the value (i.e., become better at fooling the discriminator). A standard GAN was leverage in [57] and [58] to enrich the fingerprint datasets by filling spatial gaps left by sparse crowdsourced data and a public WiFi fingerprint dataset, respectively.

Additionally, variants such as Conditional GANs (cGANs) can be used to generate fingerprints conditioned on specific locations or environmental contexts, making the synthetic data even more useful for location or region-aware positioning, as shown in Figure 7. The objective function of cGAN under the conditional label y is defined as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}[\log D(\mathbf{x}|y)] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}}[\log(1 - D(G(\mathbf{z}|y)))] \quad (18)$$

Unlike a standard GAN, a conditional GAN incorporates an additional input, such as a floor or building label, into both the generator and discriminator. This allows the generator to produce synthetic fingerprints tailored to specific locations, and the discriminator to evaluate them in conditional context. By conditioning on label information, cGANs enable controlled and location-aware data generation, making them particularly effective for augmenting data across diverse regions in complex indoor environments. In [59], a conditional GAN with conditional label of building and floor IDs was utilised. The synthetic WiFi RSS signal measures were further filtered using a distance-based algorithm to ensure only those close to real samples are retained, enhancing the WiFi fingerprint datasets without degrading model performance. A conditional GAN with 0-1 Sketch was also proposed in [60] enabling more effective and stable generation. The authors in [61] proposed Semi-Supervised GAN, levering a generator that accepted both noise vector and location labels as input to produce location-specific RSS fingerprints, avoiding generating only unlabelled data like standard GAN. In [61] a shared network architecture between the discriminator and classifier was also employed to enable simultaneous data authenticity verification and location classification.

To generate synthetic CSI data, a Amplitude Feature Deep Convolutional GAN (AF-DCGAN) method was proposed in [21] where raw CSI data is transformed into amplitude feature maps that visually encode spatial signal characteristics. DCGAN was also utilised in [62], composed of a generator, a discriminator, and a classifier with shared weights. The generator creates synthetic CSI fingerprints from random noise, which are used alongside real unlabelled samples to train discriminator and classifier. The authors in [63] converted multidimensional WiFi signal measurements into low-resolution (LR) fingerprint images, which are then augmented using an Enhanced Super-Resolution GAN to generate high-resolution (HR) images. These HR images are subsequently transformed back into augmented signal fingerprints, effectively densifying the fingerprint database.

In [64], the authors introduce Tensor-GAN, modelling Wi-Fi fingerprints as a 3D low-tubal-rank tensor, effectively capturing spatial and signal correlations. The generator is designed as a tensor completion algorithm operating on tubal-sampled data, enabling it to generate realistic synthetic fingerprints from sparse samples. Tensor-GAN was also reported in [65], which employs a transform-based 3D tensor model to represent WiFi fingerprint samples. The generator within TGAN, is designed to encode coarse-resolution fingerprint tensors into sparse representations, which are then transformed into higher-resolution fingerprint tensors.

Wasserstein GANs (WGANs) and a Pseudo Fingerprint Map (PFM) were combined to enhance Wi-Fi fingerprinting in 3D space in [66]. In WGAN, the traditional discriminator is replaced by a ”critic” that assigns real-valued scores to inputs to estimate the Wasserstein distance, a smoother and more meaningful measure of how much effort it takes to transform one probability distribution into another. To generate synthetic CSI data samples, WGANs were used in [67], clustering reference fingerprints using K-means

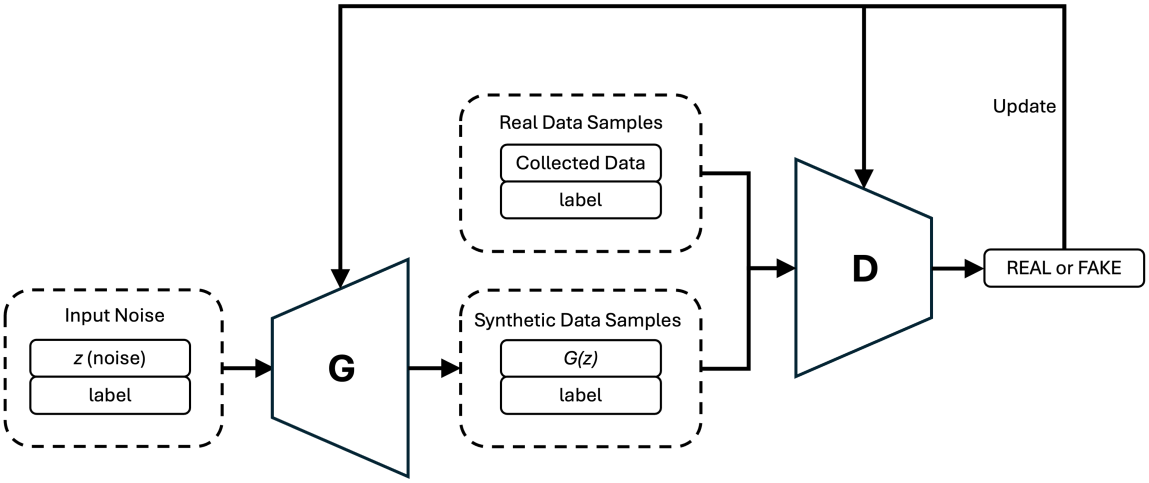


FIGURE 7: The structure of Generative Adversarial Networks. The generator (G) is trained to produce synthetic WiFi fingerprinting data samples, by transforming a random noise vector (z) into a fake fingerprint. The discriminator (D) receives two types of inputs: real fingerprints from the training dataset and synthetic fingerprints generated by G, and it learns to classify them as REAL or FAKE. Note that the labels are conditional labels for cGANs.

to identify spatial regions, determining the cluster of the point to be located. A hybrid data augmentation framework that combines Dirichlet distribution-based upsampling and Wasserstein GAN with Gradient Penalty (WGAN-GP) named extendGAN+ was proposed in [68]. The extendGAN+ first identified the location with the most data points to train a base WGAN-GP model, which was then transferred and finetuned using transfer learning to generate synthetic RSS data for unsurveyed locations. Additionally, a filtering module ensures the quality of generated samples by removing outliers based on dissimilarity thresholds.

2) OTHER DEEP LEARNING METHODS

While GANs are effective for generating realistic synthetic WiFi fingerprint data through adversarial training, Autoencoders (AEs) offer a simpler but stable alternative for data augmentation. As shown in Figure 8, AEs are unsupervised neural networks that learn efficient low-dimensional representations (encodings) of input data and can reconstruct the original input from these encodings (decodings). In the context of WiFi fingerprinting, AEs can be trained to learn compressed representations of RSS vectors and then used to generate new, diverse fingerprints by introducing controlled perturbations in the latent space before decoding. By incorporating dropout regularisation during training, a dropout AE was proposed in [69] to prevent overfitting and improve the model's ability reconstructing missing or noisy WiFi data. To learn more robust and hierarchical feature representations that are resilient to input perturbations, Stacked Denoising Autoencoder (SDAE) was used in [70] by introducing noise to the input data and stacking multiple denoising layers. Furthermore, Variational Autoencoders (VAEs) were employed in the literature. VAE is a probabilistic extension of AEs, allow for sampling from a learned latent distribution, making them particularly well-suited for generating new fingerprint samples that maintain the structural characteristics of real data. In [71], a VAE was used to learn the underlying distribution of real RSSI data collected from low-cost ESP32 devices. RSS readings from eight reference nodes were reshaped into 3×3 grayscale images and fed into a convolutional VAE with a 2-dimensional latent space, enabling the generation of synthetic WiFi RSS values. The conditional VAE extending the traditional VAE by incorporating additional conditional information (e.g., location or floor data) was leveraged in [72] enabling it to generate context-aware and more controllable synthetic data tailored to specific scenarios.

In addition to the popular GAN- and AE-based WiFi fingerprint data augmentation methods, other deep learning methods have also been explored. DeepMap, a novel system that applies Deep Gaussian Processes (DGPs) was proposed in [73] and [74] to address the challenges of indoor radio map construction. The authors identified the limitation of conventional GPR methods in modelling non-stationary RSS, particularly under conditions of sparse training samples. To overcome this, they utilised a two-layer DGP model that learns the latent, nonlinear relationship H between RSS values and spatial locations, leveraging a Bayesian variational inference method to train the model offline and optimise the marginal likelihood. In the Fingerprint Augment based on Super-Resolution (FASR) framework proposed by [19], sparse WiFi fingerprint data were converted into low-resolution images and then fed into super-resolution CNN

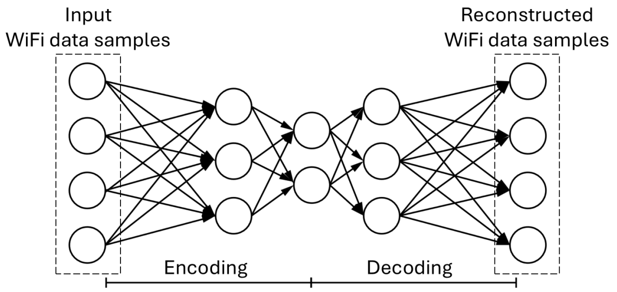


FIGURE 8: The structure of the Autoencoder methods for WiFi fingerprinting data augmentation. Autoencoders learn efficient low-dimensional representations of input WiFi data in the encoding part, and reconstruct the WiFi data samples from these input in the decoding part.

models to generate high-resolution fingerprint images. These enhanced images were then converted back into augmented WiFi fingerprint databasets. Inspired by Between-Class (BC) learning from image and sound recognition domains, the authors in [75] proposed a multi-layer regression (MLR) network to estimate RSS distributions at unsurveyed locations by modelling each access point's signal as a Gaussian distribution conditioned on location.

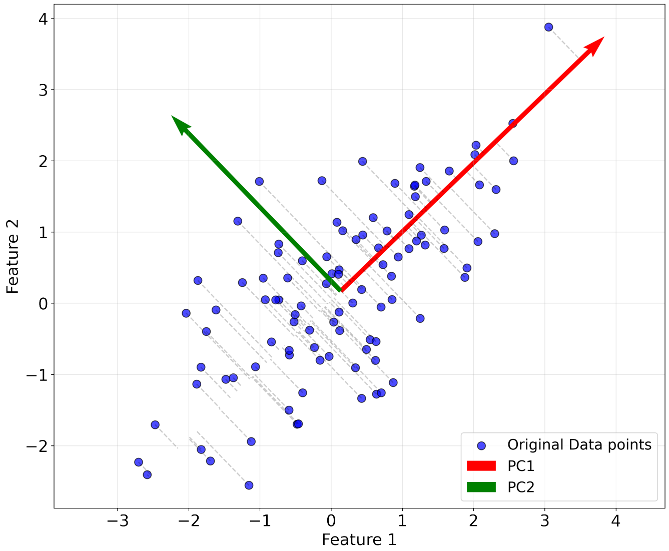


FIGURE 9: The basic implementation of PCA, a technique used to reduce the dimensionality of data by projecting it onto new axes (PC1 in red and PC2 in green) that capture the most variance. PC1 is the direction of maximum variance, while PC2 is orthogonal to it and captures the remaining variance.

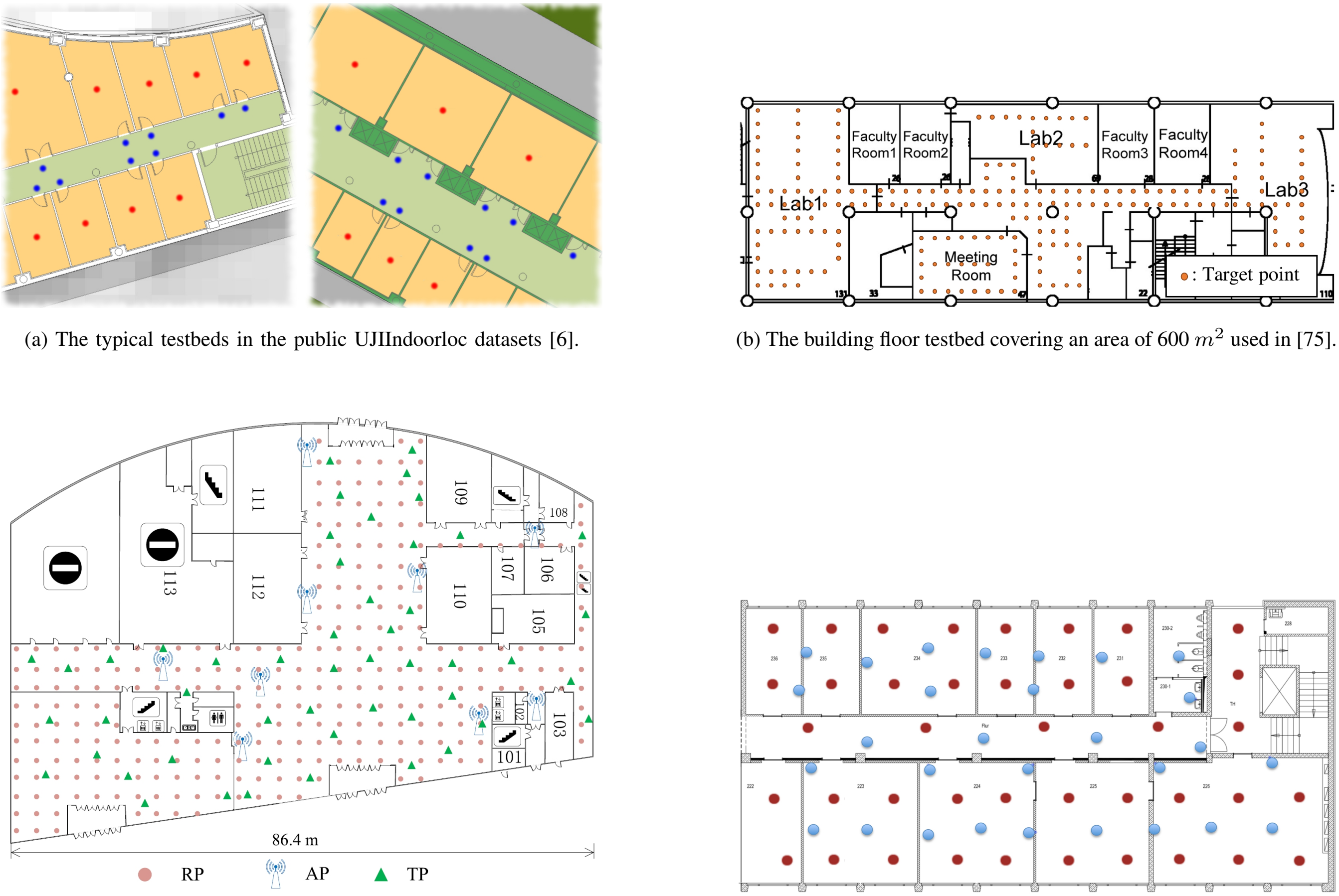
E. HYBRID METHODS AND OTHER METHODS

In addition to the previously introduced data augmentation techniques, researchers in the field often combine multiple methods to enhance the quality of the generated synthetic WiFi data samples.

Numerous methods were proposed to enhance the synthetic data generation by GPR. For instance, a novel data augmentation method leveraging Principal Component Analysis (PCA) and GPR was proposed in [76]. PCA, as shown in Figure 9, was employed to select the most informative APs while GPR was used to model the relationship between the RP coordinates and the RSS values, enabling the generation of synthetic WiFi fingerprinting data for unsurveyed locations. And in the framework proposed in [77], GPR was employed to generate initial fingerprints which were then refined using a VAE to capture inter-access-point relationships. The KMGPR (K-Means Gaussian Process Regression) was proposed in [78] that used GPR with a Gaussian mean function to more accurately model WiFi RSS values. To reduce GPR's high computational cost, K-Means was applied to divide RPs, enabling parallel GPR processing within each cluster and supporting efficient WiFi fingerprinting data updates.

Several hybrid GAN-based approaches have been proposed in the literature. These include the combination of linear interpolation with GANs [20], the integration of deep neural networks (DNNs) with GANs [79], the fusion of GPR and Least Squares GANs (LSGANs) [22], as well as the combination of AE with GANs in [23]. Specifically, DNN was trained on the original labeled data to add pseudo-label for realistic synthetic data generated by GAN in [79]. In [22], GPR was first used to provide coarse RSS estimations at unsurveyed locations (constrained spaces like cubicles and private offices). These estimations are then used as structured input to the LSGAN generator. To validate the performance of GPR-LSGAN, a mobile robot equipped with LiDAR SLAM and WiFi sensing capabilities was utilised. To adaptively update the WiFi fingerprinting dataset by filtering and incorporating new APs while preventing unnecessary expansion, WiFi RSS signal measures from a single reference floor and the spatial layout of APs on other floors were utilised in [23] to train the AE and GAN, respectively.

Other novel methods were also utilised in the literature to model the WiFi fingerprinting data augmentation from a different perspective. A RSS Geography Weighted Regression (RGWR) method was proposed in [24] by taking into account the variation in signal attenuation across different regions with weighting spatial correlations to automatically construct and update WiFi fingerprinting dataset using self-made low-power WiFi anchors alongside modified APs. The Sparsity Rank-Singular Value Decomposition (SRSVD) method was proposed in [25] to recover missing fingerprint data effectively. The SRSVD combined with the K-Nearest Neighbor (KNN) leverages the spatial and temporal correlations of fingerprints, which result in a low-rank matrix, to handle missing columns or rows in the matrix. To effectively reconstruct and update the WiFi fingerprinting datasets, Regularised Singular Value Decomposition (RSVD) method



(a) The typical testbeds in the public UJIIndoorloc datasets [6].

(b) The building floor testbed covering an area of 600 m^2 used in [75].

(c) The building floor testbed covering an area of 4000 m^2 used in [33].

(d) The building floor testbed covering an area of about 450 m^2 used in [40].

FIGURE 10: The typical testbeds for WiFi-based indoor positioning data augmentation leveraged in the literature.

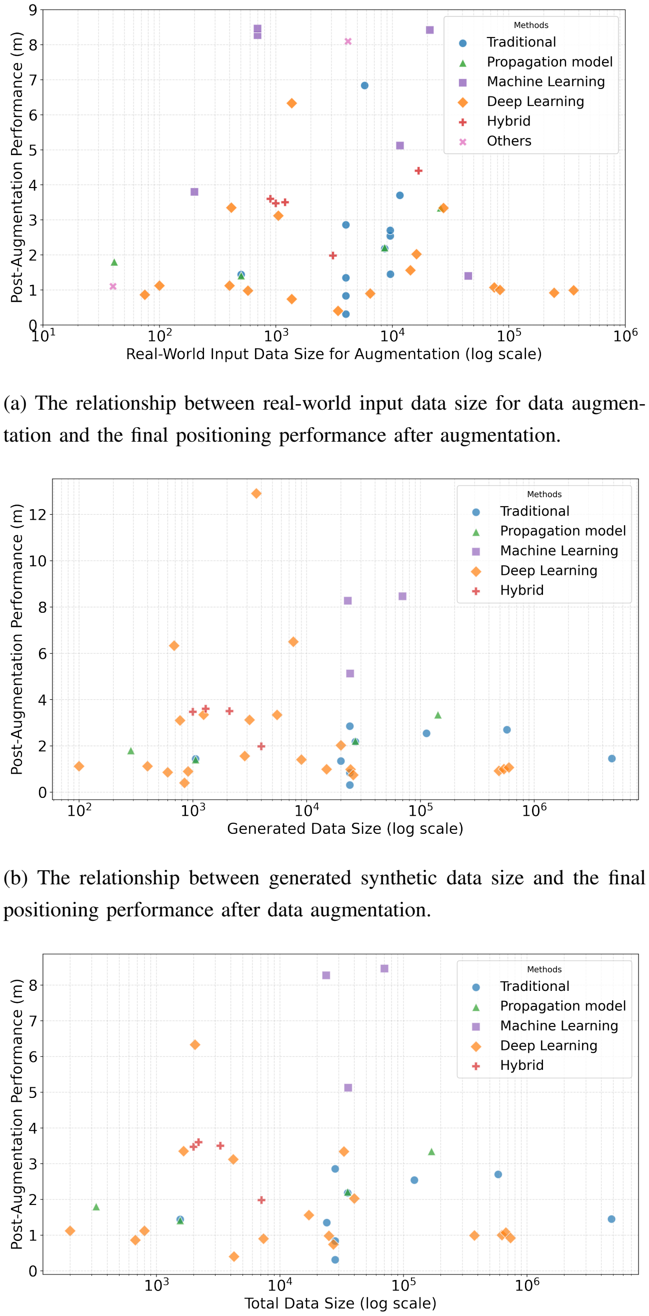
was employed in [11] that reconstruct the fingerprint matrix with measurements from a small number of RPs and employs the stability of RSS differences between neighbouring RPs and adjacent links to mitigate short-term RSS variations. The WiFi fingerprinting dataset construction problem was formulated as a Low-Rank Tensor Completion (LRTC) issue in [26]. The LRTC method leverages the strong correlations within the fingerprinting data to estimate RSS values at unmeasured RPs, hence reducing human efforts in the collection while maintaining high positioning accuracy. In the DataLoc+ proposed in [80], stochastic corruption and smoothing were combined by collecting streams of WiFi AP signals, then generating multiple augmented snapshots by randomly shuffling and sampling varying portions of the AP data. These snapshots simulate real-world signal variations (e.g., fading, shadowing) by including fewer or more APs and signal strength variations.

IV. DATA AUGMENTATION PERFORMANCE ANALYSIS

To offer valuable insights into the most widely adopted augmentation techniques, a detailed and comprehensive performance analysis of WiFi fingerprinting data augmentation methods is provided (see Table 2). Through meticulous, time-consuming, and often painstaking manual analysis and assessment, the collected data size, input data size for augmentation, generated synthetic data size, and positioning performance metrics (both before and after augmentation) are carefully extracted, recorded, and compared. This process is exceptionally difficult and labour-intensive because of incomplete or missing information, manual extraction from figures and tables and inconsistent reporting standards. As discussed in the footnote of Table 2, for papers offering positioning accuracy in Mean Distance Error (MAE) and Root Mean Square Error (RMSE), we record the results as presented. For papers measuring the augmentation performance using zone hit rate, only the positioning accuracy in percentage (%) are recorded in Table 2. For papers that didn't present their indoor positioning performance explicitly, we either manually extracted them from Cumulative Distribution Function (CDF) curve or results histograms. For systems assessing the WiFi RSS data augmentation accuracy, the error in dBm is used. This highlights a key challenge in current WiFi data augmentation methods: the absence of guidelines or standards for evaluating and presenting the performance of these methods.

Excluding studies that did not report data sizes explicitly or implicitly, 50 different data augmentation methods were left for the following performance analysis. In this analysis, the indoor positioning performances of the WiFi fingerprint datasets before and after different data augmentation are compared. In addition, we examine the relationship between positioning performance and several factors: the sample size of input data for data augmentation, the sample size of generated synthetic WiFi data, the total number of WiFi fingerprint samples, and the generation ratio of synthetic to input data. To assess the influence of data augmentation on WiFi fingerprinting, we further evaluate its contributions to positioning accuracy improvements to assess the impact of redundancy in training data. It is worth noting that only two of the included studies employed CSI [29], [81], and none utilised RTT. Consequently, this review does not examine the impact of different WiFi fingerprinting features on augmentation performance. For the same reason, namely, the insufficient number of included studies reporting hardware specifications and how often models can be updated, this review does not evaluate the hardware or computational resources required and the real-time feasibility of the included methods. Since the included research papers evenly collected the data samples per RP, the impact of class imbalance will not be discussed.

To investigate the correlation between WiFi data samples and indoor positioning performance, we manually extracted the input data size, generated data size, and overall data size (input+generated) from the included data augmentation methods. As shown in Figure 11, deep learning-based approaches generally demonstrate improved positioning accuracy with an increased number of input data samples for augmentation training. **Interestingly, most traditional method-based data augmentation approaches achieved reliable positioning accuracies of less than 4 metres, whereas some machine learning-based methods resulted in positioning accuracies exceeding 8 metres.** In addition, increasing the input to propagation model-based methods appeared to reduce positioning accuracy. This is because in propagation model-based systems, an increase in input data corresponds to a larger testbed. This expansion introduces numerous complex indoor scenarios that significantly degrade the accuracy of single WiFi signal propagation model. In examining the relationship between both generated and total (input+generated) data size, and positioning performance across various techniques, including deep learning, traditional, propagation model, and hybrid approaches, it is observed that increased data size contribute to enhanced positioning accuracy.



(a) The relationship between real-world input data size for data augmentation and the final positioning performance after augmentation.

(b) The relationship between generated synthetic data size and the final positioning performance after data augmentation.

(c) The relationship between total data size and the final positioning performance after data augmentation.

FIGURE 11: The correlation between post-augmentation performance and data sizes for WiFi fingerprinting. Performances are analysed against three key variables: input data size, generated (augmented) data size, and total data size (input + generated), all plotted on logarithmic scales. It is observed that larger WiFi fingerprint data size cannot always ensure a more robust positioning accuracy, especially below 100,000 data samples.

TABLE 2: The indoor positioning performance comparison of the WiFi dataset before and after data augmentation.

Paper	Testbed type	Testbed size † (m^2)	Collected data size	Augmentation method	Input data size*	Generated data size	Pre-Aug perf** (m)	Post-Aug perf** (m)
[28]	University building	˜36000	5785	Permutation	5785	N/A	7.5975	6.83775
[44]	Building floor	593.92	11640	GPR	11640	24120	6.914	5.121
[30]	Building floor	1664	10360	Perturbation	9602	113158	5.62	2.54
[30]	Building floor	1664	10360	Perturbation	9602	576120	5.62	2.7
[80]	Hospital	1 Floor	450	Corruption and smoothing	450	1800-2250	84.40%	100.00%
[60]	Building floor	273	120 RPs	conditional GAN	120 RPs	N/A	1.235	0.678
[19]	Office room, simulation	52.25, 64	627000, 675000	Super-resolution CNN	84000, 75000	543000, 600000	1.16, 1.11	1, 1.07
[76]	Reading room	1000	727 RPs	GPR+PCA	61 RPs	994 RPs	3.2	4
[20]	Building floor	173, 51	1369, 14300	Interpolation+GPR	1369, 14300	2m resolution, 0.1 resolution	7.44, 1.67	5.05, 1.51
[20]	Building floor	173, 51	1369, 14300	GAN	1369, 14300	684, 2860	7.44, 1.67	6.33, 1.56
[27]	Building floor	1664	9620	Resampling	9620	4819620	5.62	1.45
[79]	Building floor	˜9000	1000	GAN+DNN	1000	1000	4.1	3.47
[45]	University building	22570	697	GPR	697	Very dense	8.94	8.46
[24]	Building floor, shopping mall	550, 2875	52 RPs, 163 RPs	GWR	N/A	52 RPs, 4890	N/A	2.4, 5.1
[32]	Laboratory room	150	48 RPs	Interpolation	16 RPs	32 RPs	1.03	1.01
[36]	Office room	92.4	64 RPs	Interpolation	17 RPs	47 RPs	2.72 ± 1.12	3.06 ± 1.54
[38]	Building floor	1600	$>$ 1600	Propagation model	>1600	˜140 out of 150 RPs	N/A	0.6
[21]	Classroom	49	245000	DCGAN	245000	490000	1.34	0.92

† For data augmentation methods evaluated on different testbeds, we used comma to separate them in the same line.

* The Input Data is the size of the input WiFi data to the augmentation methods.

** The Pre-Aug perf & Post-Aug perf are the positioning accuracies of the system before and after the data augmentation. Due to the challenges in consistently reporting positioning performance in the reviewed research publications, m is used for regression systems while the hitting rates (%) are used for classification systems. For systems assessing the WiFi RSS data augmentation accuracy, dBm is used.

Paper	Testbed type	Testbed size † (m^2)	Collected data size	Augmentation method	Input data size*	Generated data size	Pre-Aug perf** (m)	Post-Aug perf** (m)
[82]	Building floor	5376	22000	cGAN	22000	5500	1.23	3.34
[22]	Building floor	700	3100	GPR+LSGAN	3100	4000	1.79	1.98
[46]	University building	22570	697	GPR	697	23,002	8.94	8.27
[73]	Building floor	2300	157 RPs	DGP	157 RPs	N/A	1.569	3.8447
[72]	University building	Two building floors	32	cVAE	≫32	3612	32.76	12.91
[57]	Corridor, laboratory	135, 120	0.5 Hz	GAN	N/A	N/A	1.8870, 2.5136	2.0902, 1.8945
[40]	Building floor	˜450	41	Propagation model	41	285	2.5	1.8
[39]	Building floor	3200	168960	Propagation model	25920	142640	3.341	3.344
[74]	Building floor, building floor	2300, 860	157 RPs, 82 RPs	DGP	157 RPs, 82 RPs	N/A	1.569, ˜2.1	3.8447, ˜6.2
[37]	University building	N/A	2000	SMOTE	1500	500	88.74%	88.10%
[59]	University building	DSI 1, DSI 2	1369, 576	cGAN	1369, 576	25712, 24350	0.88, 1.29	0.74, 0.98
[23]	University building	Four building floors	63 RPs	GAN+AE	63 RPs	N/A	89.66%	88.59%
[61]	Office building	Five building floors	1700	GAN	1700	N/A	˜90.5%	˜91.5%
[48]	Laboratory room	590.96	50 RPs	GPR	N/A	288 RPs	3.389	1.718
[63]	Conference room	12	60 RPs × 2 min	GAN	60 RPs × 2 min	10000 RPs × 1 min	2.4337	2.1094
[58]	Shopping mall	One shopping mall floor	N/A	GAN	N/A	100% of the collected	˜87.5%	˜96%
[25]	Building floor	N/A	4160	SVD	4160	N/A	˜7.5, 80% of the time	˜8.1, 80% of the time
[47]	Corridor	1050	125 RPs	GPR	125 RPs	All of the APs from 15 TPs	1.8	1.7
[52]	Two library floors	300	92 RPs	SVR	8 RPs	N/A	3.41	3.94

† For data augmentation methods evaluated on different testbeds, we used comma to separate them in the same line.

* The Input Data is the size of the input WiFi data to the augmentation methods.

** The Pre-Aug perf & Post-Aug perf are the positioning accuracies of the system before and after the data augmentation. Due to the challenges in consistently reporting positioning performance in the reviewed research publications, m is used for regression systems while the hitting rates (%) are used for classification systems. For systems assessing the WiFi RSS data augmentation accuracy, dBm is used.

Paper	Testbed type	Testbed size † (m^2)	Collected data size	Augmentation method	Input data size*	Generated data size	Pre-Aug perf** (m)	Post-Aug perf** (m)
[64]	University building	1600	N/A	Tensor GAN	Using 40% of the collected	N/A	0.32, 80% of the time	0.19, 80% of the time
[77]	University building	˜500	16850	GPR+VAE	16850	55 RPs	5.16, 95% of the time	4.4, 95% of the time
[10]	Building floor, office	918, 252	35400, 1560	Propagation model	8600, 504	26800, 1056	1.45, 1.19	2.21, 1.41
[10]	Building floor, office	918, 252	35400, 1560	Interpolation	8600, 504	26800, 1056	1.45, 1.19	2.18, 1.44
[78]	Building floors	560, 3500	2100, 1300	GPR+K-means	1200, 900	2100, 1300	˜3, 80% of the time, ∼3.2, 80% of the time	∼3.5, 80% of the time, ∼3.6, 80% of the time
[31]	Lab	11.52	15 RPs	Perturbation	N/A	N/A	1.42	0.97
[75]	Building floor	600	1050	MLR	1050, 415	3150, 1245	3.58, 3.97	3.12, 3.35
[62]	Office room	49.88	6400	DCGAN	6400 labeled, 32 labeled+6400 unlabeled	6400, 32	87.71%, 68.78%	87.84%, 85.78%
[49]	Office room, building floor	52.25, 560	627000, 2100	GPR	45000, 200	N/A, N/A	1.3, 80% of the time, 3, 80% of the time	1.4, 80% of the time, 3.8, 80% of the time
[11]	Office room	108	470	SVD	40	N/A	0.78, 50% of the time	1.1, 50% of the time
[65]	Indoor region	96	∼119 RPs	Tensor GAN	∼119 RPs	∼1080 RPs	8, 90% of the time	4, 90% of the time
[83]	Four rooms	N/A	2000	GAN	1000	1000	95.30%	97.20%
[50]	University building	108703	21049	Multi-Output GPR	21049	N/A	8.61	8.42
[71]	Classroom	151.4	1489	VAE	1489	N/A	N/A	3 dBm
[84]	Building floor, shopping mall	total 8400	N/A	cVAE	N/A	772, 7642	4.5, 8.5	3.1, 6.5
[85]	Simulation	10000	550 RPs	DCGAN	550 RPs	N/A	2, 8.23	1.68, 5.57
[81]	Laboratory room	33.75	8960000	GAN	1792000, 8960000	44800000, 44800000	71.2%, 93.3%	95.1%, 98.2%
[67]	Indoor region	100	100, 400	WGAN	100, 400	100, 400	1.376, 1.376,	1.12, 1.12,

† For data augmentation methods evaluated on different testbeds, we used comma to separate them in the same line.

* The Input Data is the size of the input WiFi data to the augmentation methods.

** The Pre-Aug perf & Post-Aug perf are the positioning accuracies of the system before and after the data augmentation. Due to the challenges in consistently reporting positioning performance in the reviewed research publications, m is used for regression systems while the hitting rates (%) are used for classification systems. For systems assessing the WiFi RSS data augmentation accuracy, dBm is used.

Paper	Testbed type	Testbed size † (m^2)	Collected data size	Augmentation method	Input data size*	Generated data size	Pre-Aug perf** (m)	Post-Aug perf** (m)
[66]	Simulation	25, 1500	75, N/A	WGAN	75, N/A	600, 9000	1.3, 2.11	0.86, 1.41
[86]	Building floor	N/A	360000	cGAN	360000	15000	1.25	0.99
[51]	Building floor	∼9000	1396 RPs	Multi-Output GPR	1396 RPs	1396 RPs	4.2	3.4
[70]	University building	N/A	23 per RP	SDAE	23 per RP	>25 per RP	9.42	8.37
[29]	Hallway, office room	46.45, 139.35	4000, 4000	Perturbation	4000, 4000	24000, 20000	0.994, 5.205	0.308, 1.349
[29]	Hallway, office room	46.45, 139.35	4000, 4000	Perturbation	4000, 4000	24000, 24000	1.413, 5.726	0.832, 2.856
[69]	Two building floors	N/A	9232, 4550	AE	6465, 3415	907, 843	1.1, 80% of the time, 0.6, 80% of the time	0.9, 80% of the time, 0.4, 80% of the time
[68]	Three building floors	46369	16157	WGAN	16157 (4-104 per location)	75 per location	2.278	2.023
[34]	Laboratory room	112	112 RPs	Interpolation	28 RPs	84 RPs	1.172	1.265
[33]	Building floor	4000	11600	Interpolation	11600	N/A	5.1	3.7
[41]	Building floor	3750	389 RPs	Propagation model	10 RPs	379 RPs	5.7	4.75
[12]	Classroom	36	48 RPs	Propagation model	32 RPs, 24 RPs	16 RPs, 24 RPs	N/A	1.8, 1.25
[35]	Building floor	∼1000	∼80 RPs	Interpolation	∼80 RPs	∼20 RPs	4.2743 dBm	3.9726 dBm
[26]	Building floor	696.54	832 RPs	Tensor Completion	20% of the collected	80% of the collected	0.9, 80% of the time	1.8, 80%of the time
[87]	Shopping mall	29400	55 RPs	Interpolation	55 RPs	275 RPs	6.38	2.86

† For data augmentation methods evaluated on different testbeds, we used comma to separate them in the same line.

* The Input Data is the size of the input WiFi data to the augmentation methods.

** The Pre-Aug perf & Post-Aug perf are the positioning accuracies of the system before and after the data augmentation. Due to the challenges in consistently reporting positioning performance in the reviewed research publications, m is used for regression systems while the hitting rates (%) are used for classification systems. For systems assessing the WiFi RSS data augmentation accuracy, dBm is used.

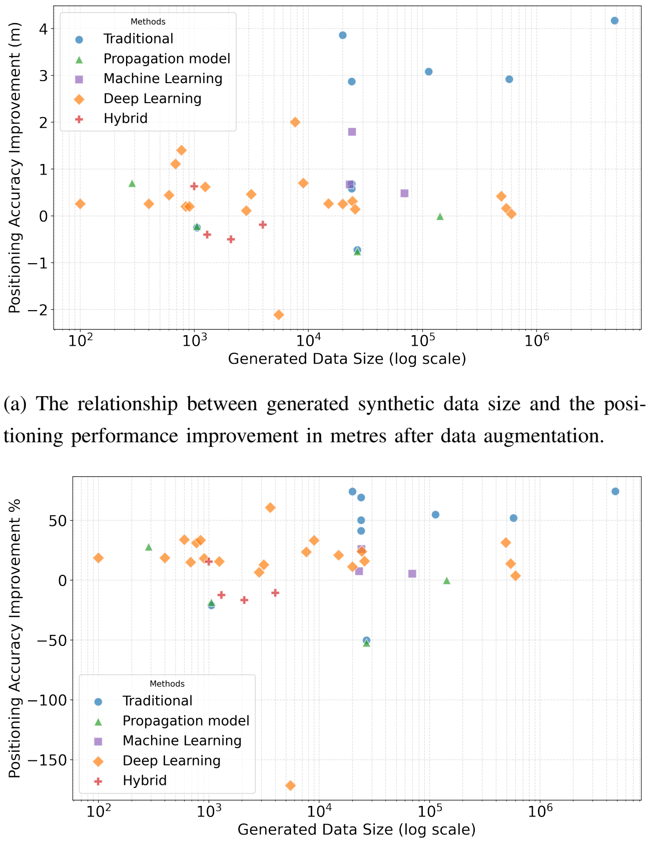
The improvements in indoor positioning achieved through data augmentation methods, focusing on the relationship between generated data sizes and the ratio of generated to input data are examined in Figures 12 and 13. Traditional methods achieve larger absolute improvements in metres with higher generated-to-input ratios, **though the growth in improvement percentage slows down after approximately sixfold augmentation, suggesting diminishing returns**. It is interesting to see that traditional methods brings the largest average improvement across all different approaches. Propagation model-based approaches initially exhibit performance degradation (even negative improvements) at low ratios but show slight accuracy recovery beyond fivefold augmentation. This is because some related studies focused on reducing human effort in WiFi fingerprint data collection and construction, thus utilising limited real data samples as input. This degradation arises because the hybrid methods used in [22] were evaluated against an entire dataset meticulously collected by human testers, whereas the approach in [78] utilised only up to 70% of the available data for augmentation training. Using only a small portion of the collected dataset for data augmentation typically leads to a decrease in positioning accuracy. However, traditional methods proposed in [32], [41], which utilised less than 30% of the collected data, still achieved post-augmentation performance improvements of 0.02 and 0.95 metres, respectively. In comparison, deep learning-based methods presented in [19], [64], [69] using up to 60% of the collected data also achieved promising post-augmentation improvements of 0.16 metres, 0.13 metres, and 0.2 metres, respectively, demonstrating a more robust and reliable performance.

Deep learning-based methods maintain steady improvement even with increasing ratios, peaking after threefold augmentation, beyond which the performance improvement stops to increase. This indicates that the performance improvement (not performance) from data augmentation using deep learning methods saturates beyond a threefold augmentation ratio. Therefore, a threefold ratio represents the most effective and efficient point for data augmentation in this context. In contrast, **machine learning methods suffer performance declines as augmentation ratios rise, potentially due to overfitting or noise from excessive synthetic data**.

V. CONCLUSION AND FUTURE WORK

This paper presents a comprehensive and detailed analysis of data augmentation techniques in WiFi fingerprinting-based indoor positioning, reviewing over 70 studies. It highlights the significant role of these techniques in reducing human labour during fingerprint dataset construction and enhancing indoor positioning performance with synthetic data. This review also proposed a novel taxonomy to categorise the most popular and trending data augmentation methods utilised in WiFi fingerprinting-based indoor positioning, including traditional methods, propagation models, machine learning methods, deep learning methods, hybrid methods and other methods. It was observed that GANs and traditional methods, such as interpolation, are frequently utilised as data augmentation techniques in the literature. Furthermore, the conversion of WiFi fingerprints into image format is a widely adopted practice.

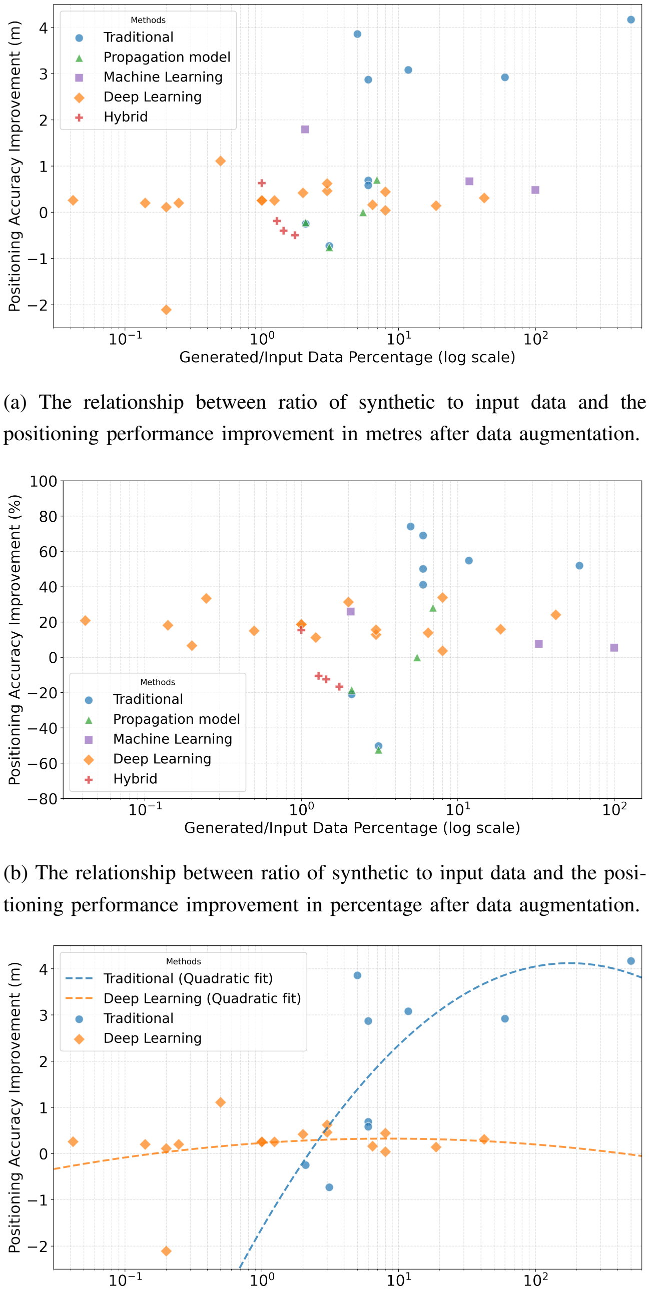
Additionally, we conducted a detailed analysis involving a comprehensive performance evaluation of WiFi fingerprinting data augmentation methods. We compared the indoor positioning performances of WiFi fingerprint datasets before and after applying different data augmentation techniques. The relationship between positioning performance and several factors, such as the sample size of input data for data augmentation, the sample size of generated synthetic WiFi data, the total number of WiFi fingerprint samples, and the generation ratio of synthetic to input data, is examined.



(a) The relationship between generated synthetic data size and the positioning performance improvement in metres after data augmentation.

(b) The relationship between generated synthetic data size and the positioning performance improvement in percentage after data augmentation.

FIGURE 12: The correlation between post-augmentation performance improvement and generated data sizes in WiFi fingerprinting indoor positioning. The post-augmentation performance is the positioning accuracy with augmented WiFi fingerprint datasets. Performances are analysed against different data augmentation methods, all plotted on logarithmic scales.



(a) The relationship between ratio of synthetic to input data and the positioning performance improvement in metres after data augmentation.

(b) The relationship between ratio of synthetic to input data and the positioning performance improvement in percentage after data augmentation.

(c) The quadratic fit of the relationship between ratio of synthetic to input data and the positioning performance improvement in metres after data augmentation for deep learning and traditional methods.

FIGURE 13: The correlation between post-augmentation performance improvement and generated-to-input data ratios in WiFi fingerprinting. Performances are analysed against different data augmentation methods, all plotted on logarithmic scales. It is observed that traditional methods and deep learning-based models are the most effective data augmentation approaches for WiFi fingerprinting, while the optimal generated-to-input ratio remains around threefold for deep learning models.

We further explored the influence of data augmentation on positioning accuracy improvements. We also investigated the correlation between WiFi data samples and indoor positioning performance across various techniques, including deep learning, traditional, propagation model, and hybrid approaches. From the analysis, we revealed that traditional data augmentation methods can achieve promising positioning accuracy with significant improvement compared to original datasets. While deep learning techniques have shown consistent advancements in enhancing indoor positioning accuracy, indiscriminately increasing data volume, particularly surpassing the threefold generated-to-input ratio, proves less effective. In terms of hardware and computational resource requirements, traditional augmentation methods offer promising performance improvements while operating efficiently on standard CPUs. In contrast, deep learning-based techniques demand significantly greater computational resources, typically requiring high-performance GPUs to achieve comparable gains [53]–[56].

While this review highlights significant advancements in WiFi fingerprinting data augmentation, several promising directions remain to be explored. First, developing hybrid models that effectively combine the computational efficiency of traditional methods with the representational power of deep learning could optimise the trade-off between augmentation quality and resource demands. Second, enhancing adaptability to dynamic indoor environments through real-time data augmentation frameworks remains critical, particularly for handling temporal signal variations and layout changes. Current WiFi fingerprinting data augmentation methods rely on pre-collected signal measurements to generate synthetic data, which limits their ability to adapt effectively in real time. Additionally, optimising deep learning architectures to maintain efficiency beyond threefold data generation ratios and integrating emerging WiFi signal measures (e.g., CSI and RTT) could further improve robustness. Finally, establishing standardised evaluation metrics and benchmark datasets would facilitate practical deployment, bridging the gap between theoretical innovation and real-world applicability. For example, researchers could either make their source code publicly available or employ well-known public WiFi indoor positioning datasets such as UJIIndoorLoc [6]. They can then evaluate system performance using widely adopted metrics such as MAE and RMSE [1]–[4]. Future research could also explore integrating WaveFlex biosensors—known for their flexibility, multi-parameter sensing, and wearable compatibility—with WiFi-based systems to enrich contextual awareness. Evaluating the feasibility of real-time augmentation represents another valuable direction for future research.

Domain generalization and adaptation represent two critical frontiers in ensuring robust performance of intelligent WiFi indoor positioning systems within dynamic indoor environments, where factors such as lighting, layout, and occupancy change over time. While these areas are foundational to addressing distributional shifts, ”environmental changes” represent a distinct technological challenge—separate from domain generalization/adaptation—that focuses on real-time system updates to accommodate continuous, non-stationary conditions. These are particularly valuable directions for future work, as they offer complementary yet distinct pathways to achieving long-term system resilience. By prioritising these methodologies, researchers and practitioners can unlock scalable solutions for more robust WiFi indoor positioning systems.

REFERENCES

[1] F. Liu, J. Liu, Y. Yin, W. Wang, D. Hu, P. Chen, and Q. Niu, "Survey on wifi-based indoor positioning techniques," *IET communications*, vol. 14, no. 9, pp. 1372–1383, 2020.

[2] S. Shang and L. Wang, "Overview of wifi fingerprinting-based indoor positioning," *Iet Communications*, vol. 16, no. 7, pp. 725–733, 2022.

[3] S. Xia, Y. Liu, G. Yuan, M. Zhu, and Z. Wang, "Indoor fingerprint positioning based on wi-fi: An overview," *ISPRS international journal of geo-information*, vol. 6, no. 5, p. 135, 2017.

[4] X. Feng, K. A. Nguyen, and Z. Luo, "A survey of deep learning approaches for wifi-based indoor positioning," *Journal of Information and Telecommunication*, vol. 6, no. 2, pp. 163–216, 2022.

[5] ——, "A review of open access wifi fingerprinting datasets for indoor positioning," *IEEE Access*, 2024.

[6] J. Torres-Sospedra, R. Montoliu, A. Martínez-Usó, 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 *2014 international conference on indoor positioning and indoor navigation (IPIN)*. IEEE, 2014, pp. 261–270.

[7] X. Feng, K. A. Nguyen, and Z. Luo, "A wifi rss-rtt indoor positioning system using dynamic model switching algorithm," *IEEE Journal of Indoor and Seamless Positioning and Navigation*, 2024.

[8] ——, "A dynamic model switching algorithm for wifi fingerprinting indoor positioning," in *2023 13th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2023, pp. 1–6.

[9] ——, "A multi-scale feature selection framework for wifi access points line-of-sight identification," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2023, pp. 1–6.

[10] V. Moghtadaiee, S. A. Ghorashi, and M. Ghavami, "New reconstructed database for cost reduction in indoor fingerprinting localization," *IEEE Access*, vol. 7, pp. 104 462–104 477, 2019.

[11] L. Chang, J. Xiong, Y. Wang, X. Chen, J. Hu, and D. Fang, "iupdater: Low cost rss fingerprints updating for device-free localization," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2017, pp. 900–910.

[12] Y.-H. Wu, Y.-L. Chen, and S.-T. Sheu, "Indoor location estimation using virtual fingerprint construction and zone-based remedy algorithm," in *2016 International Conference On Communication Problem-Solving (ICCP)*. IEEE, 2016, pp. 1–3.

[13] S.-H. Jung and D. Han, "Automated construction and maintenance of wi-fi radio maps for crowdsourcing-based indoor positioning systems," *IEEE Access*, vol. 6, pp. 1764–1777, 2017.

[14] K. Maharana, S. Mondal, and B. Nemade, "A review: Data preprocessing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022.

[15] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[16] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, p. 100258, 2022.

[17] T. Kumar, R. Brennan, A. Mileo, and M. Bendechache, "Image data augmentation approaches: A comprehensive survey and future directions," *IEEE Access*, 2024.

[18] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the prisma statement," *Bmj*, vol. 339, 2009.

[19] T. Lan, X. Wang, Z. Chen, J. Zhu, and S. Zhang, "Fingerprint augment based on super-resolution for wifi fingerprint based indoor localization," *IEEE Sensors Journal*, vol. 22, no. 12, pp. 12 152–12 162, 2022.

[20] E. M. Laó Amores, "Data augmentation models for improved indoor positioning accuracy using rss fingerprinting," 2024.

[21] 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 Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 3, pp. 468–480, 2019.

[22] 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 of Things Journal*, vol. 7, no. 8, pp. 6946–6954, 2020.

[23] J.-H. Seong and D.-H. Seo, "Selective unsupervised learning-based wi-fi fingerprint system using autoencoder and gan," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1898–1909, 2019.

[24] Y. Du, D. Yang, and C. Xiu, "A novel method for constructing a wifi positioning system with efficient manpower," *Sensors*, vol. 15, no. 4, pp. 8358–8381, 2015.

[25] Z. Gu, Z. Chen, Y. Zhang, Y. Zhu, M. Lu, and A. Chen, "Reducing fingerprint collection for indoor localization," *Computer Communications*, vol. 83, pp. 56–63, 2016.

[26] L. Ma, W. Zhao, Y. Xu, and C. Li, "Radio map efficient building method using tensor completion for wlan indoor positioning system," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.

[27] R. S. Sinha and S.-H. Hwang, "Improved rssi-based data augmentation technique for fingerprint indoor localisation," *Electronics*, vol. 9, no. 5, p. 851, 2020.

[28] L. Xiao, A. Behboodi, and R. Mathar, "A deep learning approach to fingerprinting indoor localization solutions," in *2017 27th International Telecommunication Networks and Applications Conference (ITNAC)*. IEEE, 2017, pp. 1–7.

[29] O. G. Serbetci, J.-H. Lee, D. Burghal, and A. F. Molisch, "Simple and effective augmentation methods for csi based indoor localization," in *GLOBECOM 2023-2023 IEEE Global Communications Conference*. IEEE, 2023, pp. 3947–3952.

[30] R. S. Sinha, S.-M. Lee, M. Rim, and S.-H. Hwang, "Data augmentation schemes for deep learning in an indoor positioning application," *Electronics*, vol. 8, no. 5, p. 554, 2019.

[31] C. Xiang, Z. Zhang, S. Zhang, S. Xu, S. Cao, and V. Lau, "Robust sub-meter level indoor localization-a logistic regression approach," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[32] Y. Yang, P. Dai, H. Huang, M. Wang, and Y. Kuang, "A semi-simulated rss fingerprint construction for indoor wi-fi positioning," *Electronics*, vol. 9, no. 10, p. 1568, 2020.

[33] J. Yang, "Indoor localization system using dual-frequency bands and interpolation algorithm," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11 183–11 194, 2020.

[34] H. Zhao, B. Huang, and B. Jia, "Applying kriging interpolation for wifi fingerprinting based indoor positioning systems," in *2016 IEEE Wireless Communications and Networking Conference*. IEEE, 2016, pp. 1–6.

[35] J. Racko, J. Machaj, and P. Brida, "Wi-fi fingerprint radio map creation by using interpolation," *Procedia engineering*, vol. 192, pp. 753–758, 2017.

[36] A. H. Ismail, Y. Mizushiri, R. Tasaki, H. Kitagawa, T. Miyoshi, and K. Terashima, "A novel automated construction method of signal fingerprint database for mobile robot wireless positioning system," *International Journal of Automation Technology*, vol. 11, no. 3, pp. 459–471, 2017.

[37] Y. F. Yong, C. K. Tan, and I. K. Tan, "Smote for wi-fi fingerprint construction in indoor positioning systems," in *2021 IEEE International Performance, Computing, and Communications Conference (IPCCC)*. IEEE, 2021, pp. 1–6.

[38] C. He, S. Guo, Y. Wu, and Y. Yang, "A novel radio map construction method to reduce collection effort for indoor localization," *Measurement*, vol. 94, pp. 423–431, 2016.

[39] J. Bi, Y. Wang, Z. Li, S. Xu, J. Zhou, M. Sun, and M. Si, "Fast radio map construction by using adaptive path loss model interpolation in large-scale building," *Sensors*, vol. 19, no. 3, p. 712, 2019.

[40] F. Lemic, V. Handziski, G. Caso, L. De Nardis, and A. Wolisz, "Enriched training database for improving the wifi rssi-based indoor fingerprinting performance," in *2016 13th IEEE Annual Consumer*

Communications & Networking Conference (CCNC). IEEE, 2016, pp. 875–881.

[41] T. Guan, L. Fang, W. Dong, D. Koutsonikolas, G. Challen, and C. Qiao, "Robust, cost-effective and scalable localization in large indoor areas," *Computer Networks*, vol. 120, pp. 43–55, 2017.

[42] E. Schulz, M. Speekenbrink, and A. Krause, "A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions," *Journal of mathematical psychology*, vol. 85, pp. 1–16, 2018.

[43] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian process regression for materials and molecules," *Chemical Reviews*, vol. 121, no. 16, pp. 10 073–10 141, 2021.

[44] W. Sun, M. Xue, H. Yu, H. Tang, and A. Lin, "Augmentation of fingerprints for indoor wifi localization based on gaussian process regression," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 896–10 905, 2018.

[45] Y. Dong, T. Arslan, Y. Yang, and Y. Ma, "A wifi fingerprint augmentation method for 3-d crowdsourced indoor positioning systems," in *2022 IEEE 12th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2022, pp. 1–8.

[46] Y. Dong, G. He, T. Arslan, Y. Yang, and Y. Ma, "Crowdsourced indoor positioning with scalable wifi augmentation," *Sensors*, vol. 23, no. 8, p. 4095, 2023.

[47] Y. Tao and L. Zhao, "Aips: An accurate indoor positioning system with fingerprint map adaptation," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 3062–3073, 2021.

[48] H. Zou, M. Jin, H. Jiang, L. Xie, and C. J. Spanos, "Winips: Wifi-based non-intrusive indoor positioning system with online radio map construction and adaptation," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 8118–8130, 2017.

[49] X. Wang, T. Lan, S. Zhang, and J. Zhu, "Signal distribution oriented mean functions in gpr based fingerprint augment," in *2019 11th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, 2019, pp. 1–6.

[50] Z. Tang, S. Li, K. S. Kim, and J. Smith, "Multi-output gaussian process-based data augmentation for multi-building and multi-floor indoor localization," in *2022 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2022, pp. 361–366.

[51] Z. Tang, S. Li, K. S. Kim, and J. S. Smith, "Multi-dimensional wi-fi received signal strength indicator data augmentation based on multi-output gaussian process for large-scale indoor localization," *Sensors*, vol. 24, no. 3, p. 1026, 2024.

[52] G. M. Mendoza-Silva, A. C. Costa, J. Torres-Sospedra, M. Painho, and J. Huerta, "Environment-aware regression for indoor localization based on wifi fingerprinting," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 4978–4988, 2021.

[53] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[54] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[55] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[56] C. Bowles, L. Chen, R. Guerrero, P. Bentley, R. Gunn, A. Hammers, D. A. Dickie, M. V. Hernández, J. Wardlaw, and D. Rueckert, "Gan augmentation: Augmenting training data using generative adversarial networks," *arXiv preprint arXiv:1810.10863*, 2018.

[57] S. A. Junoh and J.-Y. Pyun, "Enhancing indoor localization with semi-crowdsourced fingerprinting and gan-based data augmentation," *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 11 945–11 959, 2023.

[58] M. Mallik and C. Chowdhury, "Characteristic analysis of fingerprint datasets from a pragmatic view of indoor localization using machine learning approaches," *The Journal of Supercomputing*, vol. 79, no. 16, pp. 18 507–18 546, 2023.

[59] D. Quezada-Gaibor, J. Torres-Sospedra, J. Nurmi, Y. Koucheryavy, and J. Huerta, "Surimi: Supervised radio map augmentation with deep learning and a generative adversarial network for fingerprint-based indoor positioning," in *2022 IEEE 12th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2022, pp. 1–8.

[60] L. Chen, S. Zhang, H. Tan, and B. Lv, "Progressive rss data augmenter with conditional adversarial networks," *IEEE Access*, vol. 8, pp. 26 975–26 983, 2020.

[61] J. Yoo, "Wi-fi fingerprint indoor localization by semi-supervised generative adversarial network," *Sensors*, vol. 24, no. 17, p. 5698, 2024.

[62] K. M. Chen and R. Y. Chang, "Semi-supervised learning with gans for device-free fingerprinting indoor localization," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–6.

[63] H. Yang and L. Chen, "Improving indoor localization through data augmentation of visualized multidimensional fingerprints via enhanced generative networks," *IEEE Sensors Journal*, 2024.

[64] X.-Y. Liu and X. Wang, "Real-time indoor localization for smartphones using tensor-generative adversarial nets," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 8, pp. 3433–3443, 2020.

[65] C. Zhu, L. Xu, X.-Y. Liu, and F. Qian, "Tensor-generative adversarial network with two-dimensional sparse coding: Application to real-time indoor localization," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.

[66] J. Yang, J. Tian, Y. Qi, W. Cheng, Y. Liu, G. Han, S. Wang, Y. Li, C. Cao, and S. Qin, "Research on 3d localization of indoor uav based on wasserstein gan and pseudo fingerprint map," *Drones*, vol. 8, no. 12, p. 740, 2024.

[67] C. Li and Y. Mao, "Improved indoor localization algorithm combining k-means clustering algorithm and wasserstein generative adversarial network algorithm," in *2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. IEEE, 2023, pp. 1–5.

[68] S. Yean, W. Goh, B.-S. Lee, and H. L. Oh, "Extendgan+: Transferable data augmentation framework using wgan-gp for data-driven indoor localisation model," *Sensors*, vol. 23, no. 9, p. 4402, 2023.

[69] H. Park, C. Laoudias, S. Horsmanheimo, S. Kim *et al.*, "Dropout autoencoder fingerprint augmentation for enhanced wi-fi ftm-rss indoor localization," *IEEE Communications Letters*, vol. 27, no. 7, pp. 1759–1763, 2023.

[70] C. Zhuang and D. Zhang, "A robust wifi localization algorithm using data augmentation and stacked denoising autoencoder," in *2023 35th Chinese Control and Decision Conference (CCDC)*. IEEE, 2023, pp. 1445–1450.

[71] D. J. Suroso, P. Cherntanomwong, and P. Sooraksa, "Deep generative model-based rssi synthesis for indoor localization," in *2022 19th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. IEEE, 2022, pp. 1–5.

[72] N. Yoon, W. Jung, and H.-n. Kim, "Deeprssi: Generative model for fingerprint-based localization," *IEEE Access*, 2024.

[73] X. Wang, X. Wang, S. Mao, J. Zhang, S. C. Periaswamy, and J. Patton, "Deepmap: Deep gaussian process for indoor radio map construction and location estimation," in *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2018, pp. 1–7.

[74] ——, "Indoor radio map construction and localization with deep gaussian processes," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11 238–11 249, 2020.

[75] M. Sugasaki and M. Shimosaka, "Robustifying wi-fi localization by between-location data augmentation," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 5407–5416, 2021.

[76] B. Jia, W. Qiao, Z. Zong, S. Liu, M. Hijji, J. Del Ser, and K. Muhammad, "A fingerprint-based localization algorithm based on lstm and data expansion method for sparse samples," *Future Generation Computer Systems*, vol. 137, pp. 380–393, 2022.

[77] Y. Chan, P.-Y. Lin, Y.-Y. Tseng, J.-J. Chen, and Y.-C. Tseng, "Learning-based wifi fingerprint inpainting via generative adversarial networks," in *2024 33rd International Conference on Computer Communications and Networks (ICCCN)*. IEEE, 2024, pp. 1–7.

[78] J. Zhao, X. Gao, X. Wang, C. Li, M. Song, and Q. Sun, "An efficient radio map updating algorithm based on k-means and gaussian process regression," *The Journal of Navigation*, vol. 71, no. 5, pp. 1055–1068, 2018.

[79] 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. 98 337–98 347, 2021.

[80] A. Hilal, I. Arai, and S. El-Tawab, "Dataloc+: A data augmentation technique for machine learning in room-level indoor localization," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2021, pp. 1–7.

[81] Z. Zhang, M. Lee, and S. Choi, "Deep-learning-based wi-fi indoor positioning system using continuous csi of trajectories," *Sensors*, vol. 21, no. 17, p. 5776, 2021.

[82] W.-Y. Kim, S.-H. Tae, and D.-H. Seo, "Access-point centered window-based radio-map generation network," *Sensors*, vol. 21, no. 18, p. 6107, 2021.

[83] M. Nabati, H. Navidan, R. Shahbazian, S. A. Ghorashi, and D. Windridge, "Using synthetic data to enhance the accuracy of fingerprint-based localization: A deep learning approach," *IEEE Sensors Letters*, vol. 4, no. 4, pp. 1–4, 2020.

[84] J. Wang, Z. Zhao, J. Cui, Y. Wang, Y. Shi, and B. Wu, "Low-cost wi-fi fingerprinting indoor localization via generative deep learning," in *International Conference on Wireless Algorithms, Systems, and Applications*. Springer, 2021, pp. 53–64.

[85] C. Serbouh, W. Njima, and I. Ahriz, "Generative adversarial networks based data recovery for indoor localization," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2024, pp. 1–6.

[86] W. Wei, J. Yan, L. Wan, C. Wang, G. Zhang, and X. Wu, "Enriching indoor localization fingerprint using a single ac-gan," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2021, pp. 1–6.

[87] Y. Wang, R. Guo, W. Wang, X. Li, S. Tang, W. Zhang, L. Wang, L. Chen, Y. Li, and W. Xiu, "Near relation-based indoor positioning method under sparse wi-fi fingerprints," *ISPRS International Journal of Geo-Information*, vol. 9, no. 12, p. 714, 2020.