

# LoFi: Vision-Aided Label Generator for Wi-Fi Localization and Tracking

Zijian Zhao, Tingwei Chen, Fanyi Meng, Zhijie Cai, Hang Li, Xiaoyang Li, Guangxu Zhu

**Abstract**—Wi-Fi localization and tracking has shown immense potential due to its privacy-friendliness, wide coverage, permeability, independence from lighting conditions, and low cost. Current methods can be broadly categorized as model-based and data-driven approaches, where data-driven methods show better performance and have less requirement for specialized devices, but struggle with limited datasets for training. Due to limitations in current data collection methods, most datasets only provide coarse-grained ground truth (GT) or limited amount of label points, which greatly hinders the development of data-driven methods. Even though lidar can provide accurate GT, their high cost makes them inaccessible to many users. To address these challenges, we propose LoFi, a vision-aided label generator for Wi-Fi localization and tracking, which can generate ground truth position coordinates solely based on 2D images. The easy and quick data collection method also helps data-driven based methods deploy in practice, since Wi-Fi is a low-generalization modality and when using relevant methods, it always requires fine-tuning the model using newly collected data. Based on our method, we also collect a Wi-Fi tracking and localization dataset using ESP32-S3 and a webcam. To facilitate future research, we will make our code and dataset publicly available upon publication.

**Index Terms**—Wi-Fi Sensing, Wi-Fi Tracking, Wi-Fi Localization, Multi-modal, Channel Statement Information

## I. INTRODUCTION

Indoor localization [1] and tracking [2] tasks have long been a challenge. The most common method, GPS, often fails to work reliably indoors, as the signals are easily blocked, distorted, and weakened by building materials and structures, resulting in poor satellite visibility and geometry for accurate positioning. Vision [3], lidar [4], and bluetooth [5] are some common indoor localization methods, but they have their own shortcomings, such as low privacy and high requirements for line of sight in the case of vision, high cost of lidar, and low accuracy of bluetooth.

In contrast, Wi-Fi sensing can efficiently address these problems. When people move within a Wi-Fi-enabled indoor environment, their presence and movements can alter the propagation of Wi-Fi signals, causing changes in the Received

Signal Strength Information (RSSI) and Channel State Information (CSI) at nearby Wi-Fi access points or client devices. By analyzing these changes in Wi-Fi signal characteristics, it is possible to infer the location, movement, and even activities of the person within the indoor space. Wi-Fi sensing has been viewed as a very suitable technology for indoor localization and tracking, as it has shown promising performance in many works [6], [7]. Additionally, since Wi-Fi devices are ubiquitous in many indoor scenarios, these methods can be realized without the need for any extra devices.

Wi-Fi localization and tracking methods can be broadly classified into two categories: model-based methods and data-driven methods. Model-based methods typically require multiple receivers or antennas as they rely on leveraging spatial diversity and multi-path analysis [8]. However, this requirement may not always be practical, especially in home scenarios where a single Wi-Fi router is commonly used. In contrast, data-driven methods do not have such limitations, but they often require a large amount of training data. Currently, most Wi-Fi localization and tracking datasets rely on ground truth data obtained through manual tagging or lidar-based methods. Manual tagging can only provide coarse-grained position information or a limited number of data points, which can adversely impact the model’s generalization capacity. Additionally, manual tagging is time-consuming and labor-intensive. While lidar-based methods can provide accurate results, their high cost makes them inaccessible to many users. As a result, most current data-driven methods are limited to classification tasks that can only identify a coarse-grained region where people are located or the general trajectory shape of people’s movements, rather than precise tracking capabilities [6], [9].

To solve this problem, we propose LoFi, an easy, quick, and low-cost Wi-Fi tracking and localization dataset collection method, by using vision to extract the ground truth people location. It uses object detection methods to capture people’s coordinates in pixel space and transfer them to physical space. The main contributions of this work are:

(1) We propose LoFi, a vision-aided label generator for Location and tracking tasks in Wi-Fi sensing. Our method can help collect datasets quickly and at low cost without any requirement for a camera, Wi-Fi receiver, or people movement trajectory.

(2) We present a Wi-Fi tracking and localization dataset using our LoFi method. We provide information including CSI, RSSI, timestamp, people’s coordinates, and people ID. Our dataset can be used for many tasks including tracking, localization, people identification, and cross-domain Wi-Fi

\* Corresponding Author: Guangxu Zhu

Zijian Zhao is with Shenzhen Research Institute of Big Data, Shenzhen 518115, China, and also with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: zhaozj28@mail2.sysu.edu.cn)

Tingwei Chen, Fanyi Meng, Zhijie Cai, Hang Li, Xiaoyang Li, and Guangxu Zhu are with the Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong (Shenzhen), Shenzhen 518115, China (e-mail: tingweichen@link.cuhk.edu.cn; fanyimeng@link.cuhk.edu.cn; zhijiecai@link.cuhk.edu.cn; hangdavidli@sribd.cn; l Xiaoyang@sribd.cn; gxzhu@sribd.cn)

TABLE I  
DATASET DESCRIPTION: ‘–’ REPRESENTS INFORMATION NOT MENTIONED IN THE PAPER AND ‘\*’ REPRESENTS THE DATASET IS PUBLICLY AVAILABLE.

Dataset	Transmitter		Receiver		Scale	Modality	Device	Label Generation Method
	Site	Antennas	site	antennas				
[6], [10], [12]	1	1	1	3	56,040-300,000 frames	CSI	Intel WiFi Link 5300 NIC	Assign Previously
[13]*, [14]*, [15]–[17]	1	3	1	3	6,000-504,000 frames	CSI	Intel WiFi Link 5300 NIC	
[18]	1	–	1	3	–	CSI & RSSI	Intel WiFi Link 5300 NIC	
[19] *	172	–	4	–	25,364 frames	RSSI	mobile phone	
[20] *	620	–	1	–	3,696 frames	RSSI	mobile phone	
[7] *	1	–	2	3	216 instances	CSI	Intel WiFi Link 5300 NIC	
[8] *	1	3	3	3	60 instances	CSI	Intel WiFi Link 5300 NIC & Camera	
[21] <sup>1*</sup>	1	1	1	16	17,000 frames	CSI	USRP & SDR-equipped Vacuum-cleaner Robot	
Ours *	1	1	1	1	210,000 frames	CSI & RSSI	ESP32-S3 & Webcam	
							Vision-Aided Method	

sensing.

## II. EXISTING DATASETS

As shown in Table I, we summarize the key parameters of some popular datasets in Wi-Fi localization and tracking. In traditional Wi-Fi localization datasets, most work assigns some predetermined positions and has people stand at them to collect data [10]. This approach can only obtain data at specific points, significantly compromising the generalization capability of data-driven methods. Additionally, some other localization datasets divide the room into regions and provide only the region in which people are located, which can only be used for coarse-grained localization [6].

For Wi-Fi tracking datasets, most assign trajectories in advance, and these datasets can only be used to train models to identify the shape of the trajectories [9]. Some of them mark specific points on the ground and consider people to move uniformly between consecutive points. In this case, people’s positions can be calculated by interpolation methods using the timestamps [11]. However, the unreasonable assumption of uniform motion makes the ground truth data not accurate enough. Even though some work uses cameras to achieve accurate interpolation, the marked points still limit the people’s motion [8], which also harms the model generalization.

In contrast, our LoFi method can capture people’s motion arbitrarily by using cameras to generate their coordinates, without the limitations of predetermined positions or trajectories.

## III. DATASET COLLECTION METHOD

### A. Collection Process

The process of our LoFi dataset collection method is illustrated in Fig. 1. We first collect Wi-Fi CSI and vision image at the same time. Then we use the image to extract people’s coordinates and align it with the CSI to generate the label for Wi-Fi localization and tracking.

For the vision modality, we can first use object detection models like YOLO [22] and RCNN [23] to detect people’s position in the pixel space. We take the middle point of the bottom line of the anchor box with the highest confidence level as the people’s coordinates. Then, to transfer it to the physical space, we can use the “getPerspectiveTransform” function in

the OpenCV library [24]. It requires the coordinates of four non-collinear anchor points in the floor in pixel space and physical space respectively to calculate the transformation matrix by solving  $T$  in:

$$[X_1, X_2, X_3, X_4] = T \cdot [X'_1, X'_2, X'_3, X'_4], \quad (1)$$

where  $X_i$  and  $X'_i$  are the homogeneous coordinates of the anchor points in physical space and pixel space respectively, and  $T$  is the transformation matrix. Then, after getting the coordinate of the people’s position  $X'$  in pixel space, we can transfer it to the physical space by calculating  $X = T \cdot X'$ .

To get the coordinates of the anchor points, a naive method is to mark them by hand, use a ruler to measure the distance in the physical world, and use apps to get their coordinates in the pixel world. However, it’s also a little time-consuming and labor-intensive. As most current mobile phones have multiple cameras, we can directly use the stereo vision distance measurement method [25] to get the distance in the physical world as Eq. 2 and use mobile phone apps to get their coordinates in the pixel world. Even for mobile phones with only one camera, we can still realize the stereo vision distance measurement with the help of GPS or IMU sensors.

$$[X, Y, Z] = \left[ \frac{(x_l + x_r)B}{x_l - x_r}, \frac{(y_l + y_r)B}{x_l - x_r}, \frac{fB}{x_l - x_r} \right], \quad (2)$$

where  $X, Y, Z$  represent the 3D coordinates of the point,  $B$  is the baseline distance between the two cameras,  $f$  is the focal length of the cameras, and  $(x_l, y_l)$  and  $(x_r, y_r)$  are the coordinates of the point on the left and right imaging planes respectively.

### B. Dataset Description

According to our LoFi method, we collect a Wi-Fi tracking dataset in a laboratory as shown in Fig. 2 using a webcam and an ESP32-S3 with a single antenna, shown as Fig. 3. The ESP32-S3 receives the CSI and RSSI embedded in Ping Replies from a home Wi-Fi router, with a setting of 100Hz sampling rate and 2.4 GHz carrier frequency. The webcam is used to capture images at the same time as the ESP32-S3 with a sampling rate of 26Hz. During data collection, we recruited 7 volunteers and asked them to move arbitrarily (e.g., walk forward, walk backward, and stand) in a given region ( $1.8m \times 4.8m$ ) for 5 minutes, respectively. For privacy considerations, we do not provide the raw images. Instead, we have already processed the CSI and image data using our LoFi method and

<sup>1</sup>IEEE CTW 2019 indoor positioning competition, <https://ctw2019.ieee-ctw.org>

correlated the people's coordinates to each CSI frame. Unlike some previous datasets, we do not split the CSI sequence. As a result, users can split the data using any sliding window. As each CSI frame has its corresponding people's coordinates, the effective data amount of our dataset can reach approximately 210 thousand.

To illustrate the precision of our label generator method, we tested the label error against ground truth, where we had two volunteers stand at 10 predefined points. We test the error by setting the camera in three different views, as shown in Fig. 4. The mean errors are 17.44cm, 12.13cm, and 11.82cm, respectively. This error is acceptable because we consider people as single points, even though people occupy a given size in space.

#### IV. BENCHMARK

##### A. Model Description

In this section, we provide two common types of benchmark data-driven methods in Wi-Fi sensing: the convolution-based method and the series-based method, as shown in Fig. 5.

For the convolution-based network, we first calculate the frequency domain correlation matrix as:

$$F = M^T \cdot M , \quad (3)$$

where  $M \in R^{t,s}$  is the amplitude spectrum of CSI,  $t$  is the time length,  $s$  is the number of subcarriers,  $F$  represents the correlation between different subcarriers. Then, the convolution network can extract features and predict the people's coordinates based on this information.

For the series-based network, we first standardize each subcarrier in the time dimension, following the approach in [26]:

$$\begin{aligned} \mu_j &= \frac{\sum_{i=1}^n M_{i,j}}{n} , \\ \sigma_j &= \sqrt{\frac{\sum_{i=1}^n (M_{i,j} - \mu_j)^2}{n}} , \\ \text{Standard}(M_{i,j}) &= \frac{M_{i,j} - \mu_j}{\sigma_j + \epsilon} , \end{aligned} \quad (4)$$

where  $M_{i,j}$  represents the  $j^{th}$  dimension at the  $i^{th}$  time position of the amplitude spectrum  $M$ ,  $\mu$  and  $\sigma$  represent the mean and standard deviation, respectively, and  $\epsilon$  is a small number to avoid division by zero. Then, the standardized  $M$  can be viewed as a time series and input to the series-based network. Finally, we use a self-attention layer and a linear layer to predict the people's coordinates based on the features extracted by the series-based network.

##### B. Experiment Result

We conduct a series of experiments on the benchmarks in our dataset. We used a 1-second CSI series, corresponding to 100 packets, as the input to the model and train the model to predict the people's coordinates, where we take the mean value of the coordinates in this 1-second interval as the ground truth. The prediction error is shown in Table II, where the mean prediction error of all networks is lower than 1 meter. We then illustrate the Cumulative Distribution Function (CDF) of the error in Fig. 6.

Furthermore, as many studies have done [6], [15], we split the movement space into different regions and use the above

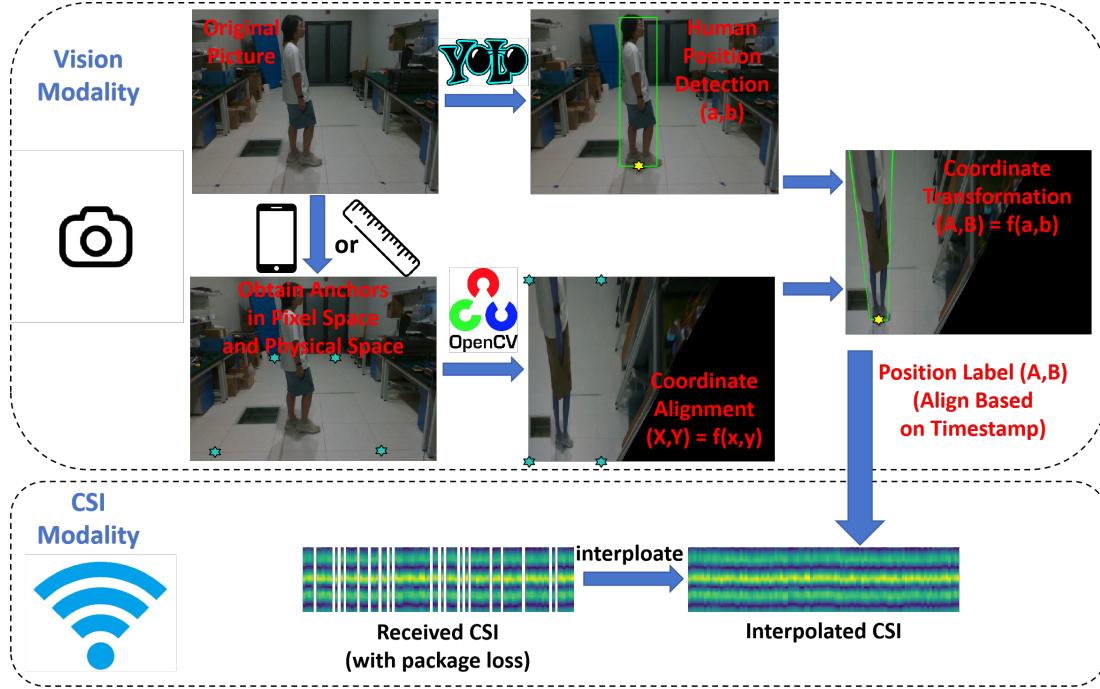


Fig. 1. Workflow: **Step 1:** Collect vision modality and CSI modality data simultaneously. **Step 2:** Detect people's positions in the image based on object detection methods. **Step 3:** Obtain the coordinates of four non-collinear anchors and construct the mapping relationship between pixel space and physical space. **Step 4:** Transfer the people's positions from pixel space to physical space. **Step 5:** Use interpolation methods to fill in the lost CSI packets. **Step 6:** Align the CSI modality and vision modality to match the people's positions to each CSI packet.

TABLE II  
EXPERIMENT RESULT: THE BOLD VALUE INDICATES THE BEST RESULT IN EACH ROW.

Metric	Methods		Convolution-based Methods		Series-based Methods		
	CNN	ResNet [27]	RNN [28]	GRU [29]	LSTM [30]	CSI-BERT [26]	
Error Mean	0.8745	<b>0.5830</b>	0.8705	0.9413	0.8643	0.6991	
Error Standard Deviation	0.3177	0.3475	0.2802	0.3217	<b>0.2475</b>	0.3063	
Classification Accuracy (6 classes)	31.99%	55.50%	53.29%	49.41%	53.50%	<b>60.07%</b>	
Classification Accuracy (4 classes)	42.03%	<b>62.54%</b>	61.92%	56.00%	62.15%	61.93%	
Classification Accuracy (2 classes)	62.84%	82.98%	<b>84.47%</b>	73.56%	73.98%	75.63%	

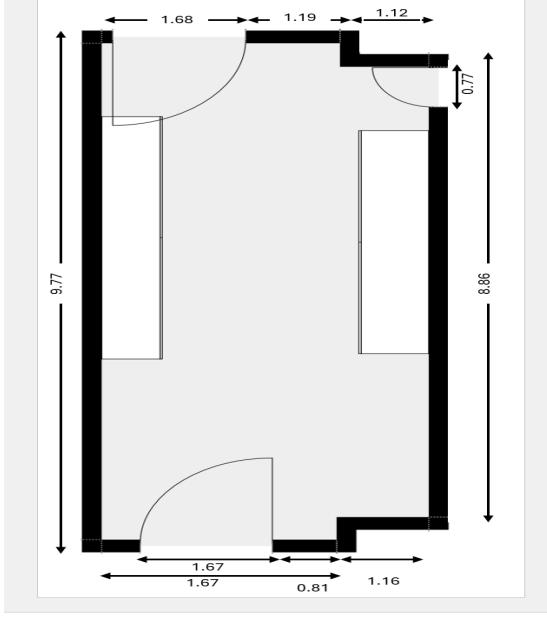


Fig. 2. Data Collection Environment: The unit in the picture is meters.



Fig. 3. Collection Device: To collect our dataset, we use a single webcam and an ESP32-S3 with a single antenna, and collect data simultaneously, linked to the same laptop.

networks to predict which region the people are located in. The classification result is shown in Table II, where we split the room into 6, 4, and 2 regions.

These results show the efficiency of our data collection method and the efficiency of data-driven based methods in single-antenna Wi-Fi localization, which, to the best of our knowledge, is the first work to illustrate this.



Fig. 4. Object Detection Result of 3 Different Views: The yellow stars represent the detected people position, and the green stars represent the anchor points.

## V. CONCLUSION

In this letter, we propose LoFi, a vision-aided automatic label generator method for Wi-Fi localization and tracking. Our method provides a simple and low-cost people coordinate extraction approach based on 2D images, which can be captured by any camera device. It offers a practical data collection technique for quickly fine-tuning relevant models. Additionally, it can collect more diverse data without any restriction on people movement, which will aid the development of data-driven methods in this field. Furthermore, we have published a dataset using our method, which can be utilized in research on Wi-Fi localization and tracking, people identification, and

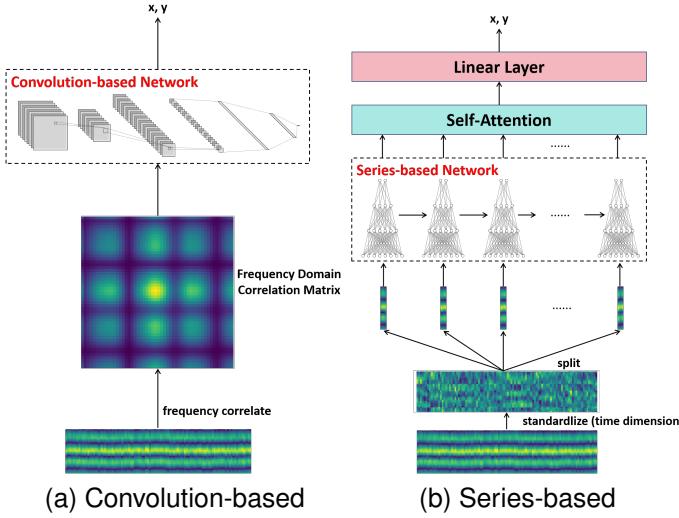


Fig. 5. Architecture of Two Types of Benchmark Networks

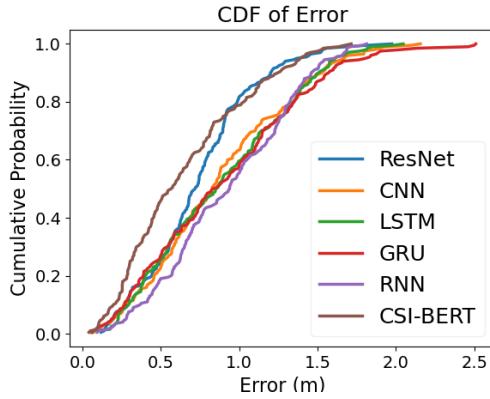


Fig. 6. CDF of Error for Benchmark Methods

cross-domain Wi-Fi sensing.

## REFERENCES

- [1] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2568–2599, 2019.
- [2] D. Dardari, P. Closas, and P. M. Djurić, "Indoor tracking: Theory, methods, and technologies," *IEEE transactions on vehicular technology*, vol. 64, no. 4, pp. 1263–1278, 2015.
- [3] A. Morar, A. Moldoveanu, I. Mocanu, F. Moldoveanu, I. E. Radoi, V. Asavei, A. Gradinaru, and A. Butean, "A comprehensive survey of indoor localization methods based on computer vision," *Sensors*, vol. 20, no. 9, p. 2641, 2020.
- [4] Y.-T. Wang, C.-C. Peng, A. A. Ravankar, and A. Ravankar, "A single lidar-based feature fusion indoor localization algorithm," *Sensors*, vol. 18, no. 4, p. 1294, 2018.
- [5] M. E. Rida, F. Liu, Y. Jadi, A. A. A. Algawhari, and A. Askourih, "Indoor location position based on bluetooth signal strength," in *2015 2nd International Conference on Information Science and Control Engineering*, pp. 769–773, IEEE, 2015.
- [6] W. Liu, Y. Chen, and H. Zhang, "Sifi: Siamese networks based csi fingerprint indoor localization with wifi," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 01–06, IEEE, 2024.
- [7] C. Li, L. Liu, Z. Cao, and M. Zhang, "Wivel: Fine-grained walking velocity estimation for wi-fi passive tracking," in *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 172–180, IEEE, 2022.
- [8] D. Wu, Y. Zeng, R. Gao, S. Li, Y. Li, R. C. Shah, H. Lu, and D. Zhang, "Witraj: Robust indoor motion tracking with wifi signals," *IEEE Transactions on Mobile Computing*, vol. 22, no. 5, pp. 3062–3078, 2021.
- [9] M. Werner, L. Schauer, and A. Scharf, "Reliable trajectory classification using wi-fi signal strength in indoor scenarios," in *2014 IEEE/ION Position, Location and Navigation Symposium-PLANS 2014*, pp. 663–670, IEEE, 2014.
- [10] X. Wang, L. Gao, and S. Mao, "Csi phase fingerprinting for indoor localization with a deep learning approach," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 1113–1123, 2016.
- [11] 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.
- [12] X. Wang, L. Gao, S. Mao, and S. Pandey, "Deepfi: Deep learning for indoor fingerprinting using channel state information," in *2015 IEEE wireless communications and networking conference (WCNC)*, pp. 1666–1671, IEEE, 2015.
- [13] X. Zhu, T. Qiu, W. Qu, X. Zhou, M. Atiquzzaman, and D. O. Wu, "Bls-location: A wireless fingerprint localization algorithm based on broad learning," *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, pp. 115–128, 2021.
- [14] X. Zhu, W. Qu, X. Zhou, L. Zhao, Z. Ning, and T. Qiu, "Intelligent fingerprint-based localization scheme using csi images for internet of things," *IEEE transactions on network science and engineering*, vol. 9, no. 4, pp. 2378–2391, 2022.
- [15] T. F. Sanam and H. Godrich, "An improved csi based device free indoor localization using machine learning based classification approach," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 2390–2394, IEEE, 2018.
- [16] T. F. Sanam and H. Godrich, "A multi-view discriminant learning approach for indoor localization using amplitude and phase features of csi," *IEEE Access*, vol. 8, pp. 59947–59959, 2020.
- [17] Y. Jing, J. Hao, and P. Li, "Learning spatiotemporal features of csi for indoor localization with dual-stream 3d convolutional neural networks," *IEEE Access*, vol. 7, pp. 147571–147585, 2019.
- [18] J. Yan, C. Ma, B. Kang, X. Wu, and H. Liu, "Extreme learning machine and adaboost-based localization using csi and rss," *IEEE Communications Letters*, vol. 25, no. 6, pp. 1906–1910, 2021.
- [19] P. Roy, C. Chowdhury, D. Ghosh, and S. Bandyopadhyay, "Juindoorloc: A ubiquitous framework for smartphone-based indoor localization subject to context and device heterogeneity," *Wireless Personal Communications*, vol. 106, pp. 739–762, 2019.
- [20] G. M. Mendoza-Silva, P. Richter, J. Torres-Sospedra, E. S. Lohan, and J. Huerta, "Long-term wifi fingerprinting dataset for research on robust indoor positioning," *Data*, vol. 3, no. 1, p. 3, 2018.
- [21] M. Arnold, J. Hoydis, and S. ten Brink, "Novel massive mimo channel sounding data applied to deep learning-based indoor positioning," in *SCC 2019; 12th International ITG Conference on Systems, Communications and Coding*, pp. 1–6, VDE, 2019.
- [22] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.
- [23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.
- [24] G. Bradski, A. Kaehler, et al., "Opencv," *Dr. Dobb's journal of software tools*, vol. 3, no. 2, 2000.
- [25] E. W. Weisstein, "Affine transformation," <https://mathworld.wolfram.com/>, 2004.
- [26] Z. Zhao, T. Chen, F. Meng, H. Li, X. Li, and G. Zhu, "Finding the missing data: A bert-inspired approach against package loss in wireless sensing," *arXiv preprint arXiv:2403.12400*, 2024.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- [28] L. R. Medsker and L. Jain, "Recurrent neural networks," *Design and Applications*, vol. 5, no. 64-67, p. 2, 2001.
- [29] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.