Double-Layer Soft Data Fusion for Indoor Robot WiFi-Visual Localization

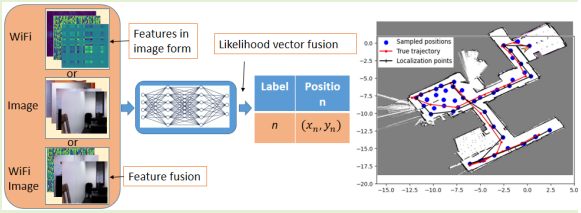
Yuehua Ding, Jean-François Dollinger, Vincent Vauchey, Mourad Zghal

Double-Layer Soft Data Fusion for Indoor Robot WiFi-Visual Localization

Yuehua Ding, Jean-François Dollinger, Vincent Vauchey, Mourad Zghal

***Abstract*— This paper presents a novel WiFi-Visual data fusion method for indoor robot (TIAGO++) localization. Long-term follow-up experiments show that this method can use 10 WiFi samples and 4 low-resolution images (58×58 in pixels) to localize an indoor robot with an average error distance of about 1.32 meters 3 months (or 1.7 meters 7 months) after training data collection.**

Instead of neural network design, this paper focuses on soft data fusion to prevent unbounded errors in visual localization. The proposed soft data fusion includes first-layer WiFi-Visual feature fusion and second-layer decision vector fusion. Firstly, motivated by the excellent capability of the neural network in image processing and recognition, temporal-spatial features are extracted from WiFi data and represented in image form. Secondly, the image-form WiFi features and the visual features taken by the robot camera are combined together, and jointly exploited by a classification neural network to produce a likelihood vector for WiFi-Visual localization. This is called first-layer fusion. Similarly, these two types of features can be separately exploited by neural networks to produce another two independent likelihood vectors. Thirdly, the three likelihood vectors are fused by Hadamard product to produce a final likelihood vector. This is called second-layer fusion. The proposed soft data fusion does not apply any threshold or prioritize any data source over the other in the fusion process. It never excludes low-probability candidate positions, which can avoid information loss due to a hard decision. The demo video is provided[a], and the code will be open to the public after the publication of this work.



***Index Terms*— Robot localization, indoor localization, WiFi localization, image, vision, neural network, data fusion, TIAGO++.**

[a]The demo video available at: https://youtu.be/__AlHGhDNSI

I. Introduction

MOBILE robots can bring enormous social and economic benefits in dealing with the challenges brought by the aging society and labor shortage. Its applications cover a wide range of areas, such as healthcare and hospitals, industry, warehousing, logistics, transportation, laboratories etc. [1]. Mobile robots can replace manual tasks such as transporting goods, surveillance patrols, operations in dangerous environments, and repetitive work. The high accuracy of indoor positioning is the premise to ensure that the mobile robot can complete tasks autonomously [2].

The problem of positioning in external environment can be solved by the ubiquitous global navigation satellite system (GNSS). However, GNSS has some difficulties in indoor environment due to complex indoor conditions, such as signal attenuation, multi-path effect, non-line of sight (NLOS) problems. On the other hand, indoor positioning accuracy, generally, needs to be at meter/sub-meter level to ensure that a robot can take a correct route.

To meet this challenge, many indoor positioning technologies have been proposed. In general, we have two categories of methods for indoor positioning: visual methods and radio methods. Visual methods are based on computer vision, image processing and artificial intelligence (AI). Radio methods are based on radio signal processing and the geometric principle or adaptation of radio data. These two types of technologies have been applied to the positioning of mobile robots, significantly increasing robots' competence. Currently, there is no technology as dominant in the field of indoor positioning as GNSS in the field of outdoor positioning. Each positioning technology has significant advantages and obvious disadvantages. The road to a true autonomous robot is still long.

Visual methods [3] [4] provide cost-effective, detailed and context-rich positioning/maps. Its image data is very suitable for neural networks processing, which achieves a great success in image recognition. However, this method requires a lot of labeled images to achieve good precision, which requires significant calculation time. Additionally, they tend to fail in homogeneous spaces with limited functionality such as long corridors or poor interior lighting conditions. One interesting example is that the glass door looks different during the day and at night. One can see the outdoor scenery from the glass door during the day, but at night the glass door reflects the indoor scene. This visual difference brought by the changes of

Yuehua Ding, Jean-François Dollinger, Vincent Vauchey and Mourad Zghal are with the laboratory CESI LINEACT UR7527, France.
Correspondence: Yuehua Ding, yding@cesi.fr

luminosity condition can significantly degrade the localization performance.

The radio methods are quite rich. There exist various technologies such as ultra-wide band (UWB) [5], frequency modulated continuous wave (FMCW) [6], WiFi [7] - [11] [13] [14] etc. UWB can achieve high accuracy localization, but it has a risk of spectral conflict, additional sensor deployment is needed. UWB based localization requires the implementation of a specific anchor architecture, its localization performance is related to the density of sensor deployment. FMCW can combine radar technology and laser technology to achieve an accurate localization, however, it is very expensive in cost, and it is complex in hardware and software development. WiFi is already ubiquitously deployed with a huge number of access points (AP), it is cheap and convenient, but its accuracy is not high. Compared to visual methods, signal processing for radio methods is not very complicated. In addition, light detection and ranging (LiDAR) [12] can provide long-range, high-resolution sensing, but it is very expensive and still suffers from limitations in the case of similar geometric environments and robot kidnapping (with unknown initial position).

Based on the previous comparison, the authors of this paper observe that both camera and WiFi are ubiquitously available with low cost. Localization using vision can achieve excellent accuracy in most of the time, however, it may produce catastrophic unbounded localization errors between the points having similar visual contexts regardless their physical distance. WiFi localization is not so accurate as the former due to significant fluctuations of WiFi signal strength. Thanks to the propagation properties of radio signal, WiFi's coverage range prevents its positioning error from being a catastrophic value.

Inspired by the complementary natures of WiFi localization and visual localization, this paper focuses on a WiFi-Visual data fusion, instead of neural network design. The objective is to combine the generally excellent localization accuracy of visual positioning and the bounded localization errors of WiFi localization. The contributions of this paper are summarized as follows:

Firstly, instead of exploiting WiFi signal strength directly, intrinsic features are extracted from WiFi data. These features reflect the temporal-spatial spectrum of different WiFi access points, and the correlation across different access points as well. Compared with the fluctuating signal strength, spectrum and correlation properties are more stable signatures of a position. They are represented in the form of image to adapt the strong capability of neural network in image recognition.

Secondly, a double-layer soft data fusion is proposed. In the process of data fusion, neither threshold nor data source priority is applied. In the first-layer fusion, the WiFi features in image form and the visual features provided by the camera are combined together, and jointly exploited by a neural network. By modeling the localization problem as a classification, a likelihood vector is produced by the neural network as a response to the input features. In the second layer fusion, another two likelihood vectors are obtained by separately exploiting the WiFi features and visual features. The decision vector fusion is realized by Hadamard product and median filtering on the three likelihood vectors to produce the final decision vector for localization.

Thirdly, the proposed method has the following distinctions from the existing methods: the proposed method does not need any additional devices except a general WiFi card and a general camera. Feature extraction and feature fusion are completed by simple classic signal processing rather than neural networking to improve computational efficiency. The data fusion process, including feature fusion and likelihood fusion, does not apply any threshold or priority.

Finally, in-field experiments with a true robot TIAGO++ are carried out in quite a general scenario at the ground floor of a teaching building. The test experiment shows that the proposed localization method can localize the robot with an average error distance about 1.32 meters by using 10 WiFi samples and 4 low-resolution images. In particular, the test experiment is 3 months after the data collection, which reflects the robustness of the proposed method.

The rest of this paper is organized as follows: Section II presents the related works; Section III describes a general model for WiFi-Visual robot localization. Section IV proposes a WiFi-Visual data fusion method for localization. The experiment results are discussed in Section V. Section VI draws the conclusions.

Notations: Capital letters of boldface are used for denoting matrices. Lowercase letters of boldface denote column vectors. $(\cdot)^T$ denotes the operation of matrix transpose. $\odot$ represents Hadamard product. $\mathbb{A}(\mathbf{M})$ (or $\mathbb{P}(\mathbf{M})$) represents the matrix obtained by element-wise operation of taking the amplitudes (or phases) of the elements in matrix $\mathbf{M}$. $\mathbf{M}^{\frac{1}{3}}$ represents the matrix generated by taking element-wise cubic root of matrix $\mathbf{M}$. $\|(\cdot)\|_2$ takes the L_2-norm of $(\cdot)$. $\mathbb{E}[\cdot]$ is the statistical expectation. $\mathbf{1}$ is a column vector with 1 as its elements.

II. Related Works

A. WiFi localization

WiFi draws increasing attentions in the domain of indoor robot localization. Its omnipresence can compensate the absence (or weak presence) of GPS signal in indoor environments. Its low-cost deployment makes its wide application economically possible. However, in technique aspect, WiFi signal strength fluctuation can significantly degrade the localization accuracy by using received signal strength indicator (RSSI) [13] [14]. To improve the accuracy, denser fingerprint sampling can be adopted to achieve higher accuracy [7]. One should note that denser fingerprint sampling needs more labor cost and time cost. [8] uses data augmentation technique to enlarge the WiFi fingerprint dataset for higher localization accuracy. Actually, we can not infinitely improve the accuracy by increasing the fingerprint sampling density due to the ambiguity caused by sensitivity level of WiFi device. WiFi signal strength fluctuation can largely increase the surface of ambiguity zone. To deal with this issue, other signal properties such as direction of arrival (DoA), time of arrival (ToA), time difference of arrival (TDoA), time of flight (ToF) are exploited [9]. [9] constructs a heat map by using these radio properties, and a machine learning method is employed for localization.

[10] [11] exploit the WiFi channel state information (CSI) for localization by using a special WiFi card whose CSI is available for users. The methods [9] [10] [11] exploiting additional radio properties, such as DoA, ToA, TDoA, ToF or CSI, can achieve very good localization performance, but they need additional or special devices, which limits their wide applications.

B. WiFi-Visual localization

The rich environment information in image motivates the research in visual localization. Its localization accuracy and low-cost feature make its wide application possible. Nevertheless, the computing burden for image matching, and the visual aliasing [15] [16] in homogeneous environments are two main challenges. To cope with these challenges, WiFi-Visual localization is a good balance. [9] realizes a WiFi localization in the way of image processing, which represents the radio properties of WiFi signal in image form to adapt the strong competence of neural network in image processing. A data fusion approach based on threshold is proposed by [17] to integrate image and WiFi. [18] proposes a multi-scale strategy for WiFi-Visual localization. An indoor localization based on sequential data fusion is proposed by [19]. According to [19], WiFi signals are used for coarse localization, and the images are exploited to refine the localization results, which is also the case in [20]. In [21], a sequential-multi-decision fusion is proposed for WiFi-Visual localization, where Gaussian process regression and hybrid whale optimization algorithm are used. To the best knowledge of the authors, most of these WiFi-visual methods are based on a sequential decision process by setting a threshold or priority, such as a coarse WiFi localization followed by a refined visual localization ([9] needs additional devices).

III. System model

A WiFi-Visual robot localization system is illustrated by Fig. 1. A robot localizes itself by processing the data collected

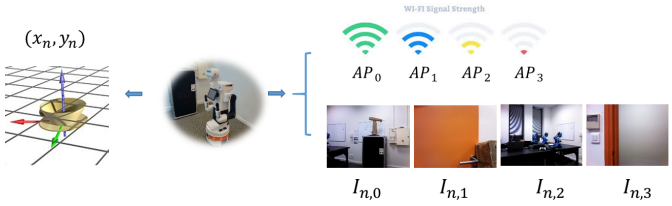


Fig. 1: WiFi-Visual robot localization system.

by its WiFi antenna and camera. Mathematically, the positioning process can be abstracted by a function as follows:

$$(\hat{x}_n, \hat{y}_n) = f_{wv}(\mathbf{S}_n, \mathbf{I}_n) \tag{1}$$

where $f_{wv}(\cdot)$ represents a WiFi-Visual localization method, such as but not limited to the methods based on KNN, LSTM or CNN. $(\hat{x}_n, \hat{y}_n)$ represents the estimated coordinate of the n^{th} position (x_n, y_n), where WiFi data $\mathbf{S}_n$ and image data $\mathbf{I}_n$ are collected for localization. The WiFi data $\mathbf{S}_n$ is usually a group of RSSI samples, $\mathbf{S}_n$ contains M RSSI samples of K access points, $\mathbf{S}_n$ can be expressed as an $M \times K$ matrix in Eq. (2)

$$\mathbf{S}_n = \begin{bmatrix} s_{n,0,0} & s_{n,0,1} & \cdots & s_{n,0,K-1} \\ s_{n,1,0} & s_{n,1,1} & \cdots & s_{n,1,K-1} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n,M-1,0} & s_{n,M-1,1} & \cdots & s_{n,M-1,K-1} \end{bmatrix} \tag{2}$$

The image data $\mathbf{I}_n$ contains a sequence of S images taken by the camera of the robot at position (x_n, y_n) . $\mathbf{I}_n$ can be represented as follows:

$$\mathbf{I}_n = \begin{bmatrix} \mathbf{I}_{n,0} & \mathbf{I}_{n,1} & \cdots & \mathbf{I}_{n,S-1} \end{bmatrix} \tag{3}$$

The objective is the minimization of localization error.

$$min \quad \|(\hat{x}_n, \hat{y}_n) - (x_n, y_n)\|_2 \tag{4}$$

In particular, the robot can also perform the WiFi localization or image localization separately, which can be respectively formulated as follows:

$$(\bar{x}_n, \bar{y}_n) = f_w(\mathbf{S}_n) \tag{5}$$

$$(\tilde{x}_n, \tilde{y}_n) = f_v(\mathbf{I}_n) \tag{6}$$

where $f_w(\cdot)$ and $f_v(\cdot)$ represent a WiFi-only localization method and an image-only localization method, respectively.

In this paper, the localization is realized by means of classification. The plan is partitioned into N small pieces of disjoint areas, the n^{th} piece is defined as class L_n with $n = 0, 1, \cdots, N-1$. The localization in (1) is re-modeled as:

$$(\hat{x}_n, \hat{y}_n) \leftarrow \hat{L}_n = f_{wv}(\mathbf{S}_n, \mathbf{I}_n) \tag{7}$$

In (7), $(\hat{x}_n, \hat{y}_n)$ is the centre position of the area labelled as class $\hat{L}_n$. $\hat{L}_n$ is the estimate of L_n

IV. Proposed method

This section presents a novel robot localization method, which jointly exploits WiFi and visual perception. In particular, the information provided by WiFi perception can be independently used for WiFi localization. This is also the case for visual perception in visual localization. The exploitation of different types of information is discussed as follows, one should note that the preprocessing on $\mathbf{S}_n$ is needed before the real processing. With a little abuse of notation, $\mathbf{S}_n$ is still used to represent the preprocessed data.

A. Theoretical motivation

The proposed method is inspired by the white spectrum of white Gaussian random noise, whose constant spectrum is much more stable than its temporal waveform, as shown in Fig. 2. Actually, white noise is a special case of random signal with zero temporal correlation. In general, the temporal correlation will influence the signal spectrum; in particular, the signal spectrum bandwidth B is proportional to the inverse of the correlation time Δ [22], as described by (8) as follows:

$$B \propto \frac{1}{\Delta} \tag{8}$$

Fig. 3 shows an example of signal spectrum with different correlation coefficient α. These features can be exploited to

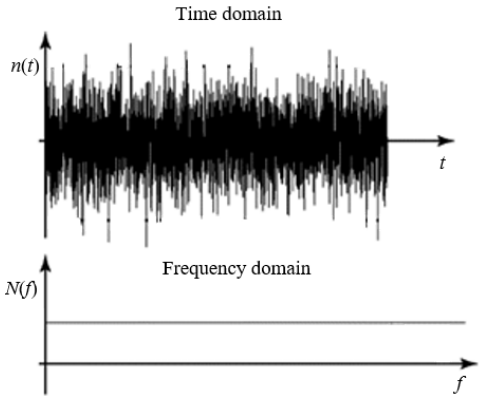


Fig. 2: Spectrum of white noise

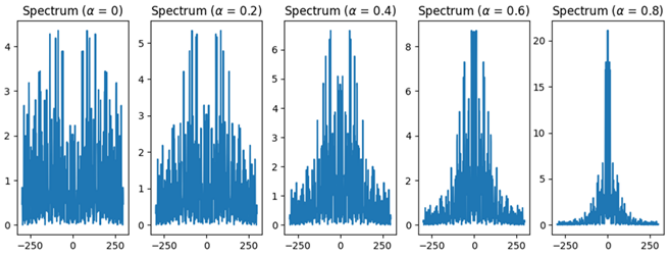


Fig. 3: Spectrum of signals with different correlations α

identify different signal types. The RSSI of WiFi signal can be modeled by log-normal distribution [23]- [25], the RSSI of WiFi signals exhibits the following properties:

Temporal correlation: Due to the continuous physical connections of users for continuous WiFi services, the RSSI of WiFi signal exhibits temporal correlation. The physical properties of an individual device used by each access point can characterize the signal transmission behavior of the access point. The physical position of the access point may also influence the density of users' connectivity.

Spatial correlation: Physical separation by large-scale obstacles can attenuate the propagation of WiFi signal from one side of the obstacle to the other side. This brings spatial correlation among the WiFi signals transmitted by the access points on the same side of the obstacle. Fig. 4 shows the spectrum of WiFi signal received at different positions from the same access point (AP0), and the correlation matrix across different AP as well.

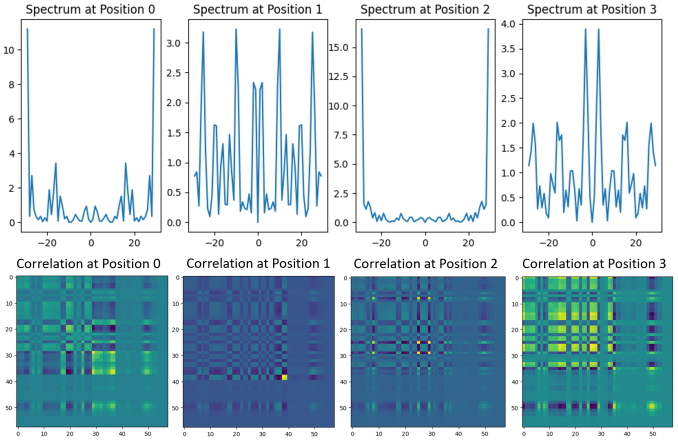


Fig. 4: Signal spectrum (AP0) and correlation matrix

These features characterized by the relative positions between the access points and user positions can be exploited for localization in the following parts of this paper.

B. WiFi localization

The main challenge of WiFi localization is the significant fluctuation of WiFi signals' strength in time domain. To cope with this problem, the intrinsic features of WiFi signals are extracted from the WiFi data in Eq. (2). These features include the temporal-spatial spectrum, the correlation matrix of access points.

1) Temporal-spatial spectrum: A row in Eq. (2) is a WiFi RSSI sample from different access points, while a column represents a sequence measuring the temporal variation of RSSI from a single access point. To characterize the features, two dimensional Discrete Fourier Transform (DFT) can be applied on $\mathbf{S}_n$.

$$\tilde{\mathbf{S}}_n = \mathbf{F}_M^T \mathbf{S}_n \mathbf{F}_K \tag{9}$$

where $\mathbf{F}_M$ (or $\mathbf{F}_K$) represents the DFT matrix. Without loss of generality, $\mathbf{F}_M$ is given by

$$\mathbf{F}_M = \frac{1}{\sqrt{M}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{M-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(M-1)} \\ 1 & \omega^3 & \omega^6 & \cdots & \omega^{3(M-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{M-1} & \omega^{2(M-1)} & \cdots & \omega^{(M-1)(M-1)} \end{bmatrix} \tag{10}$$

with $\omega = e^{-2\pi j/N}$, $j^2 = -1$.

$\tilde{\mathbf{S}}_n$ represents the temporal-spatial spectrum of the WiFi signal strength received at the n^{th} position. The m^{th} row of $\tilde{\mathbf{S}}_n$ represents the spatial spectrum with respect to different access points at the instant of the m^{th} sampling. The k^{th} column represents the temporal spectrum of the k^{th} access point.

To exploit the temporal-spatial spectrum for localization, the amplitude and phase of $\tilde{\mathbf{S}}_n$ are extracted as two features, represented by $\mathbb{A}(\tilde{\mathbf{S}}_n)$ and $\mathbb{P}(\tilde{\mathbf{S}}_n)$ respectively.

2) Correlation matrix of access points: The correlation between access point k and k' is theoretically given by

$$r_{n,k,k'} = \frac{\mathbb{E}[(s_{n,m,k} - \mu_k)(s_{n,m,k'} - \mu_{k'})]}{\sigma_k \sigma_{k'}} \tag{11}$$

where μ_k and σ_k are the mean and standard deviation of RSSI from access point k, respectively. For the preprocessed $\mathbf{S}_n$, μ_k is centered to 0, and σ_k is normalized to 1. In practice, $\mathbb{E}[\cdot]$ is replaced by averaging. The correlation matrix of access points at n^{th} position is calculated as:

$$\mathbf{R}_n = \frac{\mathbf{S}_n^T \mathbf{S}_n}{M} \tag{12}$$

In (12), $\mathbf{R}_n$ is a $K \times K$ symmetric matrix.

In WiFi perception, $\mathbb{A}(\tilde{\mathbf{S}}_n)$, $\mathbb{P}(\tilde{\mathbf{S}}_n)$ and $\mathbf{R}_n$ are three different features to be used. To keep the three matrices of the

same dimensions, additional points can be added to $\mathbf{S}_n$ or $\mathbf{R}_n$ if $M \neq K$. Fig.5 is an example of the visualization of WiFi features.

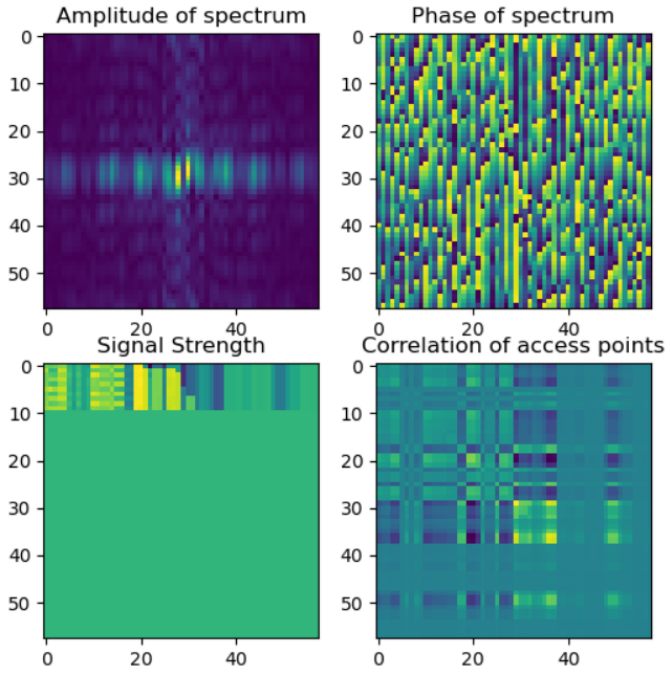


Fig. 5: WiFi RSSI features in image form

According to (5), the image-form WiFi features can be separately used as follows:

$$\mathbf{p}_w = f_w(\mathbb{A}(\tilde{\mathbf{S}}_n), \mathbb{P}(\tilde{\mathbf{S}}_n), \mathbf{R}_n, \mathbf{S}_n) \tag{13}$$

where f_w represents a neural network in this paper. $\mathbf{p}_w$ is a vector of conditional probability given by

$$\mathbf{p}_w = [p((x_1, y_1)|\mathbf{S}_n), p((x_2, y_2)|\mathbf{S}_n), \cdots, p((x_N, y_N)|\mathbf{S}_n)]^T \tag{14}$$

Its element $p(x_i, y_i|\mathbf{S}_n)$ indicates the likelihood of the i^{th} candidate position. For WiFi localization, the estimated position $(\bar{x}_n, \bar{y}_n)$ is given by

$$p((\bar{x}_n, \bar{y}_n)|\mathbf{S}_n) = \max_i \quad p((x_i, y_i)|\mathbf{S}_n) \tag{15}$$

Thanks to Bayes' Theorem, (15) is reformulated as

$$p((\bar{x}_n, \bar{y}_n)|\mathbf{S}_n) = \max_i \quad \frac{p(\mathbf{S}_n|(x_i, y_i))p((x_i, y_i))}{p(\mathbf{S}_n)} \tag{16}$$

With the assumption that the N positions (classes) are equally probable, $p((x_i, y_i))$ is a constant of $1/N$, (16) is equivalent to

$$\begin{aligned} p((\bar{x}_n, \bar{y}_n)|\mathbf{S}_n) &= \max_i \ p((x_i, y_i)|\mathbf{S}_n) \\ &\Longleftrightarrow \quad \max_i \ p(\mathbf{S}_n|(x_i, y_i)) \end{aligned} \tag{17}$$

C. Visual localization

In this paper, the visual perception is relatively simple. The photo sequence $\mathbf{I}_{n,0}, \mathbf{I}_{n,1}, \cdots \mathbf{I}_{n,S-1}$ taken at the n^{th} position are sent to the neural network $f_v(\cdot)$ in (6) for learning in the training process or for localization in test.

$$\mathbf{p}_v = f_v(\mathbf{I}_n) \tag{18}$$

where $\mathbf{p}_v$ is a probability vector produced by the neural network $f_v(\cdot)$. $\mathbf{p}_v$ is expressed by

$$\mathbf{p}_v = [p((x_1, y_1)|\mathbf{I}_n), p((x_2, y_2)|\mathbf{I}_n), \cdots, p((x_N, y_N)|\mathbf{I}_n)]^T \tag{19}$$

For the localization by vision, the coordinates corresponding to the maximum element in Eq. (19) is taken as the robot position. By employing Bayes' Theorem, the problem is reformulated as follows

$$p((\tilde{x}_n, \tilde{y}_n)|\mathbf{I}_n) = \max_i \ p((x_i, y_i)|\mathbf{I}_n) \iff \max_i \ p(\mathbf{I}_n|(x_i, y_i)) \tag{20}$$

D. First-layer WiFi-Visual feature fusion

Both WiFi features and image features can be represented in the image form. To facilitate the data fusion, it is necessary to harmonize their dimensions, in particular, $\mathbf{R}_n$, $\mathbf{S}_n$, $\mathbf{I}_n$ are reshaped as the pictures of the same dimensions. They are jointly exploited to localize the robot as follows:

$$\mathbf{p}_{wv} = f_{wv}(\mathbb{A}(\tilde{\mathbf{S}}_n), \mathbb{P}(\tilde{\mathbf{S}}_n), \mathbf{R}_n, \mathbf{S}_n, \mathbf{I}_n) \tag{21}$$

where $\mathbf{p}_{wv}$ is a probability vector produced by the neural network f_{wv}. $\mathbf{p}_{wv}$ is expressed by

$$\mathbf{p}_{wv} = [p((x_1, y_1)|\mathbf{S}_n, \mathbf{I}_n), \cdots, p((x_N, y_N)|\mathbf{S}_n, \mathbf{I}_n)]^T \tag{22}$$

Eq. (21) can provide a position estimate of the robot. The position estimate is given by

$$p((\underline{x}_n, \underline{y}_n)|\mathbf{S}_n, \mathbf{I}_n) = \max_i \ p((x_i, y_i)|\mathbf{S}_n, \mathbf{I}_n) \tag{23}$$

Similarly, $\max_i \ p((x_i, y_i)|\mathbf{S}_n, \mathbf{I}_n)$ in (23) is equivalent to

$$\max_i \ p((x_i, y_i)|\mathbf{S}_n, \mathbf{I}_n) \iff \max_i \ p(\mathbf{S}_n, \mathbf{I}_n|(x_i, y_i)) \tag{24}$$

E. Second-layer decision vector fusion

One notes that $p((x_i, y_i)|\mathbf{S}_n, \mathbf{I}_n)$ and $p(\mathbf{S}_n, \mathbf{I}_n|(x_i, y_i))$ are non-negative. (24) is equivalent to the formula as follows

$$\max_i \ [p((x_i, y_i)|\mathbf{S}_n, \mathbf{I}_n)]^2 \iff \max_i \ [p(\mathbf{S}_n, \mathbf{I}_n|(x_i, y_i))]^2 \tag{25}$$

Supposing that $p(\mathbf{S}_n|(x_i, y_i))$ and $p(\mathbf{I}_n|(x_i, y_i))$ are independent [20] [21], $p(\mathbf{S}_n, \mathbf{I}_n|(x_i, y_i))$ can be represented as:

$$p(\mathbf{S}_n, \mathbf{I}_n|(x_i, y_i)) = p(\mathbf{S}_n|(x_i, y_i))p(\mathbf{I}_n|(x_i, y_i)) \tag{26}$$

$\max_i \ [p((x_i, y_i)|\mathbf{S}_n, \mathbf{I}_n)]^2$ is equivalent to the maximization as follows:

$$\begin{aligned} \max_i \ p((x_i, y_i)|\mathbf{S}_n, \mathbf{I}_n) &\iff \max_i \ [p((x_i, y_i)|\mathbf{S}_n, \mathbf{I}_n)]^2 \\ &\iff \max_i \ p(\mathbf{S}_n|(x_i, y_i))p(\mathbf{I}_n|(x_i, y_i))p(\mathbf{S}_n, \mathbf{I}_n|(x_i, y_i)) \end{aligned} \tag{27}$$

Combining (17), (20), (24) and (27), we have

$$\max_i \ p((x_i, y_i)|\mathbf{S}_n, \mathbf{I}_n) \iff \max \ [\mathbf{p}_w \bigodot \mathbf{p}_v \bigodot \mathbf{p}_{wv}] \tag{28}$$

where $\bigodot$ represents the Hadamard product. To fuse the information in $\mathbf{p}_w$, $\mathbf{p}_v$, $\mathbf{p}_{wv}$, the Hadamard product of $\mathbf{p}_w$, $\mathbf{p}_v$, $\mathbf{p}_{wv}$ is taken as a new vector of likelihood.

$$\mathbf{p}_{wvm} = \mathbf{p}_w \bigodot \mathbf{p}_v \bigodot \mathbf{p}_{wv} \tag{29}$$

For the data fusion of WiFi data and image data, we neither prioritize different data sources nor use the prediction based on one data source to restrict the prediction based on other data source. A natural data fusion is adopted with respect to the likelihoods provided by the predictions from different data sources.

One notes that the sum of $\mathbf{p}_{wvm}$ given by (29) is not necessarily equal to 1. $\mathbf{p}_{wvm}$ should be normalized to a regular likelihood vector.

$$\mathbf{p}_{wvmn} = \frac{(\mathbf{p}_w \bigodot \mathbf{p}_v \bigodot \mathbf{p}_{wv})^{\frac{1}{3}}}{\mathbf{1}^T(\mathbf{p}_w \bigodot \mathbf{p}_v \bigodot \mathbf{p}_{wv})^{\frac{1}{3}}} \tag{30}$$

where $(\cdot)^{\frac{1}{3}}$ represents element-wise cubic root, $\mathbf{1}$ is an all-one column vector.

The principle of Eq.(23) can also be applied on Eq. (29) to localize the robot. In this paper, additional smooth filtering (such as weighted sum) is applied on $\mathbf{p}_{wvmn}$ to produce the final result.

Remarks:

1. The first-layer data fusion can enable a balance between image localization and WiFi localization. The visual similarities that may bring catastrophic localization errors can be traded-off by the WiFi features. The performance of first-layer data fusion is usually intermediate between that of WiFi localization and image localization, but is more stable.

2. The second-layer data fusion is actually a soft voting mechanism. Inspired by the Viterbi soft decoding [26] of the convolution code, the second-layer data fusion never excludes any candidate position until the final decision. One should note that the product of the three likelihoods can achieve a soft voting balance between WiFi and image localization. This can significantly weaken the incoherent candidates that have sharply contrastive different likelihoods produced by the three methods.

3. $\mathbf{p}_{wvmn}$ in (30) is actually a normalized vector of geometric mean (GM). Theoretical analysis is given in the Appendix, which shows that GM is better than arithmetic mean (AM) in enhancing consistent results and weakening divergent ones.

V. Experiments

A. Experiment platform

The experiment platform is a real physical robot TIAGO++ [27]. TIAGO++ is a fully ROS-based, customizable robot platform, it is adapted to the research needs in AI, machine learning, human-robot interaction (HRI), and manipulation. WiFi card of 802.11ax WiFi 6 and RGB-D camera are integrated into the robot platform, they can be used directly to collect WiFi and visual data. For performance evaluation, thanks to the LiDAR position and mapping system in TIAGO++, simultaneous localization and mapping (SLAM) can be performed in the experiments to provide a reliable map and real-time positions with centimeter-level accuracy, which are considered as ground-truth values in performance evaluation in the following experiments.

To examine the effectiveness of the proposed method, extensive experiments are carried out, including mapping, data collecting and testing. To keep the generality of the experimental environment, the experiments are carried out in the ground floor of a teaching building with usual activities, as shown in Fig. 6. The training process is completed in the server. In the experiment, $K = 58$ logic access points of WiFi are considered, $M = 10$ WiFi RSSI samples and $S = 4$ photos are taken for a one-time localization. The performance is evaluated in terms of root mean square error (RMSE), mean absolute error (MAE), and standard deviation (STD), respectively. They are computed as follows:

$$RMSE = \sqrt{\frac{1}{N_{test}} \sum_{i=0}^{N_{test}-1} [(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2]} \tag{31}$$

$$MAE = \frac{1}{N_{test}} \sum_{i=0}^{N_{test}-1} (|\hat{x}_i - x_i| + |\hat{y}_i - y_i|) \tag{32}$$

$$STD = \sqrt{\frac{1}{N_{test}} \sum_{i=0}^{N_{test}-1} (e_i - \bar{e})^2} \tag{33}$$

where N_{test} represents the number of test points, $(\hat{x}_i, \hat{y}_i)$ is the estimated coordinate of test point i, (x_i, y_i) is the ground-truth value, which can be provided by the LiDAR system of the platform, e_i is the error of i^{th} point of localization, $\bar{e}$ is the average localization error.

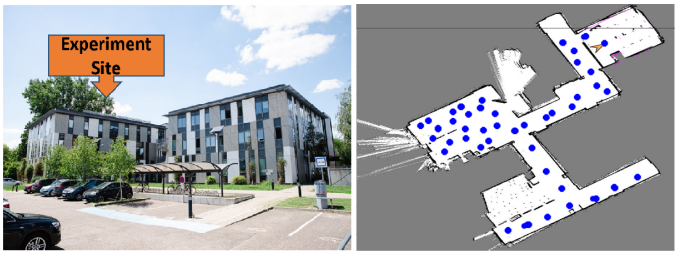


Fig. 6: Mapping in the ground floor of the teaching building.

B. Data collecting

The sampled 43 positions of data collection are marked in blue color in the map of Fig.6. For each sampled position, TIAGO++ takes 1000 WiFi RSSI samples and 100 photos around itself. In the image collecting process, TIAGO++ rotates itself by 3.6 degrees after taking each photo, as illustrated by Fig. 7. For a given sampled position, its 1000 WiFi samples are divided into 100 groups with 10 samples in each group. Its 100 photos are organized into 100 groups with 4 photos in each group. The 4 photos in a group should be uniformly spaced in terms of the camera shooting angle. Two neighboring groups are offset by a shooting angle difference of 3.6 degrees, as indicated by Fig. 8. To associate the WiFi data and the image data, all possible combinations between 100 WiFi groups and 100 images groups are taken to construct 10000 training WiFi-Visual samples.

Fig. 7: Visual perception.

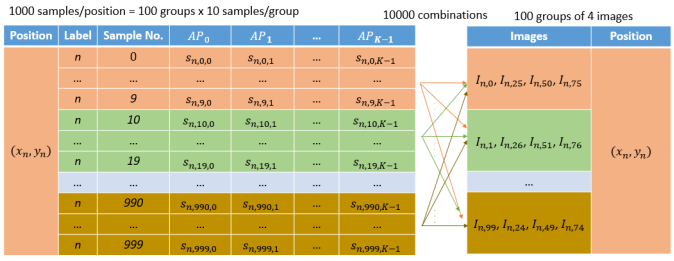


Fig. 8: Data association

C. Training

In the training process, we considered four different artificial neural networks (ANN) structures (ResNet [28], LeNet [29], LSTM [30], VLocNet [3] [4]) to adapter our dataset. Tab. I provides a comparison among these structures. ResNet is an excellent structure for image recognition, however, it is easy to have overfitting with our dataset, VLocNet is also based on ResNet. We have no overfitting with LSTM and LeNet, it is not easy to harmonize the WiFi feature and image feature to adapt the temporal sequential characteristics of LSTM. Hence, we choose LeNet as a basic structure for training. Fig. 9 illustrates the framework of training with the combinations of WiFi features and visual features. For the training with WiFi features (or image features) only, the image features (or WiFi features) can be removed directly. The input data is structured as follows:

1. For WiFi-Image localization, an input data includes four RGB images from the camera, four WiFi images (RSSI, spectrum amplitude and phase, correlation matrix) produced by 10 WiFi samples. It is represented as a $58 \times 58 \times 16$ tensor, which means that we have 16 channels.

2. For WiFi localization, an input data includes four WiFi images (RSSI, spectrum amplitude and phase, correlation matrix) produced by 10 WiFi samples. It is represented as a $58 \times 58 \times 4$ tensor, which means that we have 4 channels.

3. For image localization, an input data includes four RGB images from the camera. It is represented as a $58 \times 58 \times 12$ tensor, which means that we have 12 channels.

TABLE I: Network Comparison

	ResNet	LeNet	LSTM	VLocNet
Overfitting	yes	no	no	yes
Image adaptive	yes	yes	no	yes
Complexity	high	low	low	high

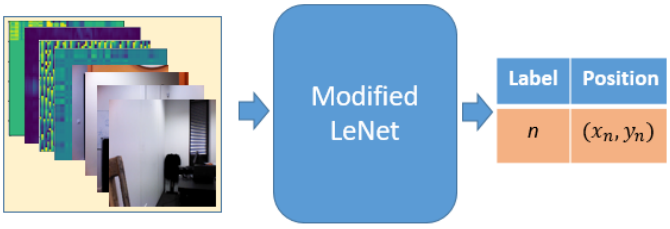


Fig. 9: Training

During the training process, we also investigated the impact of WiFi features in localization by using various trained models at different training stages, as shown in Fig. 10, which verifies the feasibility of exploitation of these features.

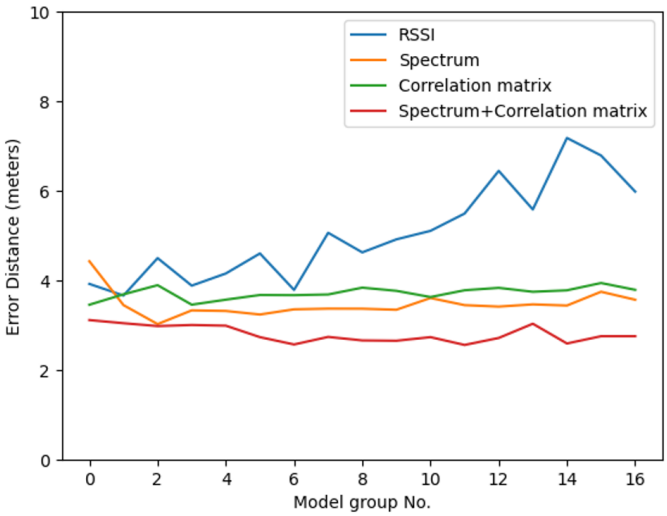


Fig. 10: Impact of WiFi features

D. "In-field" test immediately after training data collection

An immediate test is a good way to get a baseline comparison. One should note that data collection is a long process. It takes about 30 minutes for the data collection at each of the 43 labeled positions (classes). To meet the timing requirement of 'immediate test', we check the timestamp of the data for each class. In each class, the data samples are ranked from the oldest to the newest. The first 70% is taken for training, and the rest is for testing. This can guarantee an immediate test in 30 minutes. The test results are shown by Fig. 11, which clearly illustrates the performance improvements brought by each data fusion layer of WiFi-Image localization (RMSE = 0.322 m in the first layer, and RMSE = 0.177 m in the

second layer) over WiFi localization (RMSE = 0.474 m) and image localization (RMSE = 0.91 m). In addition, we can also observe that 100% of WiFi localization errors are less than 4.5 meters, which is much better than image localization. More than 96% of image localizations have smaller RMSE than WiFi localization. This immediate test shows that the proposed data fusion can successfully combine both advantages to reach an even better RMSE.

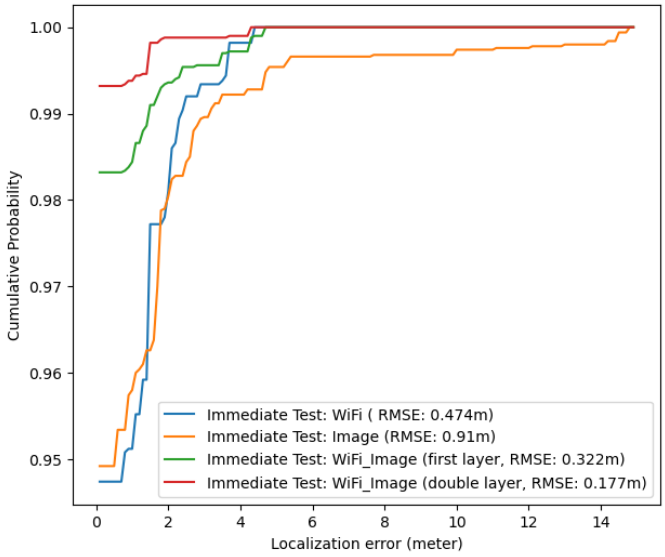


Fig. 11: CDF performance in immediate test

Remark:

In the immediate test, at least 94% of the localizations by any of the three methods are error-free, thanks to the strong temporal correlation between the data samples. The training data and the test data share almost the same features, because the feature evolution is very limited in 30 minutes. This is not the case in real applications. We need practical tests to examine its generalization capacity in a long-term environment evolution.

E. Real time in-field tests

To investigate the performance evolution of the proposed algorithm, the in-field test is carried out with TIAGO++ moving in the teaching building. It is interesting to note that the tests and data collection are temporally separated for one month, 3 months and 7 months respectively. During this process, the WiFi infrastructures are partially reorganized by the technique staff, the layout of the teaching building has been partially changed to adapt to the students' various activities. The environment evolution makes the localization more and more complicated, challenging the robustness of the method.

In the test, we choose remote interactions between a remote terminal (PC) and the robot to simulate the working environment and reduce interference, however, we don't deliberately change the natural distribution of the people around the robot in the building. Fig. 12 shows the test framework, a remote PC and the robot TIAGO++ are connected to the same WiFi, a localization request is sent from the remote PC to the robot, the robot collects $M = 10$ WiFi RSSI samples and $S = 4$ images, and sends them to the remote PC. The remote PC can use the received data and the trained model to localize the robot.

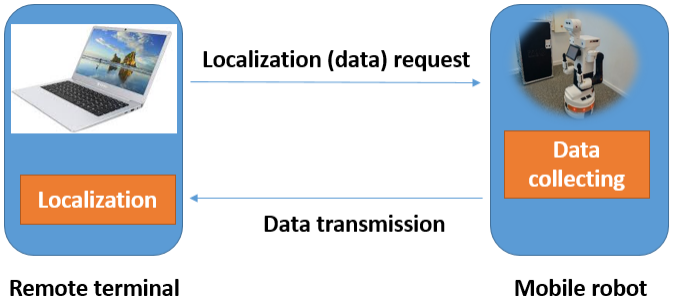


Fig. 12: The framework of in-field test

1) Test within one month after the training data collection: In the test, the robot moves in the corridors and the hall of the experiment site shown by Fig.6. Its ground truth positions are marked in red in Fig. 13, which compares the localization results of the proposed method (marked as first-layer soft fusion) with those obtained by separately using WiFi features (marked as WiFi) and image features (marked as image) in training.

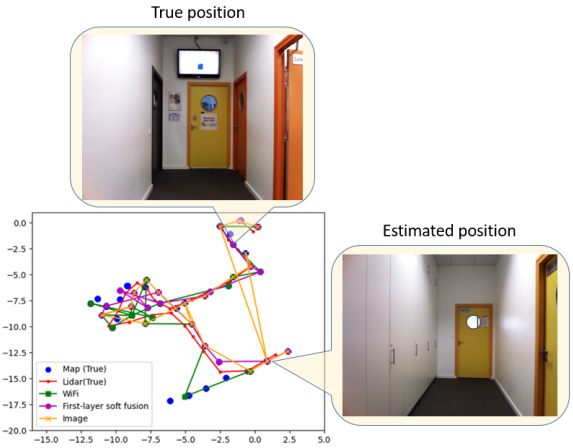


Fig. 13: Trajectory comparison (test within 1 month)

According to Fig. 13, the image localization trajectory is closest to the true trajectory except a catastrophic point, which is produced by the visual ambiguity between two similar corridors. Fig. 14 shows that this error distance is up to 12 meters. In contrast, WiFi localization and WiFi-image localization have no such visual ambiguity.

Fig. 14 quantifies the localization errors of these 3 methods. In terms of average precision without counting the catastrophic situation, the image method can reach a precision of around 1.1 meters, which is the best performance among the three methods. Unfortunately, the accuracy of the image method can be seriously degraded by homogeneous environments; the overall accuracy for the image method is the worst at 2.63 meters. The WiFi-image method ranks the first with the RMSE of 1.59 meters, and the WiFi method is the second at the

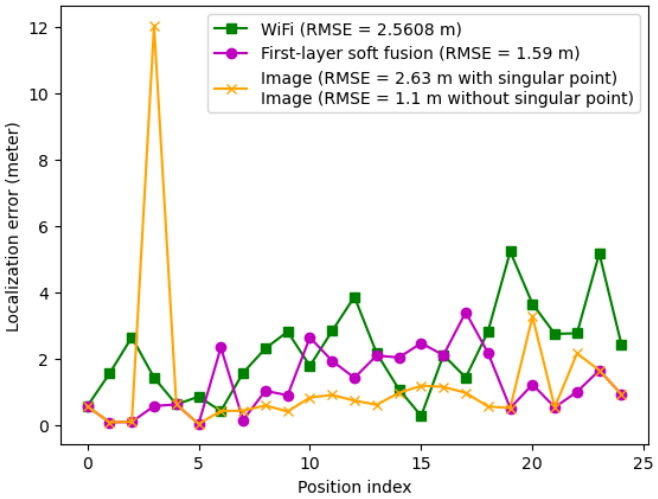


Fig. 14: Localization error comparison (test within 1 month)

level of 2.56 meters. This test shows clearly the robustness of the WiFi-image localization in a visually homogeneous environment.

Remark: In this test, double-layer soft fusion was not used. Double-layer soft fusion was still under study at that time. The comparison between the two will be made in the upcoming part.

2) Test three months after the training data collection: In one of the tests, the robot moves along a trajectory connecting halls and corridors, its positions are estimated by the remote terminal. The positions provided by LiDAR can be considered as the ground truth positions, as shown by the red trajectory in Fig. 15.

Fig. 15 compares the localization performances of the WiFi localization realized by Eq.(13), the visual localization by Eq.(18), the localization provided the first-layer soft fusion by Eq.(21) and the double-layer soft fusion localization after median filtering. We can observe that the localization RMSE of the first-layer soft fusion is 2.38 m, it is between WiFi localization (RMSE = 2.96 m) and Visual localization (RMSE = 1.29 m). The double-layer soft fusion has the best RMSE accuracy (RMSE = 1.04 m), its trajectory can closely follow the true trajectory. Detailed comparison is quantified by Fig. 16 and Fig. 17.

Fig. 16 shows the localization error distance of each estimated position. The maximum error of the double-layer soft fusion is about 3 meters, which is the same case for the visual localization. One should note that the occurrence of big-error localization of double-layer soft fusion is much less than the visual localization. For double-layer soft fusion, most of its localization errors are smaller than 2 meters. Fig. 17 shows the statistics of the localization accuracy in terms of cumulative density function (CDF). About 90% of the localization points by double-layer soft fusion can achieve an accuracy less than 2 meters, this percentage is only 80% for the visual localization. In addition, double-layer soft fusion localization has better performance than the others in terms of MAE and STD, as shown in Fig. 16.

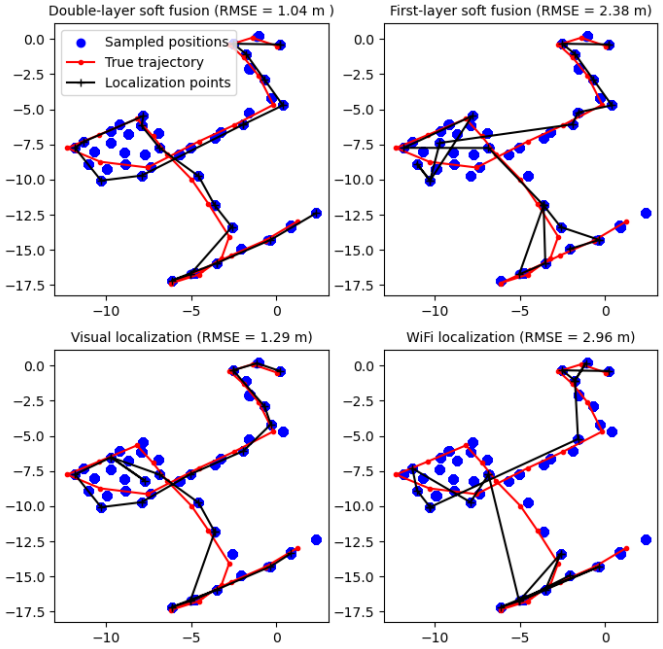


Fig. 15: Trajectory comparison among different localization modes (test after 3 months)

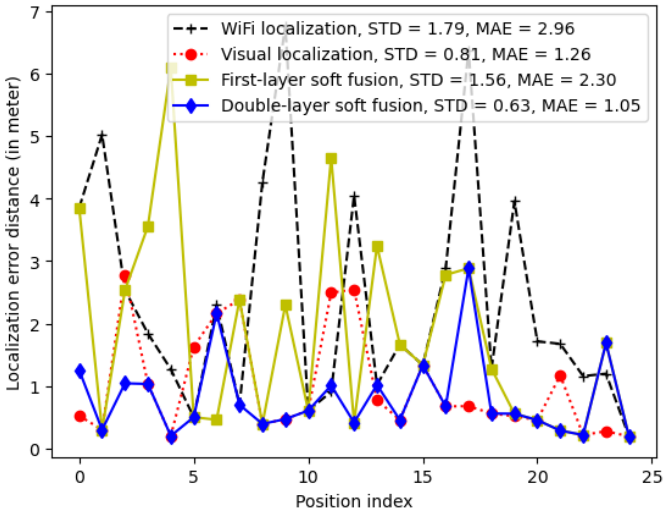


Fig. 16: Localization error comparison (test after 3 months)

3) Comparison with other data fusion methods: The double-layer soft data fusion method is compared with the following data fusion criteria:

· Data fusion based on a threshold of distance d: the threshold is a value of distance. A circle of a given radius (threshold) is chosen, the circle center is given by WIFI localization. The WiFi-Visual localization will be completely or partially restricted by this circle.

· Data fusion based on a threshold of probability γ: the threshold is a probability value. The candidates are filtered according to their individual probabilities, or the sum of their probabilities, instead of their physical distances to a reference point.

· Top-K data fusion: K is a fixed number. K best candidates

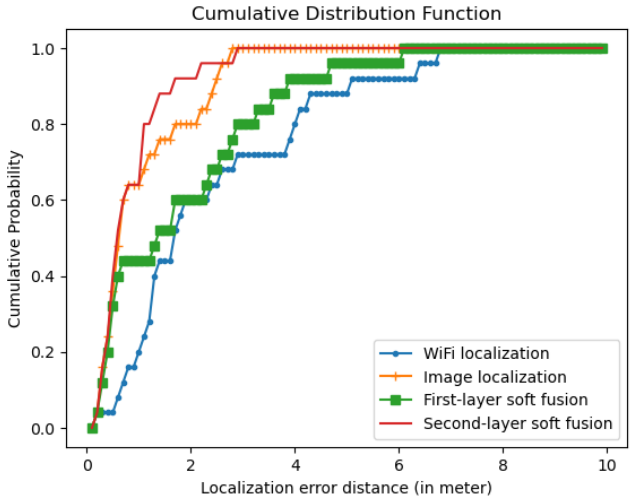


Fig. 17: Localization error comparison (CDF) (test after 3 months)

are chosen for localization refinement.

These data fusion criteria or their hybrid versions are widely used in the existing literature [17]- [21].

Another group of tests with $N_{test} = 98$ is done to compare the proposed method with the above data fusion strategies. The data fusion idea of the very recent method in [21] (named 'Tang's method') is also compared with the proposed method.

Fig.18a shows the CDF performances of the proposed method and the distance-threshold data fusion. One can note that the distance-threshold data fusion can refine the localization well when the localization error is smaller than 2 meters, in this case, its performance is a little bit better than the proposed method. However, when the localization error increases, it is not so effective to refine the accuracy. Even if a bigger threshold value such as $d = 6$ m is used, there is little performance improvement. Actually, increasing the threshold value means relaxing the constraint. Fig.18a shows that the proposed method is more effective to limit the large errors.

Fig.18b compares the proposed method with the probability-threshold data fusion. The threshold γ = 0.9 has the worst performance in error interval [0, 4], because γ = 0.9 is not sufficient to include the important candidate positions, in this case, the true positions are likely to be excluded from the candidate set. One can note that increasing the value of γ can improve the performances, but it is not very helpful in limiting the large errors. In particular, γ = 0.98 has the worst performance in limiting large errors bigger than 4 meters, because a big γ value means relaxed constraint. γ = 1 represent no constraint. The proposed method shows its advantage in limiting the large localization errors if error is bigger than 2 meters.

The proposed method is compared with Top-K data fusion in Fig.18c. For the data fusion with $K = 5$, it has the worst performance in error interval [0, 3.5], because it is likely to exclude the true position from the candidate set. Top-K with $K = 10$ can improve the performance in error interval [0, 3.5], but it is not good at limiting big errors. $K = 15$ brings even higher risk of big localization error than $K = 10$. Similarly, a big K means relaxed threshold.

Fig.18d provides a comparison between the proposed method and 'Tang's method' [21], which is a recently published hybrid data fusion criterion. The proposed method and Tang's method have very closed performances. The detailed comparison is given in Tab. II.

Tab. II compares their localization errors in terms of RMSE, MAE and STD, respectively. One can note that the RMSE and MAE of the proposed method is 1.32 meters and 1.06 meters, respectively, which are smaller than those of other methods. Thus, the proposed method can achieve the highest localization accuracy. In addition, the proposed method has the smallest STD value (1.02 meters), which means that its localization performance is more stable.

TABLE II: Performance comparison (RMSE, MAE and STD)

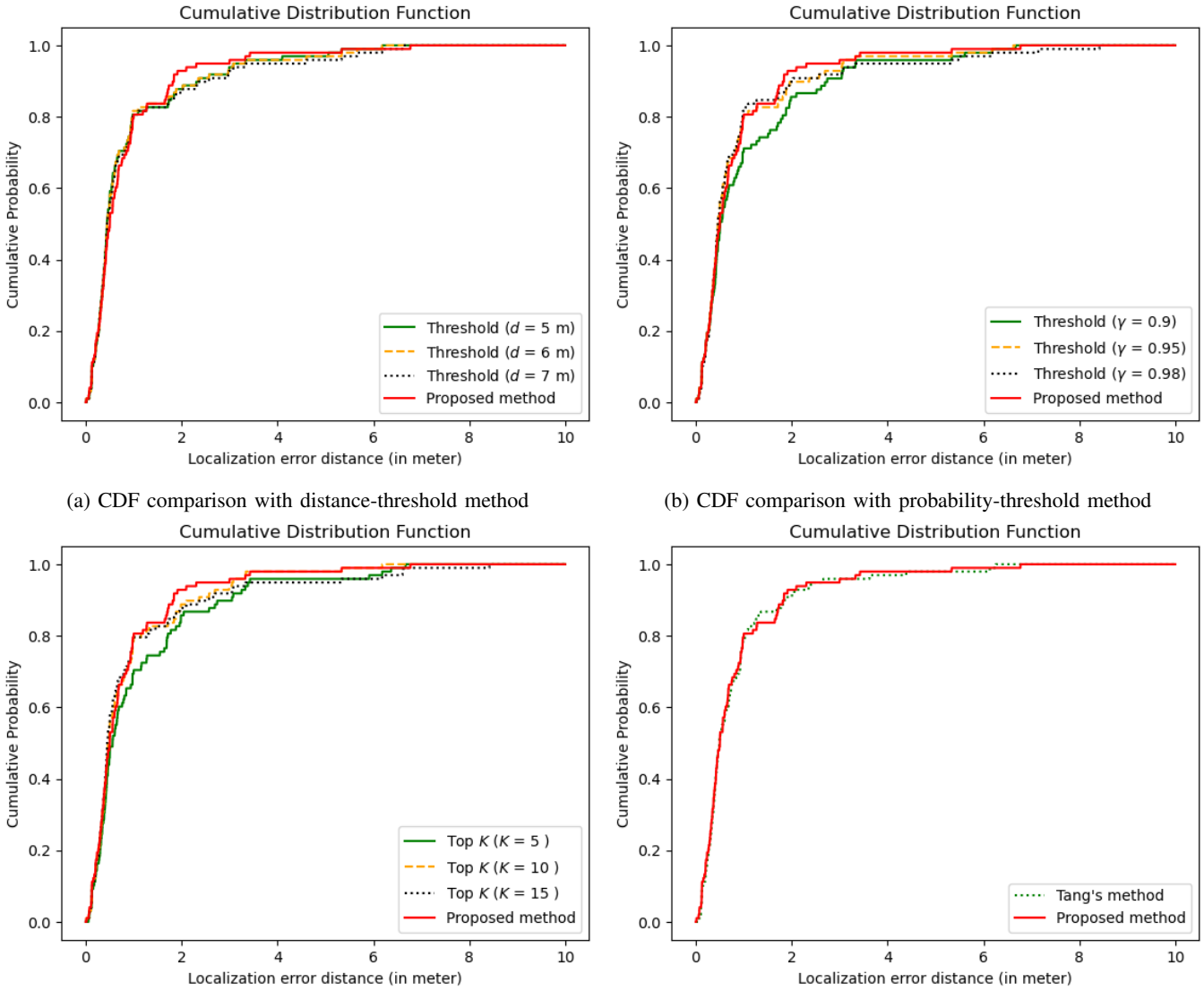
Method	RMSE	MAE	STD
Threshold ($\gamma = 0.90$)	2.013037	1.472592	1.625419
Threshold ($\gamma = 0.95$)	1.485581	1.124034	1.180261
Threshold ($\gamma = 0.98$)	1.752097	1.256527	1.459468
Threshold ($d = 5$ m)	1.453636	1.119079	1.146024
Threshold ($d = 6$ m)	1.483533	1.147843	1.180532
Threshold ($d = 7$ m)	1.626001	1.235702	1.311729
Top-K ($K = 5$)	1.768059	1.435950	1.365733
Top-K ($K = 10$)	1.382939	1.112392	1.057815
Top-K ($K = 15$)	1.808469	1.297303	1.496163
Tang's method	1.359798	1.074453	1.056154
Proposed method	1.320597	1.064133	1.019427

4) Test 7 months after the training data collection with different image sizes : 7 months after the training data collection, extra tests are done. During this period, the visual environment changed significantly, some of which are illustrated by Fig. 19 In this test, 58×58 images and 116×116 images are respectively used to examine the impact of image size on the localization performance. The Top-K ($K = 10$) and Tang's method, whose performances are close to that of the proposed method in Tab. II, are taken for comparison in Fig.20 - 21. One can observe that the visual environment change heavily degrades the image localization performance, while the proposed WiFi-Image method can still keep the robust performance, which is the best over the others. It is worth noting that the influence of the image size on the localization performance is not so significant. The RMSE produced by image-size 116×116 is 1.599 m, versus the RMSE = 1.704 m of image-size 58×58.

F. RMSE evolution

In the long-term follow-up study experiment, the authors are able to preserve most of the experiment results. In the early stage of this experiment, we used just first-layer fusion for localization. Therefore, the data of double-layer fusion is not in the history record of the first month. Fig. 22 shows the evolution process of RMSE in the passed 7 months. One can observe the following points:

- The evolution rate in the first month is the fastest. This is because the collected training data can not cover all the data distribution features. The uncovered data



(a) CDF comparison with distance-threshold method

(b) CDF comparison with probability-threshold method

(c) CDF comparison with Top-K method

(d) CDF comparison with Tang's method

Fig. 18: CDF comparison with existing fusion ideas

features in the real time will degrade the performance. One month later, the additional challenge mainly comes from human-related environment changes. For example of the image localization, the human-related visual environment change may bring dramatic perturbations to visual localization. Of course, there are possibly still a small part of uncovered features appears from time to time in the following months.

- The image localization is not stable. Its performance suffers from visual similarity even if there is little visual environment change, one can refer to the second red point in Fig. 22. Besides, its performance may degrade rapidly as the time going with visual environment change, as shown by the last red point in Fig. 22. The third red point is a lucky point with limited performance degradation compared to the immediate test (the first red point).
- The RMSE of WiFi localization changes moderately. The RMSE value at the 7-th month in Fig. 22 is even smaller than that of the 3-rd month. This can be interpreted by our newly used model training technique.
- The WiFi-image localization including first-layer fusion and double layer fusion has acceptable RMSE increments, the performance gap between the first-layer fusion and double layer fusion is relatively stable, about 1 meter.
- For the double layer data fusion, we don't have RMSE value for the first month. It was still under study at that time.

VI. Conclusions

This paper proposes a soft WiFi-Visual data fusion method for indoor robot localization. The localization problem is modeled as a classification. The WiFi features are represented in image form in order to fuse the WiFi features and visual features together. The fused WiFi-Visual features are jointly exploited by a classification neural network to produce a likelihood vector, which can classify the input features to the

Fig. 19: Examples of visual environment change

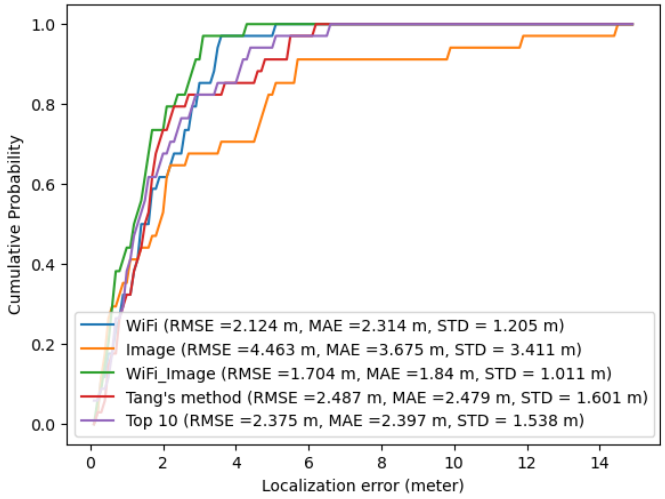


Fig. 20: CDF performance 7 months after training data collection (58 x 58 image)

most probable position. The visual features and WiFi features are also separately exploited by neural networks to produce another two likelihood vectors. The three likelihood vectors are fused by Hadamard product to produce the final likelihood vector. The proposed method is tested on an indoor robot (TIAGO++) in different scenarios, which are immediately after, within one month after, 3 months after and 7 months after the training data collection respectively. The tests show that the proposed method can use 10 WiFi samples and 4 low-resolution images (58×58 in pixels) to localize the robot with an average error distance about 1.32 meters (or 1.7 meters) 3 months (or 7 months) after training data collection. The experiment tests are in a general teaching building, whose WiFi and visual environments are always in dynamic evolution. This shows the robustness of the proposed method. This localization mode can initialize other source-based localization operating in a non-optimal manner in the event of a robot kidnapping, such as kidnapped LiDAR.

One should note that the localization results degrade along with the environment evolution. Based on our observation in this paper, it is very interesting to develop algorithms to update the fingerprint database partially and adaptively rather

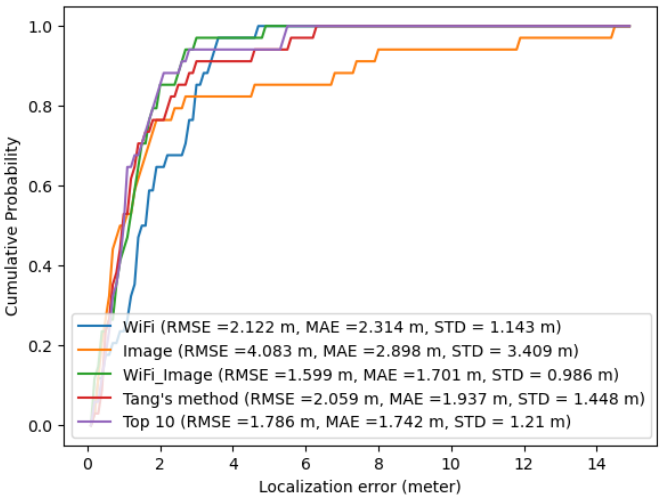


Fig. 21: CDF performance 7 months after training data collection (116 x 116 image)

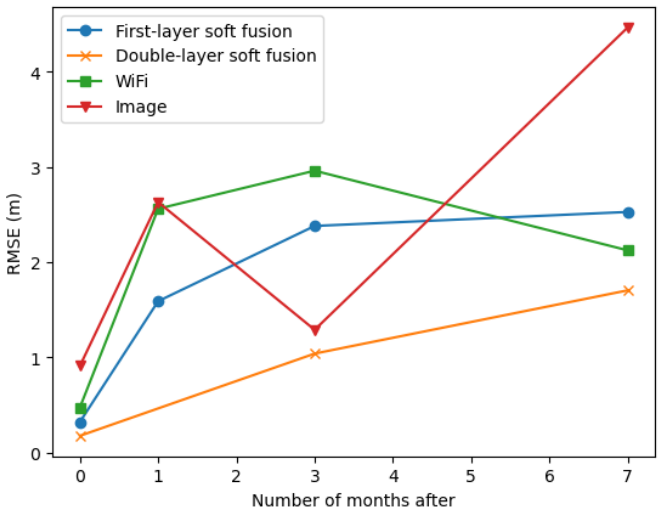


Fig. 22: RMSE evolution with time (58 x 58 image)

than updating the whole database, by analyzing the evolution curve and its corresponding environment change. Intelligent algorithm making the database following the environment local changes is a promising way, which is a research topic of our future work.

APPENDIX

The effectiveness of the proposed likelihood vector fusion method can be explained by the inequality between geometric mean (GM) and arithmetic mean (AM).

Let's place this inequality in the scenario of machine learning, where $p_1(X_i)$, $p_2(X_i)$ ...,$p_S(X_i)$ represent S different likelihood measurements of a class X_i $(i = 1, 2, \cdots, N)$ respectively. According to the GM-AM inequality [31], we have

$$\frac{\sum_{s=1}^{S} p_s(X_i)}{S} \geq \left(\Pi_{s=1}^{S} p_s(X_i)\right)^{\frac{1}{S}} \tag{34}$$

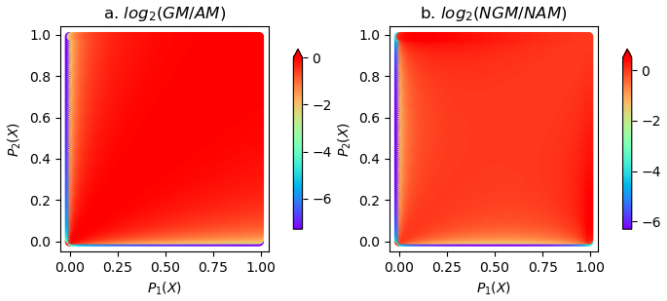


Fig. 23: GM versus AM

The equality holds if and only if $p(x_1)$=$p(x_2)$=$p(x_3)$=$\cdots$=$p(x_S)$. More specifically, if the difference among $p_1(X_i)$, $p_2(X_i)$, $\cdots$, $p_S(X_i)$ is not significant, the right side (GM) and the left side (AM) of (34) are very close. Otherwise, the difference among $p_1(X_i)$, $p_2(X_i)$, $\cdots$, $p_S(X_i)$ will enlarge the gap between GM and AM.

In WiFi-Visual localization, the right side the used to to enhance the consistent results and weaken divergent results. One notes that

$$\max_i \ (\Pi_{s=1}^S p_s(X_i))^{\frac{1}{S}} \iff \max_i \ \Pi_{s=1}^S p_s(X_i) \qquad (35)$$

Computing the product of the right side of (35) is enough for taking the maximum.

We take $S = 2$ as an example to illustrate its physical meaning in a classification problem. Suppose that $S = 2$ methods are independently used to solve a classification problem. Method 1 gives a likelihood as follows:

$$[p_1(X_1), p_1(X_2), \cdots, p_1(X_N)]$$

while method 2 gives a likelihood vector as

$$[p_2(X_1), p_2(X_2), \cdots, p_2(X_N)]$$

For $p_1(X = X_i)$ and $p_2(X = X_i)$, their GM and AM are studied by numerical simulations. Without loss of generality, we set $i = 1$. An indicator $log_2(\frac{GM}{AM})$ is defined to make a heatmap in a two-dimensional $(p_1(X = X_1),\ p_2(X = X_1))$ plane, as shown by Fig.23.

Fig.23a shows that the ratio of $\frac{GM}{AM}$ is usually between ¼ and 1 except the boundary area. In the blue-purple boundary area of Fig.23a, one probability is very weak and the other takes a normal value, the ratio $\frac{GM}{AM}$ can be $\frac{1}{64}$ or even smaller. We have similar conclusions for the normalized GM (NGM) and normalized AM (NAM), as shown by Fig.23b.

The boundary area in Fig.23 is corresponding to the case of WiFi-image localization with visual ambiguity, where WiFi localization (method 1) is not coherent with the image localization (method 2). In this case, GM is more efficient to weaken the controversial candidate.The localization performance comparison is provided by Fig.24, which shows the advantage of GM in enhancing the consistent results and weakening the divergent ones.

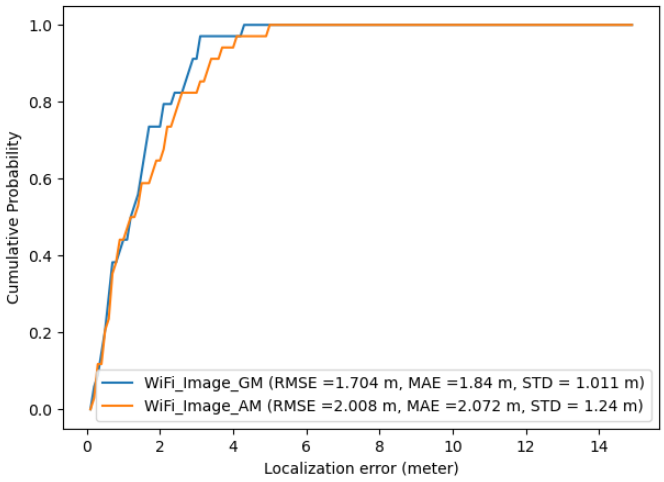


Fig. 24: GM fusion versus AM fusion

ACKNOWLEDGMENT

The authors would like to thank Ilinykh Arthur for his participation in our project as an engineering student by programming the robot to take photo sequence.

REFERENCES

[1] T. Umetani, Y. Kondo, and T. Tokuda, "Rapid development of a mobile robot for the Nakanoshima Challenge using a robot for intelligent environments," *Journal of Robotics and Mechatronics*, vol. 32, no. 6, pp. 1211-1218, 2020.

[2] T. Lee, C. Kim and D. Cho, "A Monocular Vision Sensor-Based Efficient SLAM Method for Indoor Service Robots," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 1, pp. 318-328, Jan. 2019.

[3] N. Radwan, A. Valada, W. Burgard, VLocNet++: Deep Multitask Learning For Semantic Visual Localization And Odometry, *IEEE Robotics And Automation Letters (RA-L)*, 3(4):4407-4414, 2018.

[4] A. Valada, N. Radwan, W. Burgard, Deep Auxiliary Learning For Visual Localization And Odometry, *Proceedings Of The IEEE International Conference On Robotics And Automation*, Brisbane, Australia, 2018.

[5] S. -H. Bach, P. -B. Khoi and S. -Y. Yi, "Global UWB System: A High-Accuracy Mobile Robot Localization System With Tightly Coupled Integration," *IEEE Internet of Things Journal*, vol. 11, no. 9, pp. 16618-16626, 1 May, 2024.

[6] A. Venon, Y. Dupuis, P. Vasseur and P. Merriaux, "Millimeter Wave FMCW RADARs for Perception, Recognition and Localization in Automotive Applications: A Survey," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 533-555, Sept. 2022.

[7] J. Jun, L. He, Y. Gu, W. Jiang, G. Kushwaha, A. Vipin, L. Cheng, C. Liu, and T. Zhu, "Low-overhead wifi fingerprinting," *IEEE Transactions on Mobile Computing*, vol. 17, no. 3, pp. 590–603, 2017.

[8] 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.

[9] R. Ayyalasomayajula, A. Arun et al. "Deep learning based wireless localization for indoor navigation," *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020.

[10] A. Arun, R. Ayyalasomayajula, W. Hunter and D. Bharadia, "P2SLAM: Bearing Based WiFi SLAM for Indoor Robots," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3326-3333, April 2022.

[11] H. Chen, Y. Zhang, W. Li, X. Tao, and P. Zhang, "ConFi: Convolutional neural networks based indoor Wi-Fi localization using channel state information," *IEEE Access*, vol. 5, pp. 18066–18074, Sep. 2017.

[12] V.Vauchey, Y. Dupuis, P. Merriaux, X. Savatier, "Particle filter meets hybrid octrees: an octree-based ground vehicle localization approach without learning," *Applied Intelligence*, 7 Avril, 2023.

[13] M. T. Hoang, B. Yuen, X. Dong, T. Lu, R. Westendorp and K. Reddy, "Recurrent Neural Networks for Accurate RSSI Indoor Localization," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10639-10651, Dec. 2019.

[14] L. Zhang et al., "WiFi-Based Indoor Robot Positioning Using Deep Fuzzy Forests," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 10773-10781, Nov. 2020.
[15] I. El Bouazzaoui, S. A. R. Florez, and A. El Ouardi, "Enhancing rgb-d slam performances considering sensor specifications for indoor localization," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 4970–4977, 2021.
[16] Y. Zhao, J. Xu, J. Wu, J. Hao, and H. Qian, "Enhancing camera-based multimodal indoor localization with device-free movement measurement using wifi," *IEEE Internet of Things Journal*, vol. 7, no. 2, pp. 1024–1038, 2019.
[17] M. D. Redzic, C. Laoudias, and I. Kyriakides, "Image and wlan bimodal integration for indoor user localization," *IEEE Transactions on Mobile Computing*, vol. 19, no. 5, pp. 1109–1122, 2019
[18] G. Huang, Z. Hu, J. Wu, H. Xiao, and F. Zhang, "Wifi and visionintegrated fingerprint for smartphone-based self-localization in public indoor scenes," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 6748– 6761, 2020.
[19] Y. A. Almansoub, M. Zhong, Z. Hu, G. Huang, M. A. Al-qaness, and A. A. Abbasi, "Multi-scale vehicle localization in underground parking lots by integration of dead reckoning, wi-fi and vision," *2020 6th International Conference on Big Data Computing and Communications (BIGCOM)*, pp. 41–49, 2020.
[20] C. Tang, W. Sun, X. Zhang, J. Zheng, W. Wu and J. Sun, "A Novel Fingerprint Positioning Method Applying Vision-Based Definition for WIFI-Based Localization," *IEEE Sensors Journal*, vol. 23, no. 14, pp. 16092-16106, 15 July15, 2023.
[21] C. Tang, W. Sun, X. Zhang, J. Zheng, J. Sun and C. Liu, "A Sequential-Multi-Decision Scheme for WiFi Localization Using Vision-Based Refinement," *IEEE Transactions on Mobile Computing*, vol. 23, no. 3, pp. 2321-2336, March 2024.
[22] D. Tse, V. Pramod, Fundamentals of wireless communication. Cambridge university press, 2005.
[23] Z. Wei et al., "Integrated Sensing and Communication Channel Modeling: A Survey," *IEEE Internet of Things Journal*, doi: 10.1109/JIOT.2024.3449377. Early Access, 2025.
[24] M. Botta, M. Simek. "Adaptive Distance Estimation Based on RSSI in 802.15.4 Network," *Radio engineering*, 2013, 22, pp.1162–1168.
[25] Zawar Shah and R. A. Malaney, "Particle Filters and Position Tracking in Wi-Fi Networks," 2006 IEEE 63rd Vehicular Technology Conference, Melbourne, VIC, Australia, 2006, pp. 613-617.
[26] J. Hagenauer and P. Hoeher, "A Viterbi algorithm with soft-decision outputs and its applications," *1989 IEEE Global Telecommunications Conference and Exhibition 'Communications Technology for the 1990s and Beyond'*, Dallas, TX, USA, 1989.
[27] https://pal-robotics.com/
[28] K. He, X. Zhang, S. Ren, et al. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016. p. 770-778.
[29] Y. Lecun, L. Bottou, Y. Bengio et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998, vol. 86, no 11, p. 2278-2324.
[30] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural computation*, 1997, vol. 9, no 8, p. 1735-1780.
[31] S. Boyd, L. Vandenberghe, Convex optimization, 2004, Cambridge university press.