INDOOR LOCALIZATION USING TRANSFORMER ENSEMBLE REGRESSION AND LED SIGNALS

Huy Q. Tran[1]*, Huy Le-Quoc[2]

[1]*Robotics and Mechatronics Research Group, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam*
[2]*The University of Danang - University of Science and Technology, Vietnam*

*Corresponding author: tqhuy@ntt.edu.vn

(Received: July 02, 2025; Revised: January 05, 2026; Accepted: January 22, 2026)

DOI: 10.31130/ud-jst.2026.24(1).356E

Abstract - Improving the localization accuracy under the influence of multipath reflections and signal interference in indoor localization systems using visible light signals remains a complex and challenging task. In this study, we propose utilizing the transformer architecture in a localization system based on light intensity signals of 16 LEDs. By leveraging the self-attention mechanism, the model can detect and focus on the locations with the highest relevance. The predictions are aggregated thanks to an ensemble strategy. Simulation results show that the proposed method achieves a Root Mean Square Error of approximately 0.334 m for the entire room (5x5 m), 0.14 m for the central region (3×3 m), and 0.39 m for the remaining areas near the walls and corners.

Key words - Localization; ensemble learning; transformer encoder; LED

1. Introduction

Previously, indoor localization technologies using Wi-Fi [1], [2], Bluetooth [3], 5G [4], etc., have been widely studied and applied, ranging from smart homes to modern factories. Each technology has its own advantages related to coverage range, localization accuracy, cost, or energy consumption. However, these solutions also face difficulties related to electromagnetic interference, accuracy degradation in complex environments, or infrastructure limitations. To reduce design costs, in the study [5], the authors used Wi-Fi signal strength (RSS) to locate and navigate a robot indoors. This system helps to determine the robot's position in real-time by analyzing the RSS values from multiple access points.

In addition to using signals independently, some studies have explored the combination of signals together. For example, the combination of Wi-Fi and Bluetooth (BLE) can improve the accuracy of indoor localization systems. In [6], the authors proposed an integrated approach that employs Kalman filtering, the K-nearest neighbor (KNN) classifier, and a recurrent neural network (RNN) to reduce noise in the RSS signals. At the same time, this study also models the complex relationship between signal strength and physical location. From there, it allows for high -accuracy location estimation. In another approach, some researchers replaced the traditional localization range with public Wi-Fi signal data. This reduces dependence on static infrastructure by creating a wireless map from RSS data collected by smartphones and Wi-Fi-enabled robots [7]. Recent studies have attempted to address the issue of localization without the need for a transmitter. Researchers have used only passive infrared sensors instead of conventional wireless signals [8]. That study proposed a deep learning-based approach that combines channel separation and template matching. Specifically, the model combines Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). The results show that this solution achieves high localization accuracy, with an average error of only 0.55 m, and 80% of the errors are within 0.8 m. Similarly, to develop lightweight, cost-effective, and high-accuracy systems, other studies have tested the integration of infrared sensors and single-chip radars for robot localization [9]. In that study, the authors combined Doppler radar-based motion modeling and a hybrid sensor fusion technique into a Monte Carlo localization framework. This method enables real-time localization with impressive accuracy. Therefore, it is considered an effective alternative to traditional LiDAR or camera-based systems.

As an alternative to RF-based localization, Visible Light Communication (VLC) has emerged as a promising solution by utilizing existing LED lighting infrastructure to support high-speed data transmission and accurate indoor localization. VLC technology uses the existing LED lighting system to transmit data, providing many advantages such as wide bandwidth, energy saving, and no electromagnetic interference. At the same time, visible light localization (VLL) technology also allows for accurately determining device location in spaces where traditional GPS signals are difficult to operate effectively, such as hospitals, warehouses, shopping malls, or smart home systems. The combination of VLC and VLL, namely VLCL, has been one of the potential trends. VLCL is increasingly being integrated into Industry 4.0 and Internet of Things (IoT) environments. These technologies support human-centric systems by enabling accurate tracking for operators interacting with augmented reality, virtual reality, or collaborative robots. This is particularly crucial in smart factories, where real-time localization enhances human-machine collaboration and operational efficiency while overcoming challenges posed by electromagnetic interference common in industrial settings [10].

However, current VLC-based localization systems still have difficulty maintaining accuracy when the receiver is tilted randomly, which often occurs in practice. To address this issue, a new study has developed an integrated visible light localization and communication system that simultaneously provides real-time, precise localization and

high-speed data transmission. In particular, an advanced differential phase difference of arrival algorithm was developed to simplify the hardware design and increase stability when the device moves. Experiments showed that the VLCL system performs better than existing solutions and maintains high accuracy even when the receiver is tilted in real-world scenarios such as two-dimensional and three-dimensional space [11].

To perform LED-based localization tasks, the authors in the study [12] employed an optimal optical omnidirectional angle estimator combined with a novel three-dimensional (3D) angle-of-arrival localization algorithm to reduce localization error and computational cost. Additionally, two supplementary methods were proposed to enhance accuracy when more than two white LEDs are used, with experimental results demonstrating centimeter-level precision. To improve localization accuracy, some studies have focused on strengthening both the transmitter and receiver; for example, study [13] proposed a weighted least squares algorithm combined with optimal calibration of LED tilt and gain, significantly reducing localization errors compared to traditional Gaussian Processes and multi-lateration methods. Meanwhile, to perform the localization process without installing complex infrastructure, the Spectral-Loc solution is proposed [14]. In that study, the author used variations in the spectral distribution of the received light at different locations. This solution achieves superior accuracy compared to conventional light intensity sensors.

Recently, deep learning models have played a significant role in improving the localization capabilities of LED localization systems. This solution makes the system more accurate and flexible than traditional methods. Research indicates that integrating advanced AI techniques with high-quality datasets and suitable evaluation methods can substantially reduce localization errors in smart cities and IoT applications [15, 16]. In addition to traditional optical sensors, the study [17] proposed a localization system that uses only a single PTZ camera. At the same time, the authors used a CNN model trained on synthetic images to detect an LED marker as an ellipse. This approach enables accurate 3D localization in both indoor and outdoor environments, without the need for GPS. Moreover, instead of using an LED array, the author proposed using a single photodiode and deep learning [18]. As a result, the proposed model achieved sub-meter accuracy thanks to transformer-based denoising, GNN-based error estimation, and PSO optimization. All the above solutions apply IoT technology at a low-cost.

This study proposes a novel approach based on the transformer architecture to enhance the accuracy of the indoor localization system in multipath reflection environments. Specifically:

- Applying the transformer architecture to indoor localization using LED light signals, this study represents pioneering efforts in leveraging attention-based models to capture the complex spatial relationships among RSS signals collected from multiple transmitters.
- Applying the self-attention mechanism, the model can dynamically identify the most relevant RSS inputs, thereby enhancing localization performance, especially in complex environments such as Non-Line-of-Sight (NLOS) areas.
- Training multiple transformer models with different initialization seeds and ensembling their predictions helps reduce random errors, improving the system's stability and generalization capability.

2. LED-based localization model and proposed method

This section is divided into two main parts to help readers better understand the proposed method. First, we present the indoor localization model using LED lights and the RSS data collection process. Next, we introduce the regression method using the ensemble-based transformer model, which is the core component of the proposed localization system.

2.1. LED-based localization model

This study is carried out in an enclosed space measuring $5 \times 5 \times 2.3$ meters. This represents a standard indoor room environment, as shown in Figure 1. For the lighting system, 16 LED lights are arranged at 1-meter intervals and placed 1 meter from the walls. This arrangement forms a uniform grid on the ceiling. This ensures consistent illumination across the space.

The walls of the room are painted with white paint, and the main door is made of frosted glass. The texture enhances the reflection of light and naturally creates multipath propagation. This is a common phenomenon in real indoor environments. Based on the texture, the average reflectance of the walls and the doors is approximately 0.8. The above setup ensures that the received signal at any point includes both direct and reflected light components. Since both types of paths are typically present in real-world environments, the collected data are used to evaluate the performance of the proposed transformer ensemble model.

Alongside the LED lighting system, a photodetector is employed to capture the variation in light intensity from each LED source. The sensor is placed close to the floor and can move freely across a predefined grid to collect data at various locations within the room. At each location, it measures the RSS values emitted from the overhead LED lights.

To ensure the photodetector can distinguish the signals from each LED, each light is modulated at a unique frequency. Upon receiving the light signal, the photodetector performs frequency-domain analysis to identify the corresponding source. The modulation frequencies used are as follows: 16 LEDs starting from 2 kHz, with each subsequent LED spaced 2 kHz apart. The LED lighting system is configured with a semi-angle at half power of 70°, corresponding to a Lambertian emission order m. Each LED operates at 3 watts of power. The receiver has a photodetector area of 10^{-4} m^2, with an optical filter gain of 1 and a concentrator gain determined by a refractive index of 1.5 and a field-of-view (FOV) of 60°.

Using different modulation frequencies helps the system clearly separate the light signals from each LED, even when they overlap, and reduces the impact of noise from the environment. As a result, the collected RSS data are easier to distinguish and can be used reliably as input for the proposed transformer ensemble-based Regression model.

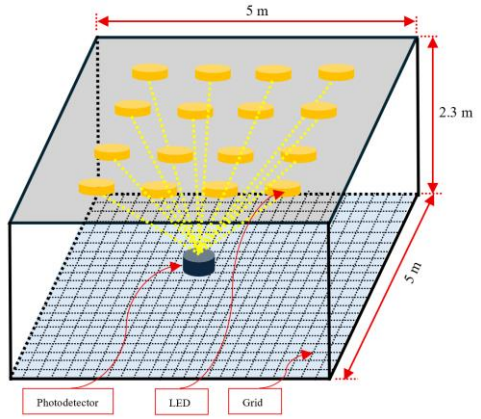

Figure 1. *LED-based localization model*

In VLL systems, the optical power received at the photodetector consists of two main components: the direct path signal (Line-of-Sight, denoted as P_{LOS}) and the indirect reflection signal (Non-Line-of-Sight, denoted as P_{NLOS}) from surrounding surfaces, especially from walls. To model these components, mathematical expressions are used to capture geometric and physical factors such as distance, projection angle, surface reflectivity, and the optical characteristics of the receiver.

The P_{LOS} component represents the direct optical power transmitted from the LED to the sensor without obstruction, while the P_{NLOS} component accounts for the first-order reflections of light from walls or doors before reaching the receiver. These two received power models are formally expressed in equations (1) and (2) [19].

$$P_{LOS} = \begin{cases} P_s \frac{(m+1)A_{rx}}{2\pi d^2} cos^m(\phi) T_f(\psi) G(\psi) cos(\psi), if\ 0 \le \psi \le \psi_{max} \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad ,if\ \psi > \psi_{max} \end{cases} \tag{1}$$

where:

P_s: Transmitted optical power from the LED source.

A_{rx}: Physical area of the photodetector (receiver).

m: Lambertian order of emission, which defines the directionality of the LED radiation pattern. It is calculated as $m = ln(2)/ln(cos(\Phi_{1/2}))$, where $\Phi_{1/2}$ is the LED's half-power angle.

d: Distance between the LED source and the receiver.

ϕ: Angle of irradiance.

ψ: Angle of incidence.

$T_f(\psi)$: Optical filter transmission gain, modeling the effect of a filter placed over the photodetector.

$G(\psi)$: Optical concentrator gain, which amplifies received power depending on the incidence angle.

$cos(\psi)$: Lambert's cosine law component representing the effective projection area of the photodetector.

ψ_{max}: Field-of-view (FoV) limit of the photodetector; outside this range, no signal is received.

$$P_{NLOS} = P_s \frac{(m+1)A_{rx}}{2\pi d_1^2 d_2^2} \rho cos^m(\phi) \cos(\alpha) \cos(\beta) A_{surf} \cos(\psi_r) T_f(\psi_r) G(\psi_r), \tag{2}$$

$for\ 0 \le \psi_r \le \psi_{max}$

where:

d_1: Distance from the LED source to the reflection point on the wall or surface.

d_2: Distance from the reflection point to the photodetector.

ρ: Reflection coefficient of the surface.

α: Angle of incidence at the reflecting surface (from the LED to the surface).

β: Angle of reflection from the surface toward the receiver.

A_{surf}: Reflecting surface area involved in the indirect light path.

ψ_r: The incidence angle of the light from the wall.

2.2. Transformer ensemble-based Regression Model

As mentioned in the previous section, in indoor localization systems based on LED lighting, the noise caused by multipath reflection significantly reduces the localization accuracy. These traditional localization algorithms rely on simple linear relationships. Therefore, they often struggle to estimate positions in the presence of significant noise. With its self-attention mechanism, the transformer algorithm can learn complex relationships between input features. Additionally, multiple independently trained transformer models are combined to enhance the stability and generalization capabilities of the entire system. To perform the training and position estimation process based on data obtained from the optical sensor, we configure the parameters of the proposed transformer ensemble model as described in detail in Table 1. This table presents information related to Model architecture, training configuration, ensemble strategy, data processing, and evaluation.

Table 1. *Hyperparameters and training configuration of the proposed transformer-based localization model*

Category	Parameter	Value
Model Architecture	Input dimension	16 RSS values
	Embedding dimension	64
	Number of attention heads	4
	Number of encoder layers	2
	Input projection layer	Linear (1 to 64)
	Feedforward regressor	128 to 2
Training Configuration	Optimizer	Adam
	Learning rate	1×10^{-3}
	Loss function	MSE
	Batch size	64
	Epochs	200
	Weight initialization	PyTorch default
Ensemble Strategy	Number of models	5
	Seeds used	0, 1, 2, 3, 4
	Ensemble method	Mean averaging
Data Processing	Feature scaling	Standardization
	Train–test split	80% – 20%
Evaluation	Metric	RMSE

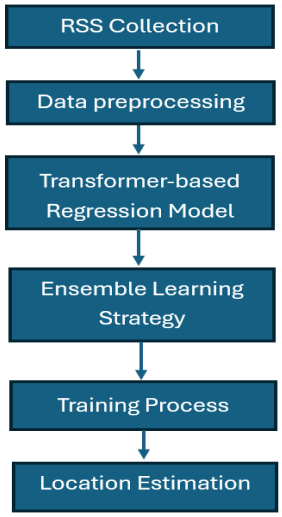

Figure 2. *Transformer ensemble-based regression model*

The overall procedure involves the main steps: data preprocessing and preparation, model definition, ensemble inference, training, and evaluation. The details of each step are depicted in Figure 2 and presented as follows:

• Data collection and preprocessing

The floor is divided into a grid of size 25×25, creating 625 reference points. At each grid point, RSS values from sixteen LEDs are recorded under NLOS conditions. These measurements form the input vector of received optical power as:

$$X = [P_1,\ P_2,\ P_3, \dots,\ P_{16}] \tag{3}$$

where P_i is the optical power received from the i^{th} LED. The corresponding output is the true location of the photodetector:

$$Y = [x_r,\ y_r] \tag{4}$$

Prior to model training, both the input features and output labels are normalized using standardization:

$$x_{norm} = \frac{x - \mu_x}{\sigma_x}, y_{norm} = \frac{y - \mu_y}{\sigma_y} \tag{5}$$

This normalization process improves model convergence and ensures that all features are on a comparable scale. The data are then partitioned into training and testing sets using an 80:20 ratio.

• Transformer-based Regression Architecture

The proposed regressor utilizes a transformer encoder architecture to model the nonlinear spatial relationship between RSS signals and physical location. Each scalar input P_i is first projected to a higher-dimensional space via a linear transformation [20]:

$$P_i' = Linear(P_i) \in R^{dmodel} \tag{6}$$

The resulting sequence $[P_1',\ P_2',\ P_3', \dots, P_{16}']$ is then passed through multiple layers of the transformer encoder. The self-attention mechanism within the encoder allows the model to learn contextual dependencies between the input RSS signals, which may carry correlated spatial information due to multipath effects.

The encoder output is flattened and passed to a fully connected feedforward network to produce the estimated 2D coordinates [21]:

$$y' = f(flatten(Transformer(x))) \tag{7}$$

• Ensemble Learning Strategy

To improve the robustness and generalization of the model, we adopt an ensemble learning approach. Specifically, the transformer model is trained five times independently using different random seeds. Each training instance follows the same hyperparameters and optimization procedure.

At inference time, the predictions from all five models are averaged element-wise to form the final output:

$$y'_{final} = \frac{1}{n}\sum_{j=0}^{n} y'^{(j)} \tag{8}$$

where $y'^{(j)}$ is the prediction from the j^{th} model.

• Training and Loss Function

All models are trained using the Adam optimizer with a learning rate of 1×10^{-3}. The loss function is defined as the mean squared error (MSE) between predicted and ground truth positions:

$$L = \frac{1}{N}\sum_{i=1}^{N} ||y'_i - y_i||^2 \tag{9}$$

Training is conducted for 200 epochs using a mini-batch size of 64. All models are implemented in PyTorch and trained on a standard computing platform.

• Evaluation Metrics and Results

After ensemble inference, predictions are inverse-transformed to return to the original coordinate system. The localization accuracy is evaluated using Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} ||y'_i - y_i||^2} \tag{10}$$

3. Results and Discussion

To evaluate the effectiveness of the proposed transformer ensemble approach, we compare errors in predicted positions with the actual positions for four different algorithms: LSTM, linear regression, gradient boosting, and the proposed transformer regression method, as illustrated in Figure 3 to Figure 6.

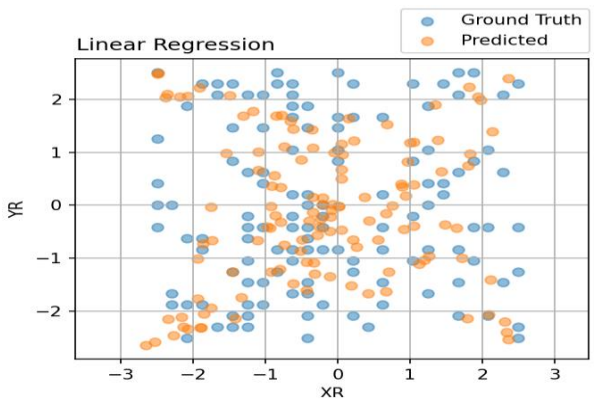

Figure 3. *Simulation results with Linear Regression*

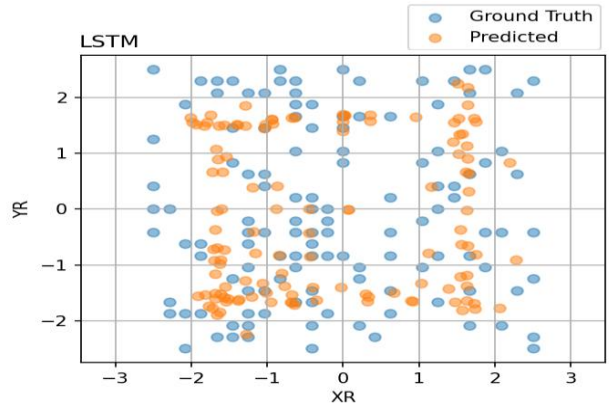

Figure 4. *Simulation results with the LSTM algorithm*

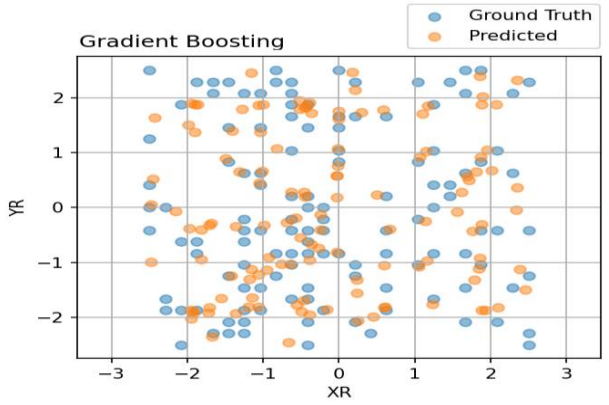

Figure 5. *Simulation results with the gradient boosting algorithm*

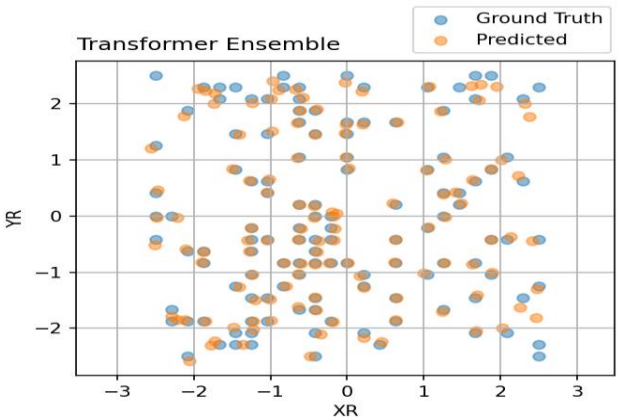

Figure 6. *Simulation results with Transformer ensemble-based Regression Model*

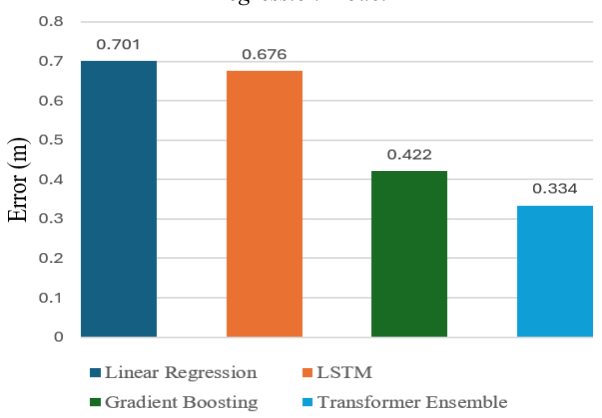

Figure 7. *RMSE comparison*

In the Linear Regression model, the predicted results deviate significantly from the ground truth, especially at the edges. The LSTM method improves the localization results in the middle region but incorrectly predicts the results in the edge region. In the third approach using the gradient boosting algorithm, the localization error is significantly reduced compared to the previous two methods; however, it remains 21% higher than that of the transformer ensemble method. Thus, our proposed model has the best ability to learn spatial features

Comparing the four algorithms, the proposed solution, namely the transformer ensemble model, outperforms traditional methods in accuracy. Specifically, the RMSE of linear regression is 0.701 m (Figure 3), while LSTM is 0.676 m (Figure 4), and gradient boosting is 0.422 m (Figure 5). Meanwhile, the transformer ensemble only reaches 0.334 m (Figure 6), 52.3% lower than linear regression and 50.6% lower than LSTM, and 20.9% lower than gradient boosting, as presented in Figure 7.

These numbers show that the transformer ensemble reduces the localization error by half compared to popular machine learning models, thanks to its ability to automatically identify actual signals in noisy environments and multipath reflections. This clearly demonstrates the potential of attention architectures in indoor localization applications using visible light.

In addition to analyzing the localization results throughout the entire room, the analysis and comparison of localization errors at the center of the room (3x3 m) and the edge area near the four walls and corners were also conducted, as shown in Table 2. The results show that in the center area, the transformer ensemble algorithm also achieved the highest accuracy, with a value of 0.14 m. In comparison, the edge area had a value of 0.39 m. Other algorithms also demonstrated better accuracy in the central area than in the peripheral area. Specifically, in the center area, Gradient boosting, LSTM, and Linear regression achieved errors of 0.39 m, 0.57 m, and 0.62 m, respectively. Meanwhile, the errors at the edge of those algorithms were 0.49 m, 0.72 m, and 0.76 m, respectively. Thus, in all cases, the transformer ensemble learning algorithm also achieved the lowest localization error.

Table 2. *Localization accuracy in central and rim regions*

Algorithm	RMSE in central region	RMSE in rim region
Transformer ensemble	0.14	0.39
Gradient boosting	0.39	0.49
LSTM	0.57	0.72
Linear regression	0.62	0.76

4. Conclusion

In this study, we proposed a method that combines the transformer and ensemble approaches to address the challenges caused by multipath reflection and signal interference. Thanks to the self-attention mechanism, the model automatically selected and focused on the most critical RSS signals. Simulation results demonstrated that the model significantly reduces the localization error in

both the central region and the surrounding areas of walls and corners. Results showed that the method achieves a root mean square error (RMSE) of approximately 0.14 m in the central region (3 x 3 m) and 0.39 m in the rim regions. Therefore, our proposed approach achieved an overall error of 0.334 m.

The future research direction will focus on deployment in real environments to verify practical applicability and improve the attention algorithm to increase accuracy and adaptability to more complex situations.

Acknowledgement: We acknowledge Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam, for supporting this study.

REFERENCES

[1] N. S. Ahmad, "Recent advances in WSN-based indoor localization: A systematic review of emerging technologies, methods, challenges, and trends," *IEEE Access*, vol. 12, pp. 180674–180714, 2024. DOI: 10.1109/ACCESS.2024.3509516.

[2] S. Wang, S. Zhang, J. Ma, and O. A. Dobre, "Graph neural network-based WiFi indoor localization system," in *Proc. IEEE Global Communications Conf. (GLOBECOM),* Cape Town, South Africa, 2024, pp. 116–120. DOI: 10.1109/GLOBECOM52923.2024.10901150.

[3] T. Shi and W. Gong, "A survey of Bluetooth indoor localization," in *Proc. 2024 IEEE 10th Conf. on Big Data Security on Cloud (BigDataSecurity),* New York City, NY, USA, 2024, pp. 71–77. DOI: 10.1109/BigDataSecurity62737.2024.00020.

[4] X. Zhou, L. Chen, Y. Ruan and R. Chen, "Indoor Localization With Multi-Beam of 5G New Radio Signals," in *IEEE Transactions on Wireless Communications*, vol. 23, no. 9, pp. 11260-11275, Sept. 2024, doi: 10.1109/TWC.2024.3380737.

[5] S. A. Kharmeh, E. Natsheh, R. Nasrallah, and M. Masri, "Triangulation-enhanced WiFi-based autonomous localization and navigation system: A low-cost approach," in *Proc. 2024 22nd Int. Conf. on Research and Education in Mechatronics (REM),* Amman, Jordan, 2024, pp. 69–74. DOI: 10.1109/REM63063.2024.10735691.

[6] K. Beigi and H. Shah-Mansouri, "An intelligent indoor positioning algorithm based on Wi-Fi and Bluetooth Low Energy," in *Proc. 2024 IEEE Wireless Communications and Networking Conf. (WCNC)*, Dubai, United Arab Emirates, 2024, pp. 1–6. DOI: 10.1109/WCNC57260.2024.10570531.

[7] Z. Yi *et al*., "Multimodal indoor localization using crowdsourced radio maps," in *Proc. 2024 IEEE Int. Conf. on Robotics and Automation (ICRA),* Yokohama, Japan, 2024, pp. 13666–13672. DOI: 10.1109/ICRA57147.2024.10610683.

[8] S. Yongchareon, J. Yu, and J. Ma, "Efficient deep learning-based device-free indoor localization using passive infrared sensors," *Sensors*, vol. 25, no. 6, Art. no. 1362, 2025.

[9] D. Wang, M. Masannek, S. May, and A. Nüchter, "Infradar-localization: Single-chip infrared- and radar-based Monte Carlo localization," in *Proc. 2023 IEEE 19th Int. Conf. on Automation Science and Engineering (CASE),* Auckland, New Zealand, 2023, pp. 1–8. DOI: 10.1109/CASE56687.2023.10260572.

[10] L. Danys *et al.*, "Visible light communication and localization: A study on tracking solutions for Industry 4.0 and the Operator 4.0," *Journal of Manufacturing Systems*, vol. 64, pp. 535–545, 2022. [Online]. Available: https://doi.org/10.1016/j.jmsy.2022.07.011.

[11] H. Yang et al., "An advanced integrated visible light communication and localization system," *IEEE Transactions on Communications*, vol. 71, no. 12, pp. 7149–7162, Dec. 2023. DOI: 10.1109/TCOMM.2023.3309823.

[12] Y. Yu, B. Zhu, Z. Zhang, L. Wang, L. Wu, and J. Dang, "Indoor visible light localization algorithm with the optimal optical angle-of-arrival estimator," in *Proc. 2021 2nd Information Communication Technologies Conf. (ICTC)*, Nanjing, China, 2021, pp. 194–198. DOI: 10.1109/ICTC51749.2021.9441621.

[13] F. Wu, N. Stevens, L. De Strycker, and F. Rottenberg, "Enhancing RSS-based visible light positioning by optimal calibrating the LED tilt and gain," *arXiv preprint arXiv:2404.18650*, 2024. [Online]. Available: https://arxiv.org/abs/2404.18650

[14] Y. Wang, J. Hu, H. Jia, W. Hu, M. Hassan, A. Uddin, B. Kusy, and M. Youssef, "Spectral-Loc: Indoor localization using light spectral information," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.,* vol. 7, no. 1, Art. no. 37, pp. 1–26, Mar. 2023. DOI: 10.1145/3580878.

[15] O. Kerdjidj *et al*., "Uncovering the potential of indoor localization: Role of deep and transfer learning," *IEEE Access*, vol. 12, pp. 73980–74010, 2024. DOI: 10.1109/ACCESS.2024.3402997.

[16] I. Cappelli *et al.*, "Enhanced visible light localization based on machine learning and optimized fingerprinting in wireless sensor networks," **IEEE Transactions on Instrumentation and Measurement**, vol. 72, pp. 1–10, 2023, Art. no. 9503410. DOI: 10.1109/TIM.2023.3240220.

[17] X. Oh, R. Lim, S. Foong, and U.-X. Tan, "Marker-based localization system using an active PTZ camera and CNN-based ellipse detection," *IEEE/ASME Transactions on Mechatronics*, vol. 28, no. 4, pp. 1984–1992, Aug. 2023. DOI: 10.1109/TMECH.2023.3274363.

[18] X. Cao, Y. Zhuang, X. Wang, T. Yu, J. Zhou, and J. Jiang, "Deep-learning-enhanced visible light positioning system based on the LED array," *IEEE Internet of Things Journal*, vol. 11, no. 12, pp. 21985–21995, Jun. 15, 2024. DOI: 10.1109/JIOT.2024.3377506.

[19] Z. Ghassemlooy, W. Popoola, and S. Rajbhandari, *Optical Wireless Communications: System and Channel Modeling with MATLAB*. Boca Raton, FL, USA: CRC Press, 2012.

[20] A. Vaswani *et al.*, "Attention Is All You Need," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017. DOI: 10.48550/arXiv.1706.03762.

[21] Z. Zhang, R. Tian, and Z. Ding, "TrEP: Transformer-Based Evidential Prediction for Pedestrian Intention with Uncertainty," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, pp. 3534–3542, 2023. DOI: 10.1609/aaai.v37i3.25463.