2022 International Conference on Indoor Positioning and Indoor Navigation (IPIN), 5 - 7 Sep. 2022, Beijing, China

Leveraging Transfer Learning for Robust Multimodal Positioning Systems using Smartphone Multi-sensor Data

Xijia Wei
UCL Interaction Centre
University College London
xijia.wei.21@ucl.ac.uk

Valentin Radu
Department of Computer Science
University of Sheffield
valentin.radu@sheffield.ac.uk

***Abstract*—Indoor positioning has been widely researched in recent years due to its high demand for developing localization services and its complexity in GPS-denied environments. However, the diversity of indoor spaces and temporal variation of local conditions impose the need for building specific and periodic calibrations at high cost for deployment and maintenance of these localization systems. A robust positioning solution that overcomes these challenges is yet to be available. Previous systems achieve good performance when specializing their solution to the unique characteristics of the deployment site. The drive is now to automatically model these localization solutions on the sensor data from each site with the least amount of effort. We propose to accelerate the model adaptation to new deployment sites by using transfer learning of a multimodal deep neural network architecture. We demonstrate that the required training data is drastically reduced compared to training the model from scratch, while also boosting its accuracy, due to the additional knowledge from pretraining on other sites. The resulting model is also fault-tolerant, showing good performance in missing modalities experiment. Our research opens the way toward scalable and cost efficient localization systems.**

***Index Terms*—AI-based Positioning, Multi-sensor Systems, Multimodal Machine Learning, Transfer Learning**

I. Introduction

The increasing adoption of wearable and mobile devices is improving the way we interact with the world, owing to their advanced sensing capability. The Global Positioning System (GPS) on mobile devices has been widely used in various outdoor scenarios to provide information and navigation for our current geo-spatial digital world. However, the GPS is unreliable inside buildings and underground due to GPS satellite signals not reaching many such environments. Alternative indoor locatization methods have been explored, relying on the smartphone built-in sensors, such as WiFi received signal strength, magnetic field intensity, and inertial sensors (e.g., accelerometer, gyroscope) [1]. Despite this effort, a reliable and scalable indoor localization solution is not yet available.

Conventional engineering based localization solutions mainly include Pedestrian Dead Reckoning (PDR) and WiFi Fingerprinting [2]. Both of these approaches use custom methods with precise mathematical formulations for processing the sensor signals that are specific to subsets of deployment sites. However, these over-engineered solutions often perform suboptimal when deployed to new scenarios. This is due to condition changes away from the lab settings, and varying sensor sensibility between the devices used when engineering the solution and those used in deployment. All these aspects contributes to low system robustness across deployments. As a result, human intervention for periodic calibration is essential for maintaining the accurate functioning of these systems, which makes wide adoption currently unattainable.

In the age of big data, we believe that relying on data for an end-to-end machine learning approach is the only promising solution for robust indoor localization, instead of conventional over-engineered solutions. Many Machine Learning (ML) based solutions have already been proposed for indoor localization [3], however, these still hold a narrow focus specialising only on a small set of deployment sites. We consider that ML approaches should have the ability to carry the learnt knowledge across multiple deployment sites for increased robustness and fast deployment. This is achievable by reducing the amount of training data that is required from each new deployment site.

Inspired by the success of multimodal machine learning in many modality-fusion tasks and the effectiveness of using transfer learning to strengthen machine learning models [4], in this work, we propose an end-to-end hybrid multimodal architecture integrated with transferable sub-components. Our Model Transferable Localisation system (MTLoc) is formed of independent components operating on distinct smartphone sensor data, joined by a final stage of knowledge fusion to produce the location estimation.

We classify the sensor data into two types: the infrastructure-free sensing modality (IMU data) and the infrastructure-based modalities (magnetic and WiFi RSS scans). For processing infrastructure-free samples, we pretrain a Long Short-Term Memory (LSTM) model, IMU-LSTM, as a feature extractor, part of the MTLoc architecture. This is then refined using transfer learning. For extracting infrastructure-based features, we construct another LSTM model, MAG-LSTM, and a deep neural network (DNN) to extract multi-

978-1-7281-6218-8/22/$31.00 © 2022 IEEE

sensor features. All extracted modality-specific features are then joined in a one-dimensional vector, followed by additional multi-layer perceptrons to produce the final location estimation. The transferable component of the IMU-LSTM is pretrained on the source site data (lab conditions). After pretraining, we integrate this model in the MTLoc architecture to bring the learnt infrastructure-free representations to the target deployment.

For our evaluation, we collect two multimodal datasets from two indoor scenarios. Both datasets contain time-sequential IMU sensors and magnetic samples as well as WiFi Received Signal Strength (RSS) fingerprints, along with ground truth location annotations. We explore the impact of data volume during the fine-tuning stage of our MTLoc by varying the amount of training data available from the target domain (deployment site). To evaluate the robustness of the fine-tuned model we corrupt valid multi-sensor samples. In this process, we find that the MTLoc can tolerate missing data of one or more modalities from the target site, being bootstrapped with just a small number of samples. MTLoc predicts the trajectory with good fidelity, over 80% of the estimations being within 3 meters of error. Benefiting from transferred knowledge, MTLoc fine-tuned with a small amount of data shows robustness compared to training the model entirely on the target site data from scratch.

The contributions of this work are as follows:

- We introduce transfer learning to multimodal machine learning based location estimation. The model shares the knowledge learnt in the infrastructure-free modality across deployment sites.
- We offer insights into the best options to fine tune the multimodal neural network with a small amount of data from a target deployment site. With less than a quarter of the available training data for a new deployment site, the model achieves a strong performance, median estimation error being within 1.56 metres. This is actually better than training from scratch on the full amount of data, without transfer learning (2.39 metres median error).
- The method we propose here is also evaluated for robustness to noisy and missing modality data. We show it can handle 40% of the modality variation. The model is bootstrapped with a small amount of data to achieve a prediction mean error of just 2.92 metres.

II. Motivation and Related Work

Due to the poor reliability of GPS in indoor environments, many solutions have been proposed for tracking subjects and devices based on alternative signal sources such as WiFi received signal strength (RSS), Bluetooth, magnetic field and inertial movement unit (IMU). However, conventional engineering-based solutions such as pedestrian dead reckoning (PDR) and WiFi Fingerprinting are often designed with a target building in mind. However, for deployment on new sites and adapting to indoor environment changes requires costly refinement to cope with new environment characteristics, which is prohibitive.

In recent years, machine learning based positioning systems have become a research hotspot pursuing data-driven localization approaches for minimal human intervention and deployment cost [5]. Current solutions mainly exploit single-modality data (single sensor), which often depend on the indoor infrastructures. For instance, a WiFi Fingerprinting based positioning system fails to work when WiFi signals are absent. Hence, the robustness of a single modality based localization system is greatly impacted by deployment conditions.

Multimodal machine learning has been investigated in multiple modality tasks such as audio-visual speech recognition [6]. It has the advantage of modeling complementary modalities and strengthening essential features that would otherwise be hard to spot in single modality settings. Inspired by the success of multimodal machine learning, we explored a multimodal hybrid deep neural network for location tracking in our previous work [7]. In that work we proposed a purely data-driven end-to-end robust localization approach by utilising multi-sensor inputs (IMU, magnetic and WiFi RSS data).

Fundamentally, each building has its unique radio and magnetic propagation characteristics, due to building materials, furniture and occupancy. This unique fingerprinting allows an association between signals and locations. However, some characteristics are shared across multiple buildings and deployment environments, which can be learned and transferred across sites. Hence, in this work, we aim to answer the following questions: *i) How to construct an architecture that is robust to deployment site variability? ii) How much effort of data collection is reduced by using transfer learning and fine-tuning? iii) How robust is our solution to sensor data alteration at inference stage?*

Transfer learning has been applied to many AI problems with often good performance, such as migrating the learnt vision recognition or natural language processing ability from a large trained model to a new deployed model for processing tasks in the target settings [8]. However, to date, this technical solution has not been validated for accelerating the deployment of localization systems. To address the aforementioned issues of data scarcity in new deployments and reusing knowledge across sites, we highlight transfer learning as a viable and effective solution for creating robust localization systems with reduced deployment cost.

A. Engineering-based Positioning

Engineering-based indoor positioning systems commonly rely on two approaches, Pedestrian Dead Reckoning (PDR) [9] and WiFi Fingerprinting [10], which work on a set of handcrafted formulations to identify the mobility frame including step counting, orientation estimation and fingerprints matching for localization [11]. Systems are usually designed to be building-specific. When indoor environment characteristics change or on new site deployment, system recalibration of the model is needed to fit these new distribution of data, resulting in additional tuning cost and lower efficiency.

B. AI-based Positioning

Instead of engineering-based solutions, artificial intelligence based positioning system shows its advantage in low deployment cost without requiring accurate mathematical equations, though moving the focus to the quality and quantity of training data itself [12]. For instance, HiMLoc integrates IMU sensors with WiFi RSS samplings through prior observations of Gaussian processes for direction estimation, distance estimation and correlation, and detected human activity [2]. CamLoc [13] uses computer vision to identify the tracking target, feet position and pedestrian skeleton for obtaining the location.

C. Transfer Learning-based Positioning

To date, transfer learning based localization is largely unexplored. There are a few examples based on single modality positioning systems with transfer learning techniques. One previous adoption of transfer learning is that of Pan et al. [14] for WiFi-based positioning to address the challenge of WiFi signal distribution variation, which changes across time and devices. Another implementation is that of Werner et al. [15] bringing transfer learning to a vision-based localization system that assists positioning by migrating the image recognition ability from deep convolutional neural networks to the system for identifying indoor symbolic targets. These examples evaluate the model performance under different working conditions such as time variations and device variations under the settings of the same building, but lack the evaluation of moving to new buildings as deployment sites.

III. Methodology

A. Multimodal Neural Network

The architecture of our proposed multimodal transfer learning model, the MTLoc, is shown in figure 1. Here, the network contains three parallel modality-specific sub components performing feature extraction from each modality input. These rely on LSTM networks, each operating on the IMU signal and on the magnetic field samplings respectively, and a DNN model extract features from WiFi fingerprints. All extracted modality-specific features are then joined in a one-dimensional vector, followed by additional multi-layer perceptrons to produce the joint location estimations.

Table I presents the structure of our MTLoc model. This contains 295,810 trainable parameters.

TABLE I
MTLoc Model Construction and Parameter Settings

Layer	Shape	Trainable Param
Input Layer.1 (WiFi)	750	0
FC Layer.1 (WiFi)	128	96128
FC Layer.2 (WiFi)	128	16512
Input Layer.2 (IMU)	10*2	0
IMU-LSTM (Transfer)	128	67072
Input Layer.3 (MAG)	10*1	0
MAG-LSTM (MAG)	128	66560
Fusion Layer (W/I/M)	384	0
FC Layer.3 (Fusion)	128	49280
FC Layer.4 (Fusion)	2	258

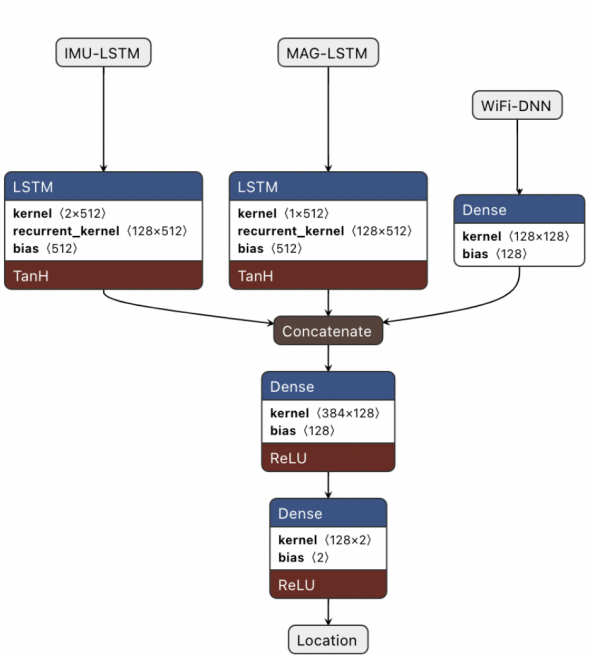


Fig. 1. MTLoc Model Architecture

B. Transfer Learning

Figure 2 illustrates the procedures for implementing transfer learning for a new deployment. By pretraining a model based on the IMU dataset from the source scenario, we derive an IMU-LSTM regression model. This sub network behaves as a transferable component, which holds the learnt IMU sampling features and representations from the previous scenario. When deploying the multimodal network to the target scenario, the trained IMU-LSTM sub network is transferred and integrated into the multimodal network. Here, the IMU-LSTM component is not trained further on new depoyment sites due to the general nature of locomotion across sites. Its weights and bias are frozen during the MTLoc model training process to extract new IMU sampling inputs feature based on learnt knowledge from the source site. The other two branches of the MAG-LSTM and the WiFi-DNN sub networks are trained from scratch to understand multimodal dataset representations from the new deployment. All model parameters, except the transferred model parameters, are updated during the gradient descent to allow the whole model to fit on the new scenario.

C. Fine Tuning

To allow the new deployed model to better fit on the characteristics of the new scenario, we implement fine tuning of the pre-trained model. The IMU-LSTM component behaves as a weights initializer which allows the model to update its weights and bias based on the transferred parameters. By setting the IMU-LSTM sub component parameter as trainable, all model parameters are updating during the training process based on the transferred knowledge.

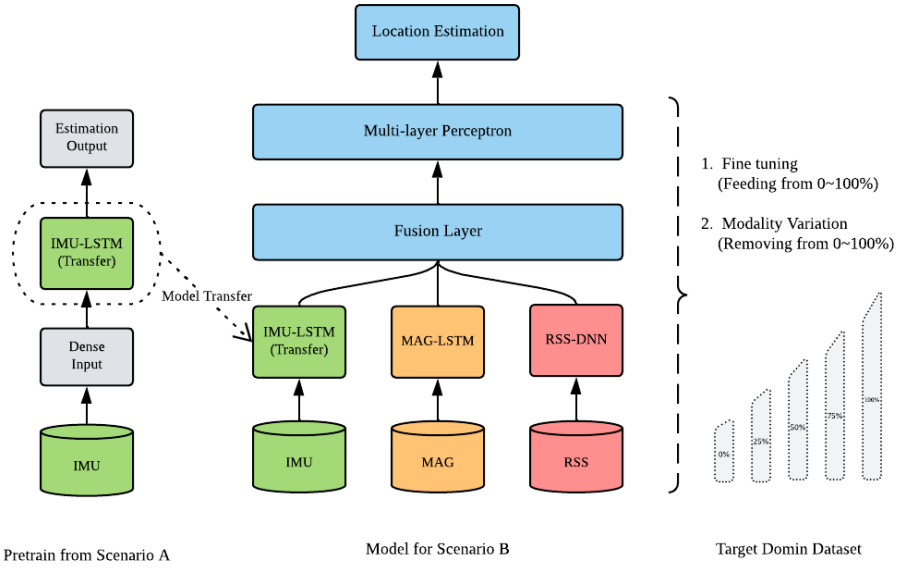


Fig. 2. Schematic representation of the components of our MTLoc, applying transfer learning and fine-tuning on the target data (deployment environment).

TABLE II
MULTIMODAL DATASET FORMAT

	Infrastructure-free		Infrastructure-based					Label	
Time	**Accelerator**	**Gyroscope**	**Magnetometer**	**AP0**	**AP1**	...	**APn**	**X**	**Y**
T0	a(*0∼999*)	g(*0∼999*)	m(*0∼999*)	null	-86	...	null	X*0*	Y*0*
T1	a(*999∼1999*)	g(*999∼1999*)	m(*999∼1999*)	null	null	...	null	X*1*	Y*1*
T2	a(*1999∼2999*)	g(*1999∼2999*)	m(*1999∼2999*)	-70	null	...	-65	X2	Y2
...	...	...	...	...	...	...	...	...	...
Tn	a(*n∼n+999*)	g(*n∼n+999*)	m(*n∼n+999*)	null	null	...	null	X*n*	Y*n*

TABLE III
MULTIMODAL DATASET STATISTICS

Dataset	IMU Samples	Mag Samples	RSS Samples	Access Points	Time
Source Scenario	24,450 * (10 * 2)	24,450 * (10 * 1)	25,541	102	407 Mins
Target Scenario	29,836 * (10 * 2)	29,836 * (10 * 2)	8,390	750	497 Mins

IV. EXPERIMENTS

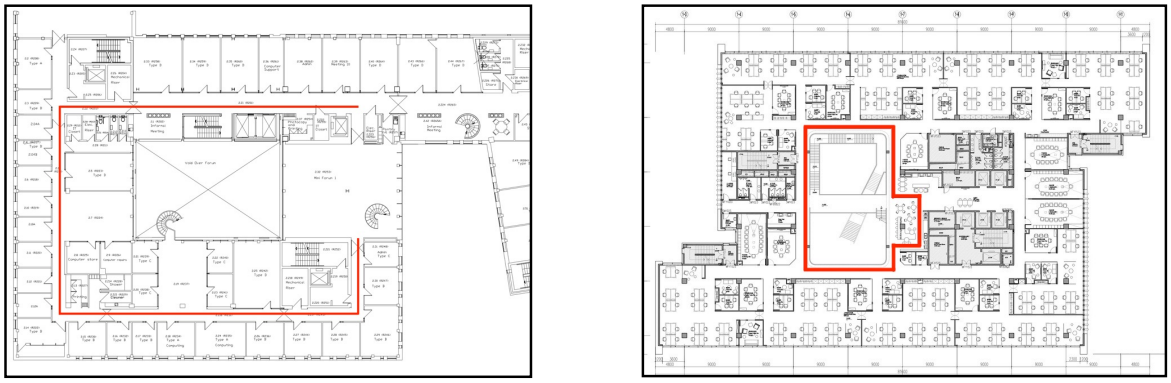
A. Data

For model training and evaluating, we use a multimodal dataset collected from two indoor scenarios (source and target) shown in Figure 3. Data from each scenario is collected by different persons using the OnePlus 7 and HUAWEI P40 smartphone respectively. Both datasets contain time-sequential IMU sensors and magnetic samplings as well as WiFi RSS fingerprints collected when walking along corridors with the ground truth location annotation. During the data collection process, multiple variations are included to increase dataset complexity and generalisation, including different time of day, walking postures and speeds. In addition, we keep the occasional short-term occurrence of WiFi hotspots to replicate real situations.

Table II presents the samples distribution collected from our two experiment sites. Specifically, the source scenario dataset contains 24,450 inertial measurement units (IMU) and magnetic sensor samples as well as a boosted number of WiFi samples, to 25,541 accessed from 102 access points mounted in the building. The target scenario dataset holds fewer WiFi samples of 8,390 sensed from 750 access points, and the IMU and magnetic sensors of 29,836 samples. As both datasets collected from the two scenarios contain 14 rounds of data, we split the whole dataset into an 8:5:1 ratio used for training, validation and testing through all experiments.

Table III presents the datapoint format. Each timestep (sample in one millisecond) contains time-sequential IMU and magnetic samplings within one second time window and the WiFi RSS scans at the current point. If there are no WiFi updates at a certain timestep, we use a *'null'* value to represent the missing value in the dataset. We record the ground truth location when passing by certain landmark such as corners, elevators and stairs during data collection, which are frequent occurrences. All other location labels assigned to each timestep are generated by linear interpolating with static samples at precise locations to create the full labelled dataset.

We categorise multiple modalities into two types: the infrastructure-free and the infrastructure-based modality. Specifically, the infrastructure-free sampling indicates to the samplings which have minimum variations caused by building-specific settings. Such as the motion samplings (e,g. walking, running, climbing stairs) captured by IMU sensors are less related to the building infrastructures but individual's movement gestures and behaviours. By contract, modalities including the magnetic field and WiFi RSS samplings are more related to geographical factors and physical forms of scenarios. For instance, buildings located at different geographical locations with different WiFi access points deployment strategies result



(a) Source Scenario Trajectory

(b) Target Scenario Trajectory

Fig. 3. Trajectories selected for gathering dataset from our two indoor evaluation scenarios.

in distinctive magnetic field samplings and WiFi Fingerprint datasets, hence, regarded as the infrastructure-based modality.

The purpose of categorising the multimodal dataset is to select which types of modalities are appropriate for implementing transfer learning techniques. Here, we consider the IMU samplings as the infrastructure-free modality for implementing transfer learning while the magnetic and RSS scans as the infrastructure-based modality that requires the network to extract building-specific features from new deployment sites from freshly collected on-site data.

B. Model Pre-training

Before performing transfer learning, we pretrain a transferable model that learns the infrastructure-free modality representations of human motion features (e.g., walking straightforward, turning around) captured by IMU sensors. We take the same strategy of constructing an LSTM network, proposed in [5]. Precisely, we construct an IMU-LSTM model that contains an input layer taking IMU sensor data (*timestep * 10 * 2*). Here, *timestep* represents the time window of the LSTM. In our situation, we consider time window of one second span. We implement a downsampling strategy to select samples with a period of 100 milliseconds. Hence, each datapoint capturing one second time window is of the shape 10 samples multiplied by 2 features (accelerator and gyroscope). A sliding window with an overlapping of 900 ms is implemented to allow the model to better learn the feature representations in between each two sampling inputs. For instance, if the first timestamp fed into the network starts from 0 to 999 ms, the next input sample is from 99 to 1,099 ms instead of from 1,000 to 1,999 ms. The output is a 2-dimensional regression layer that predicts the geo-spatial coordinates in x and y. We use the IMU dataset gathered from the source scenario to pretrain the model. The settings are illustrated in table IV.

C. Model Transfer

After pretraining, we extract the LSTM layer from the model. This transferable component carries the IMU knowledge learnt from the source scenario, behaving as a feature extractor for IMU samplings. This component is integrated in our MTLoc architecture using transfer learning. The multimodal network architecture contains the transferred IMU-LSTM model for infrastructure-free IMU samples, a MAG-LSTM and a DNN network for reading infrastructure-based samples. All modality-specific features are then fused by concatenation to a one dimensional vector and fedforward through a multi-layer perceptron for making the location estimation.

TABLE IV
PRETRAIN NETWORK PARAMETER SETTINGS

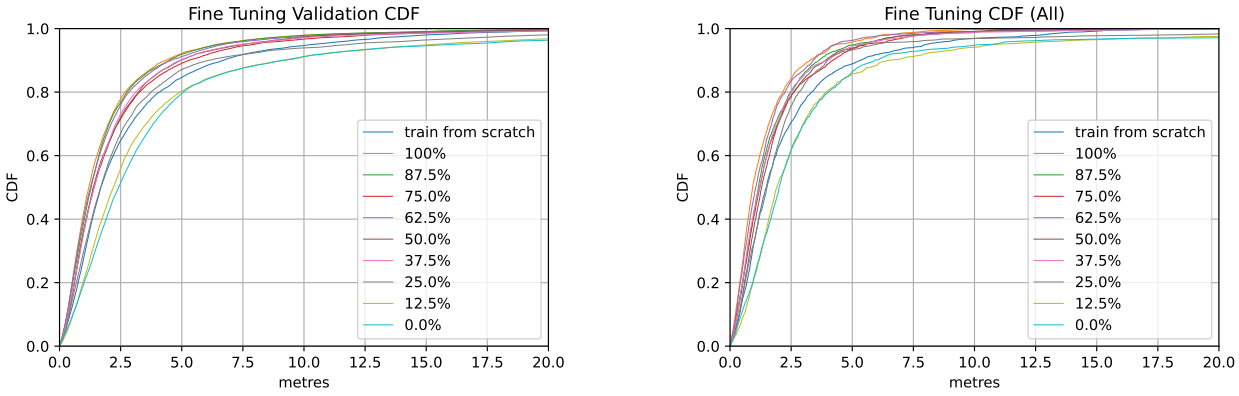
Parameter	Settings
Epoch	100
LSTM Layer	1 Layer
LSTM Hidden Units	128
LSTM Transferable Parameters	67072
Learning Rules	RMSprop
Learning Rate	0.005

D. Model Fine-tuning

To specialize the model after transfer learning, we fine-tune with a small amount of data collected from the new deployment site. Here, we vary the amount of data required for fine tuning to observe its impact. We take the parameters of the IMU-LSTM component pretrained on the source site, and set this sub network as trainable during the fine-tuning process. We gradually increase the amount of training data from 0% to 100% (total training sets include 8 rounds of data). Figure 4 illustrates the comparison results. Here, the line of 0% represents the transferred model (train from scratch) with freezing the IMU-LSTM weights and biases when feeding new inputs from the target domain data, while the line of 100% represents the transferred model with fine tuning based on the whole training set from the target domain.

Table V presents the results in median error, mean error and standard deviation obtained by the two training approaches, training from scratch and with a varying amount of fine-tuning data for transfer learning.

As expected, the model with transfer learning and fine-tuned with 100% dataset of the target domain data outperforms all



(a) CDF plotted based on the validation set

(b) CDF plotted based on the testing set

Fig. 4. Cumulative distribution function (CDF) plots showing the location prediction errors of models with different fractions of fine-tuning training data collected from the target scenario. The 0.0% is equivalent to the transferred model without any site adaptation.

TABLE V
PREDICTION ERRORS WITH DIFFERENT FINE-TUNING RATIO.

		Amount of fine-tuning ratio								
	Train from scratch	**100%**	**87.5%**	**75%**	**62.5%**	**50%**	**37.5%**	**25%**	**12.5%**	**0%**
Median Error	1.48	0.92	1.16	1.22	1.05	1.29	1.21	1.51	1.91	2.01
Mean	2.39	1.46	1.75	1.81	1.56	1.85	1.80	2.36	3.34	3.35
STD	2.77	1.63	2.07	1.92	1.82	1.96	2.17	3.70	4.75	5.09

other models. The model without fine tuning and the one with 12.5% (1/8 rounds) fine-tuning rates have lower prediction accuracy, compared to the model without implementing transfer learning. The results indicate that the IMU-LSTM component transfers the IMU representation learnt from the source scenario to the target scenario. Despite data from the new deployed scenarios being collected by different hardware and variations, the new model still benefits from the transferred information and keeps improving its inference accuracy by fine tuning with increasing amount of data. Here, the model with 62.5% fine-tuned configuration offers an accurate performance with minimum data demand, which only requires over half of the dataset but outperforms the model trained with the full dataset. It inherits the learnt knowledge of the infrastructure-free IMU samplings from pretraining on other sites and needs just a small amount of data for fine-tuning from the new deployment site.

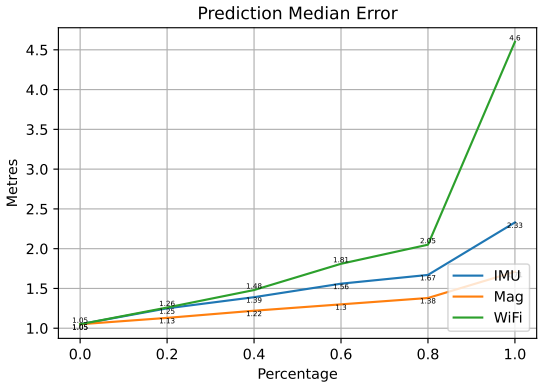
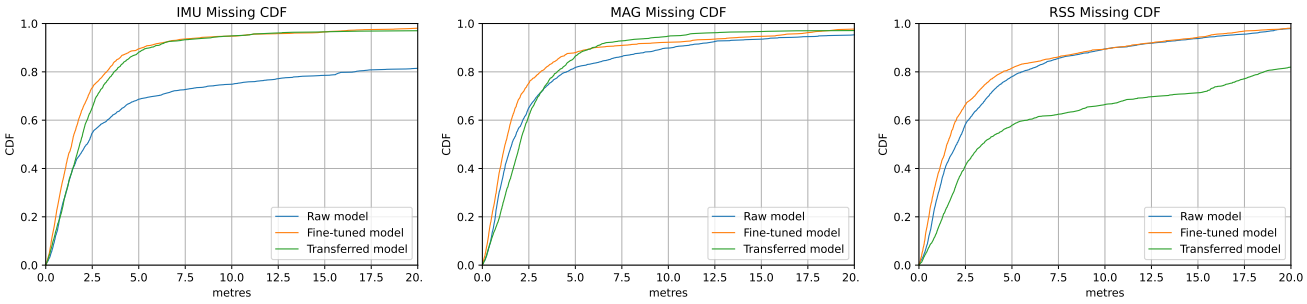


Fig. 5. MTLoc component performance in varying amount of available training data from the test set.

E. Modality Variability

Modality variability between the ones used to build computational models and those used by people during deployment significantly affects the model's inference accuracy and system robustness. Reasons for this variability mainly include that *i*) sensor malfunctions resulting in modality missing, *ii*) sensor network variations due to acceptability and privacy concerns, and *iii*) sensor hardware quality and user wearing preferences. All these factors bring the modality variability and modality missing challenges for transferring the model to new localization scenarios, resulting in multisensory based systems losing stability and robustness when a subset of the sensor networks fails to operate.

To evaluate the impacts of modality variations on model performance that what types and how dense the modalities are more contributive and correlated to localization prediction, we test the fine-tuned model without implementing the additional human intervention or system recalibration. We randomly remove the data points in the testing set to simulate the real-time situations of modality missing and irregular samplings. Specifically, we shelter each of the modality inputs from 0% (keeping whole inputs) to 100% (removing entire inputs) to the model during the online phase.



(a) IMU Missing CDF (b) MAG Missing CDF (c) RSS Missing CDF

Fig. 6. CDF Plots: Evaluating model robustness in missing data experiment by removing 40% of modality data in the target scenario.

We choose the model fine-tuned with 62.5% of data from the deployment site. Figure 5 shows the performance of this trained model under various modality missing situations. By increasingly removing valid data from 40% to 100%, the model's prediction accuracy drops from an acceptable precision of approximately 1.36 metres on average to over 2.88 metres median error. Estimation errors increase with valid data being dropped gradually. When removing the same amount from each modality, the absence of WiFi inputs has the most significant drawbacks to the model robustness, followed by the impact of magnetic field and IMU samplings. Hence, WiFi modality, containing the building-specific representations, contributes the most to localization estimation quality, while the IMU samplings offer rough information about the movement.

F. Model Robustness

To evaluate our proposed fine-tuned model, we further compare the fine-tuned model, the transferred model (without fine-tuning), and the raw model (without implementing transfer learning) under the same modality missing situation of removing 40% of each modality input.

Table VI shows a numerical comparison between the transferred model without fine-tuning and the fine-tuned model under the same situation that 40% of each modality input is being wiped. We observe that the fine-tuned model outperforms the transferred model with one-metre higher precision accuracy.

TABLE VI
MODEL PREDICTION MEDIAN ERRORS WHEN 40% OF EACH MODALITY BEING WIPED

Model	**Transferred Model**	**Fine-tuned Model**
Mag	2.01	1.22
IMU	1.81	1.40
WiFi	3.50	1.55
Average	2.44	1.39

Figure 6 represents the comparison results. In general, fine-tuned model outperforms all other models showing a robust performance for localization. In figure 6(a), when removing IMU sensor inputs, the fine-tuned model performs slightly better than the transferred model. It indicates that the IMU sampling representations are shared across scenarios and this knowledge is learnt and transferred from source scenario to target scenario through transfer learning. The new deployed model learns the complementary multimodal features by fine-tuning with the building-specific dataset to further improve inference accuracy.

In figure 6(b), the absence of the magnetic inputs has similar drawbacks to all models that transferred knowledge contributes a little to the model's prediction. It indicates that the magnetic scans are relatively isolated from the other sensing information. Even so, the fine-tuned model is still approximately 1 metre more accurate than the others. It is likely to explain that with transferred knowledge of the IMU samplings, the model boosted its ability to capture communicative features from the multisensory dataset as the transferred model holds not only the IMU features from the source scenario but also the deployment scenarios.

In figure 6(c), the transferred model without fine tuning shows an reduced performance compared to the raw model. After fine tuning, the model outperforms the raw model again. It indicates that the multimodal network makes location estimations majorly based on the deep insights from the WiFi and IMU features, instead of capturing the modality-specific information from each feature extractor. Through fine-tuning, the model re-captures the communicative representations of the RSS and IMU samplings from the deployment scenario, based on the precondition that the model has already held the transferable IMU knowledge from source scenarios.

We observe that in our deployment scenario, the WiFi modality contributes the most to the localization, followed by the magnetic field samplings and the human motion IMU samplings. However, this situation can be varied significantly from scenario to scenario depending on the information utility of each sensing modality. From a robust positioning perspective, a model should be tolerant to different localization feature representations by not only understanding modality-specific features independently but also learning the joined features comprehensively. We believe that the proposed multimodal network can identify these complementary features automatically and select the most discriminative fusion of sensor features.

V. Conclusion

In this work we leverage transfer learning for preparing our indoor localization system built on a multimodal deep neural network architecture (MTLoc) for deployment to new sites. Our approach requires just a minute amount of training data from the target spaces. The infrastructure-free component of our model is pretrained on source sites and integrated in the multimodal architecture. Then we fine-tune the model with a small amount of data from the target deployment site. Our MTLoc achieves an accuracy of 1.05 metres median error. This outperforms the model trained from scratch on data from the target site alone (trained with more data, but without transfer learning). Furthermore, our model is fault-tolerant, showing a robust performance when evaluated with 40% of modality data missing, still achieving a prediction median error of 1.39 metres without any system recalibration. Our system benefits from transfer learning with knowledge brought from source sites to the target sites, which makes our solution generalizable and scalable. This work takes our community closer to fast deploying and easily maintainable indoor localization systems.

References

[1] Hakan Koyuncu and Shuang Hua Yang. A survey of indoor positioning and object locating systems. *IJCSNS International Journal of Computer Science and Network Security*, 10(5):121–128, 2010.

[2] Valentin Radu and Mahesh K. Marina. Himloc: Indoor smartphone localization via activity aware pedestrian dead reckoning with selective crowdsourced wifi fingerprinting. In *Proc. IEEE IPIN 2013*, 2013.

[3] Priya Roy and Chandreyee Chowdhury. A survey of machine learning techniques for indoor localization and navigation systems. *Journal of Intelligent & Robotic Systems*, 101(3):1–34, 2021.

[4] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[5] Xijia Wei and Valentin Radu. Calibrating recurrent neural networks on smartphone inertial sensors for location tracking. In *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–8. IEEE, 2019.

[6] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G Okuno, and Tetsuya Ogata. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42(4):722–737, 2015.

[7] Xijia Wei, Zhiqiang Wei, and Valentin Radu. Mm-loc: Cross-sensor indoor smartphone location tracking using multimodal deep neural networks. In *2021 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–8. IEEE, 2021.

[8] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128, 2006.

[9] Stephane Beauregard and Harald Haas. Pedestrian dead reckoning: A basis for personal positioning. In *Proceedings of the 3rd Workshop on Positioning, Navigation and Communication*, pages 27–35, 2006.

[10] Esmond Mok and Günther Retscher. Location determination using wifi fingerprinting versus wifi trilateration. *Journal of Location Based Services*, 1(2):145–159, 2007.

[11] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D Lane, Cecilia Mascolo, Mahesh K Marina, and Fahim Kawsar. Multimodal deep learning for activity and context recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–27, 2018.

[12] Xijia Wei, Zhiqiang Wei, and Valentin Radu. Sensor-fusion for smartphone location tracking using hybrid multimodal deep neural networks. *Sensors*, 21(22):7488, 2021.

[13] Adrian Cosma, Ion Emilian Radoi, and Valentin Radu. Camloc: Pedestrian location estimation through body pose estimation on smart cameras. In *2019 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–8. IEEE, 2019.

[14] Sinno Jialin Pan, Vincent Wenchen Zheng, Qiang Yang, and Derek Hao Hu. Transfer learning for wifi-based indoor localization. In *Association for the advancement of artificial intelligence (AAAI) workshop*, volume 6. The Association for the Advancement of Artificial Intelligence Palo Alto, 2008.

[15] Martin Werner, Carsten Hahn, and Lorenz Schauer. Deepmovips: Visual indoor positioning using transfer learning. In *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pages 1–7. IEEE, 2016.