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

Leveraging Transfer Learning for Robust Multimodal Positioning Systems based on Smartphone Multisensory 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 complexity in GPS-denied scenarios and its acute demand in localization services. However, the variation of indoor scenarios and temporal conditions impose the need for building specific and periodic calibrations at high costs for deployment and maintenance of localization systems. A robust positioning solution that overcomes these challenges is yet to be available. Most previous solutions achieve good performance, but this is due to hard-coding the unique characteristics of the deployment site. We believe that in the current age of big data relying on end-to-end machine learning is the right path for generalizable localization systems, taking minimal human intervention. We build our solution on a multimodal deep neural network architecture, which is ideal for transfer learning in new deployments. Our transfer learning extension boosts the quality of location estimation at a reduced cost for fine-tuning compared to training the model from scratch. Our experiments demonstrate that training data can be reduced by halving with minimal impact on the location estimation quality in our transfer learning solution. Our research opens the way from 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 enhancing our life, improving the way we interact with the world. The Global Positioning System (GPS) on our mobile devices has been widely used in various outdoor scenarios to provide information and navigation, thus flourishing a geo-spatial digital world of goods and services. However, the GPS fails in indoor and underground environments due to GPS satellite signals being distorted or missing entirely in buildings and underground. Alternative solutions for indoor location estimation have been proposed relying on smartphone built-in sensors, such as WiFi received signal strength, magnetic field, and inertial sensors (e.g., accelerometer, gyroscope) [1]. This has been the focus of intensive research over the past years, but with moderate effect.

Conventional engineering based localization solutions mainly include Pedestrian Dead Reckoning (PDR) and WiFi Fingerprinting [2]. Both of these alternatives require well engineered solutions built on accurate mathematical formulations to process sensor signals. However, these heavily-engineered solutions often lose efficiency when deployed in new scenarios. This is due to environment change 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 affect the system's robustness. As a result, human intervention for periodic calibration is essential for maintaining the accurate functioning of these systems, which makes wide adoption unattainable.

In the age of big data, we believe that relying on data itself to deliver an end-to-end data-driven machine learning approach is the only promising solution for robust indoor localization, instead of conventional dedicated engineering solutions. Furthermore, the model should have the ability to carry the learnt knowledge across multiple deployment sites for better robustness, instead of requiring a big amount of dataset collected from the new deployment site for training from scratch.

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 [3], in this study, we propose an end-to-end hybrid multimodal architecture integrated with transferable sub-components, named here MTLoc, to deliver robust location estimation from smartphone sensor-fusion data.

We use a multimodal dataset collected by ourselves from two indoor scenarios. Both datasets contain time-sequential IMU sensors and magnetic samples as well as WiFi Received Signal Strength (RSS) fingerprints from corridors, along with ground truth location annotations. We qualify the multi-sensor dataset 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 recurrent model (IMU-LSTM) as a feature extractor and then integrate this model component into MTLoc architecture using transfer learning methods. For extracting infrastructure-based features, we construct another long short-term memory model (MAG-LSTM) and a deep neural network (DNN) to extract multi-sensor features. All extracted modality-

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

specific features are then joined in a one-dimensional vector, followed by additional multi-layer perceptrons to produce the joint location estimation. The transferable component of the IMU-LSTM is pretrained on the dataset collected from the source scenario (lab conditions). After pretraining, we integrate this model into the MTLoc architecture to allow the model to carry the learnt infrastructure-free representations to the target deployment. We explore the utility of data amount in the fine-tuning stage of our MTLoc by varying the amount of training data available from the target settings (deployment site). Furthermore, we corrupt valid multi-sensor samples in order to evaluate the fine-tuned model's stability and robustness under the target missing and noisy data. We find that the MTLoc can achieve better results under missing data of one or more modalities from the target site, being bootstrapped instead with just a small number of samples. MTLoc predicts the trajectory with good fidelity of ground truth, over 80% of the estimations being within 3 meters of error. Benefiting from the transferred knowledge, MTLoc fine-tuned with a small amount of data, shows robustness in performance 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 only half the amount of data, the model achieves a good performance, median estimation error being within 1.56 metres, which is actually better than training without transfer learning, but with full amount of the target site data alone (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 by a minimal amount of data to achieve a prediction mean accuracy of 2.92 metres without human intervention for system recalibration.

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. When indoor environmental changes and when the model is deployed in a new scenario, system refinement is required for re-fitting on the new data variation, resulting in additional tuning costs and lower efficiency.

In recent years, machine learning based positioning system has become a research hotspot for data-driven localization approaches, without heavy human intervention and deployment costs [4]. Current solutions are mainly explored on single-modality data, though limited by 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 can be varied in different deployments, depending on the available signal sources.

Multimodal machine learning has been investigated in many modality-mixed tasks such as audio-visual speech recognition [5]. It shows the advantages of understanding correspondences between complementary multimodalities and capturing comprehensive features from multisensory representations instead of relying on a single modality. Inspired by the success of multimodal machine learning, we explored a multimodal hybrid deep neural network for location tracking in our previous work [6]. This is a first step towards offering a purely data-driven end-to-end robust localization approach by utilising multi-sensors inputs of the IMU, magnetic and WiFi RSS data together.

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 modality variation present in the sensor data?*

Transfer learning has been applied to many AI problems with promising performances, 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 scenarios [7]. However, to date, this technical solution has not been validated for accelerating the deployment of effective localization systems. To address the aforementioned issues of data scarcity in new deployments and reusing knowledge across sites, we believe transfer learning of multimodal sensing is viable and effective in creating robust localization systems with reduced deployment costs.

A. Engineering-based Positioning

Engineering-based indoor positioning systems commonly rely on two mechanism called Pedestrian Dead Reckoning (PDR) [8] and WiFi Fingerprinting [9], which work on a set of hand-crafted formulations to identify the mobility frame including step counting, orientation estimation and fingerprints matching for localization [10]. Systems are usually designed to be building-specific. When indoor environmental signal distribution changes or for new scenario deployment, system recalibration is needed for the model to fit with the variations, resulting in additional tuning costs and lower efficiency.

B. AI-based Positioning

Instead of engineering-based solutions, artificial intelligence based positioning system shows its advantages of low deployment cost without requiring accurate mathematical equations, though moving the focus to the quality and quantity of the dataset itself [11]. For instance, HiMLoc integrates IMU sensors with WiFi RSS samplings through prior observations of Gaussian processes for direction estimation, distance estimation and correlation, and admissible human activity [2]. CamLoc [12] uses computer vision to identify the tracking target, feet position and pedestrian skeleton for obtaining target locations.

C. Transfer Learning-based Positioning

To date, transfer learning based localization is still under well exploration. There are a few examples based on single modality positioning systems with transfer learning techniques. One example is that Pan et al. implement transfer learning on a WiFi-based positioning system to address the challenge that WiFi signal distribution is varied across changing time and device differences [13]. Another example is that Werner et al. integrate transfer learning to a vision-based localization system to assist positioning by migrating the image recognition ability from deep convolutional neural networks to the system for identifying indoor symbolic targets [14]. These experiments evaluate the model's performance under different working conditions such as time variations and device variations in the same building but lack the evaluation when a model is transferred to other scenarios.

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. This relies 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 describes the model architecture construction. It contains 295,810 trainable parameters in total.

B. Transfer Learning

Figure 2 illustrates the procedures for implementing transfer learning to the 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 performs as a non-trainable sub network.

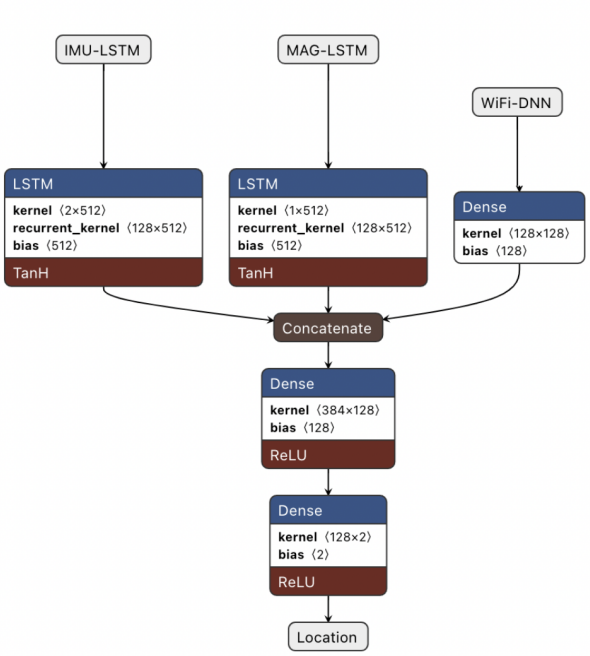


Fig. 1. MTLoc Model Architecture

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

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's parameters, are updated during the gradient descent to allow the whole model to fit with the new scenario.

C. Fine Tuning

To allow the new deployed model to better fit with the new scenario samplings feature with a small portion of the dataset, we implement fine tuning to the 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 network's parameter trainable, all model parameters are updating during the training process based on transferred knowledge.

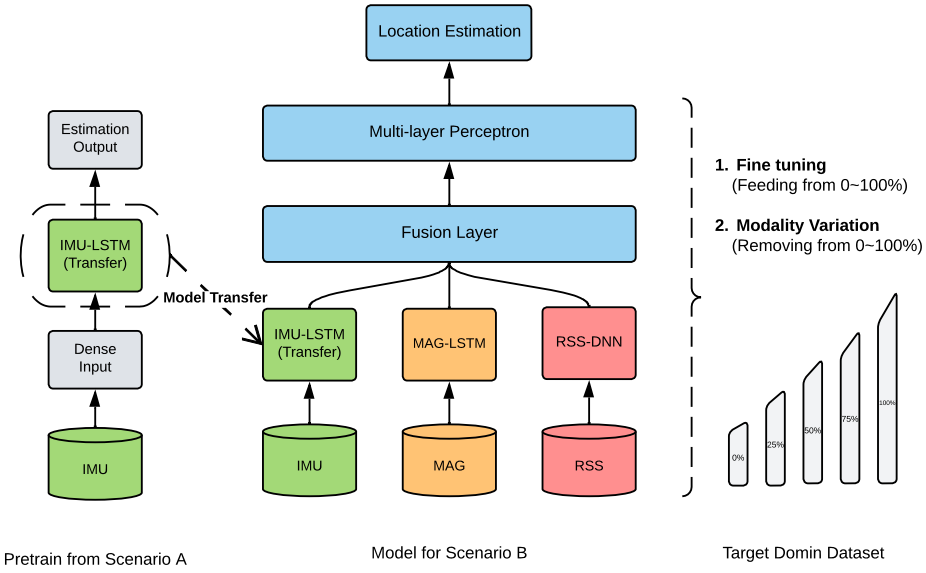


Fig. 2. MTLoc Transfer Learning and Fine Tuning Procedure

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 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 collecting times, walking postures and speeds. Meanwhile, the occasionally appearing WiFi hotspot are kept to simulate real usage situations that contain signal interference.

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	X*2*	Y*2*
...	...	...	...	...	...	...	...	...	...
Tn	a(*n∼n+999*)	g(*n∼n+999*)	m(*n∼n+999*)	null	null	...	null	X*n*	Y*n*

Table II presents the samples distribution collected from two experimental scenarios. Precisely, the source scenario dataset includes 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 collect from two scenarios contain 14 rounds of data, we split the whole dataset into an 8:5:1 ratio for training, validation and testing through all experiments.

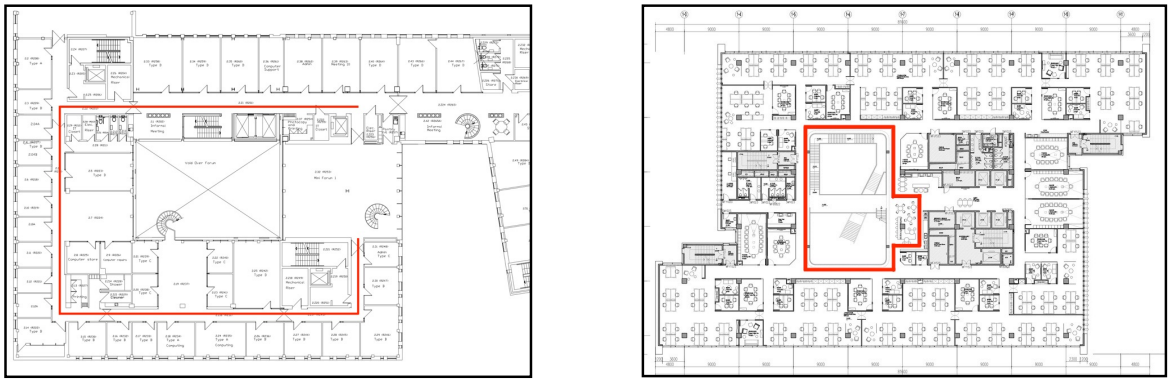
Table III represents the dataset format. Each timestep (sample in one millisecond) contains time-sequential IMU and magnetic samplings within one second time window and the WiFi RSS scans at the current point. If there are no WiFi updates at certain timesteps, we use a *'null'* value to represent the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

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

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 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



(a) Source Scenario Trajectory

(b) Target Scenario Trajectory

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

select which types of modalities are appropriate for implementing transfer learning techniques. Here, in our situation, 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 deployed scenarios by feeding new dataset.

B. Model Pre-training

Before implementing transfer learning, we first need to pretrain a transferable model that learns the infrastructure-free modality representations which contain the 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 [4]. Precisely, we construct an IMU-LSTM model that contains an input layer for accepting IMU sensor data (timestep * 10 * 2). Here, the timestep represents the time window of the LSTM. In our situation, we consider a one second time coverage. We implement a downsampling strategy that pick one sampling every 100 milliseconds. Hence, in each datapoint of a one second time window, the shape is 10 samplings multiplied by 2 features (accelerator and gyroscope). Meanwhile, 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 geographical coordinates of x and y. We use the IMU dataset gathering from the source scenario to pretrain the model, the parameter settings are illustrated in table IV.

C. Model Transfer

After pre-training, 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. We integrate this model component into MTLoc architecture using transfer learning methods. The multimodal network architecture contains the transferred IMU-LSTM model for accepting infrastructure-free IMU samplings, a MAG-LSTM and a DNN network for reading infrastructure-based multi-sensor inputs. All modality-specific extracted features are then fused by concatenation to a one-dimension vector and feedforward to top multi-layer perceptrons for making location predictions.

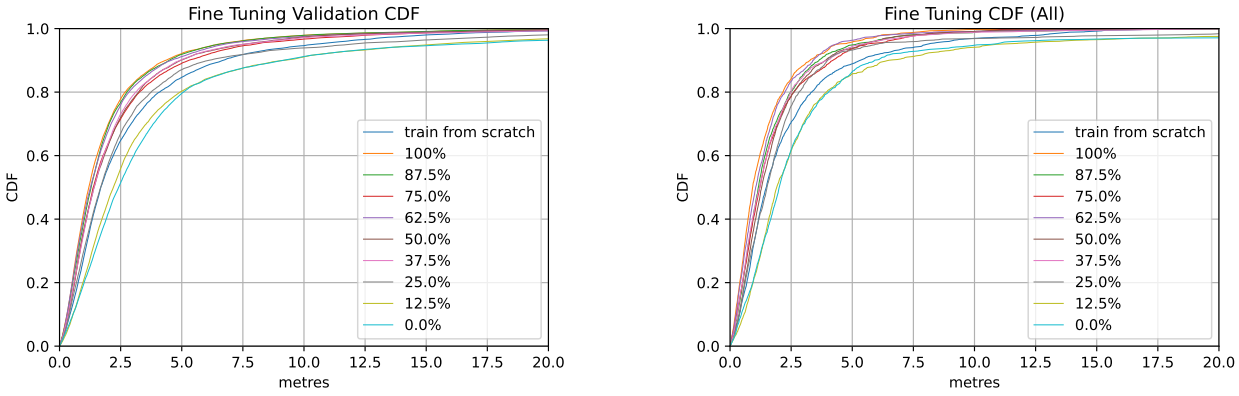
TABLE IV
PRETRAIN NETWORK PARAMETER SETTINGS

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 get a better prediction performance after implementing transfer learning, we feed the model with an increasing amount of dataset collected from the new deployed scenario to explore the minimum amount of data required by fine tuning. We start by freezing the weights and bias updates of the transferred IMU-LSTM component during the model training process. Alternatively, we initialise the parameters of the IMU-LSTM component by the transferred model's parameters, and set this sub network as trainable during the fine-tuning process. We gradually increase the amount of the training set 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 component's weights and bias when feeding new inputs from the target domain dataset, while the line of 100% represents the transferred model with fine tuning based on the whole training set from the target domain.

Table V summarises the statistic results of the mean and standard deviation of the model's prediction errors based on the testing dataset gathering from the target scenario. By



(a) CDF plotted based on the validation set

(b) CDF plotted based on the testing set

Fig. 4. Cumulative distribution function (CDF) plot shows the location predictions of models with different fine tuning amount of the data of the target scenario.

TABLE V
PREDICTION ERRORS WITH DIFFERENT FINE-TUNING RATES

Fine-tuning rate	**Train from scratch**	**100%**	**87.5%**	**75%**	**62.5%**
Mean	2.39	1.46	1.75	1.81	1.56
STD	2.77	1.63	2.07	1.92	1.82
Fine-tuning rate	**50%**	**37.5%**	**25%**	**12.5%**	**0%**
Mean	1.85	1.80	2.36	3.34	3.35
STD	1.96	2.17	3.70	4.75	5.09

observation, we find the model fine-tuned based on 100% (8/8 rounds) dataset outperforms all other fine-tuned models, though its accuracy is slightly higher than the model fine-tuned with 62.5% (5/8 rounds) dataset. 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 methods but trained from the beginning. The results indicate that the IMU-LSTM component transfers the IMU representation learnt from the source scenario to the target scenario. Despite that the data from the new deployed scenarios is collected by different hardware and variations, the new model can still benefit from the transferred information and keep improving its inference accuracy by fine tuning with increasing dataset quantities. 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 previous scenarios and benefits from a small portion of the new deployment dataset.

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,

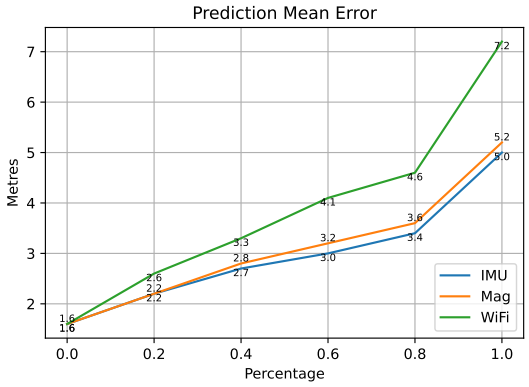
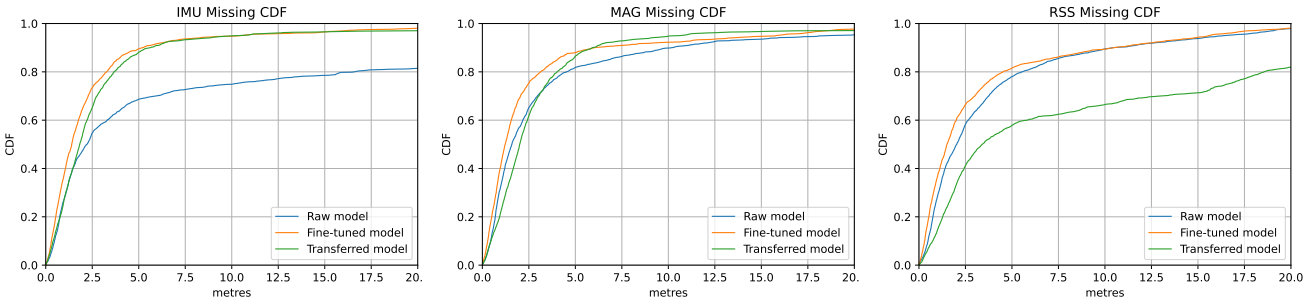


Fig. 5. Model's performances based on different dropping rates of datapoints in the testing set.

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.

By using 62.5% of the dataset from the deployment scenario for fine tuning, we finalise the transferred model. Figure 5 shows this fine-tuned model's performances under various modality missing situations. By increasingly removing the



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

Fig. 6. CDF Plot: Evaluating model's robustness by removing 40% of the valid data of each modality in the target scenario.

valid data from 40% to 100%, the model's prediction accuracy drops from an acceptable precision of approximately 2.9 metres on average to over 5.6 metres shifting from the ground truth. Estimation errors increase with valid data being dropped steadily. When removing the same amount of 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, while the IMU samplings offer rough information of the movement representations to assist the model for positioning, though we have approved that the IMU-based localization can still provide an acceptable inference accuracy [4].

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 higher precision accuracy (over two metres).

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

Model	**Transferred Model**	**Fine-tuned Model**
Mag	3.93	2.76
IMU	7.99	2.66
WiFi	3.72	3.33
Average	5.21	2.92

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.

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), it is interesting that the transferred model without fine tuning shows an unsatisfied 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 communicative information of the WiFi and IMU sampling features via the fusing sub component, 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.

It needs to be mentioned that, in our deployment scenario, the WiFi modality contributes 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 that what types of multisensory samplings are more mutually communicative for localization. From a robust positioning perspective, a model should 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 to select what types of the multisensory combination are more distinctive and crucial for making robust estimations under different modality variations.

V. Conclusion

In this work, we present how to leverage transfer learning in a multimodal deep neural network (MTLoc), to produce a robust localization system with a minimal amount of fine-tuning data from the target deployment site. First, we pretrain the infrastructure-free component of our model on samples from the source site and then integrate this transferable component into the multimodal architecture. Then we fine-tune the model with a small amount of the data from the target site (at deployment). The MTLoc achieves an accuracy of 1.56 metres median error. This outperforms the model that is trained based on data from the target site alone trained from scratch, without transfer learning. Furthermore, the model shows a robust performance when evaluated with 40% of modality data missing, still achieving a prediction mean error of 2.92 metres without any system recalibration. The proposed solution greatly benefits from the training conditions of transfer learning, bringing knowledge from the source site to the target site, which makes the solution generalizable and scalable. This work takes our community one step closer to fast and cheap deployment of robust indoor positioning 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] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.