A Multi-Device, Multi-Modal, Convolution-Augmented Transformer-Based Model for User Localization

Tian Qiu
Device and Services Department
Amazon.com Inc.
Sunnyvale, United States
tiantsu@amazon.com

Rohit Kumar
Device and Services Department
Amazon.com Inc.
Sunnyvale, United States
rrohk@amazon.com

Sai Prashanth Chinnapalli
Device and Services Department
Amazon.com Inc.
Sunnyvale, United States
saic@amazon.com

***Abstract*—Transformer architecture has shown great success in many challenging problems such as natural language processing. It excels at discovering features and correlations in large amounts of data. Sensor fusion is another technology that requires this kind of processing power and reasoning capability. Researchers have applied transformer architecture in sensor fusion model technologies. However, most of them focus on self-driving applications where camera and Lidar are the main data sources. In this paper, we propose a novel transformer-based model that focuses on wireless and ultrasound sensors. Specifically, this mainly targets user localization in a smart home context. The model is trained on data collected in a real home setup and is compared with several baseline algorithms. The localization accuracy is presented as a performance index. Hyperparameters of the model are also investigated to optimize the trade-off between power consumption and model performance.**

***Index Terms*—transformer, distributed system, heterogeneous sensor fusion, localization, convolution neural network.**

I. Introduction

User location information is valuable in consumer electronics industry. Applications include spatial audio rendering [1], device arbitration [2], health monitoring [3], and scene detection [4]. Localization in this setting, however, is difficult for three reasons. First, indoor environments are cluttered, which degrades the propagation of both radio waves [5] and acoustic waves [6]: GPS signals are unreliable indoors, while radar and ultrasound returns are corrupted by strong reflections from walls, furniture, and other surfaces. Second, the bandwidth and transmit power available to consumer-grade sensors are capped by regulatory limits [7], which in turn bound the achievable signal-to-noise ratio and sensing range. Third, most consumers are price-sensitive, so keeping the bill-of-materials (BOM) cost low is a hard requirement [8] that constrains component and vendor selection at the hardware design stage.

There are many sensing technologies for target range, angle, or location estimation. Wireless signals have been used for localization purposes for decades [9], [10]. One of the main signals used is received-signal-strength-indicator (RSSI) [11], [12]. Distance and trilateration-based localization can achieve a certain level of accuracy, but that would require the user to carry a mobile device to constantly communicate with stationary anchors [13]. This would be a customer friction. A large-scale RSSI-based localization system is proposed in [14]. However, in the consumer electronics domain, this level of system complexity would result in a prohibitively expensive product solution. Wi-Fi has seen tremendous growth over the past couple of decades and its signal has been used for localization in many ways. One method is called channel state information (CSI) [15]. However, the algorithms developed are usually specific to the environment and normally do not generalize well. Frequency modulated continuous wave radar technology was developed [16]. It has seen applications such as heart rate estimation [17]. It has proved to be effective when implemented on Wi-Fi hardware [18]. The performance of FMCW architecture heavily depends on the hardware setup. In the consumer electronics domain where bandwidth is heavily regulated, performance is usually challenged. Moreover, by sharing hardware with normal Wi-Fi data traffic, the air-time for FMCW is also limited. That means traditional classic signal processing pipelines would perform poorly in those conditions.

Acoustic-based localization has seen application in robot positioning and many other ranging tasks [19]. It relies on sending mechanical vibrations and collecting the bounced-back signal, computing their time shift or correlation to estimate the range; frequency modulated continuous wave is a commonly used method [20]. Time of flight is also another classic method to estimate range between source and microphone [21]. However, this would require either the user or the device to produce sound, creating customer engagement friction. Ultrasound signals are another popular method for localization [22] and imaging purposes [23]. Their coverage is usually limited compared to wireless signals. Researchers have spent significant effort to improve their coverage [24]. Cameras can achieve great precision with modern convolution neural network-based models [25]. Information from some models is so granular that even body pose can be derived [26]. The main drawback of computer vision is the privacy

concern.

Every sensing modality has its advantages and disadvantages, therefore sensor fusion is important for a lot of applications [27], such as the range based user localization shown in Fig. 1. Transformer, introduced in [28], has shown great value in terms of feature extraction [29] and dimension reduction [30]. It has great potential for extracting semantics from the raw data. For example, round-trip time in a satellite network is estimated using transformer [31]. Transformer can also be used for CSI matrix compression as discussed in [32]. It has also provided a novel way to analyze the state-action pairs for unmanned aerial vehicle positioning purposes [33].

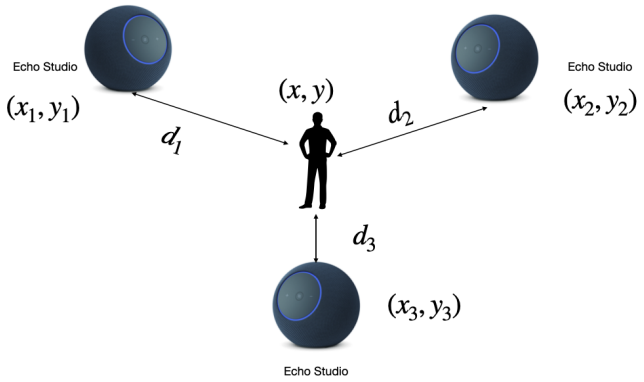


Fig. 1: An example network of devices with user localization capability (Three-device example)

The paper is organized as follows: Section II provides a brief review of relevant technologies and model architectures. Section III explains the main method used for the proposed model and relevant analysis methods. Section IV reviews the data collection procedure, key performance indices and other analysis results. Section V summarizes the main findings and future opportunities.

II. Background

A. Frequency Modulated Continuous Wave Radar

A Frequency Modulated Continuous Wave (FMCW) Radar sends out modulated radio waves and analyzes the frequency shift between the sent and received signal to determine the target distance and velocity. When the frequency follows a linear curve, the signal is called a "chirp", shown in Eq. 1. The target range and the frequency follow a linear mapping.

$$s(t) = A\cos(2\pi(f_c t + \frac{B}{2T}t^2)) \tag{1}$$

Where t is time, A is signal amplitude, f_c is carrier frequency, B is bandwidth, and T is period. When the signal bounces back to the receiver, it is mixed with the transmitted signal and goes through a low pass filter to extract the beat frequency f_b, then the target distance can be computed with Eq. 2

$$d = \frac{f_b T c}{2B} \tag{2}$$

where c is the speed of light. If given dedicated hardware and sufficient bandwidth, FMCW can achieve very high accuracy with very basic fast-fourier transform.

In Wi-Fi radar context, f_c is the Wi-Fi traffic carrier frequency and B is regulated by various local agencies. The signal is noisy due to the low-cost hardware, the synchronization between the received and sent signal is poor and the bandwidth is limited by various laws and regulations. This signal is used as the "IQ Samples" input in Fig. 3.

B. Ultrasound Proximity Sensing

Ultrasound has been used for near-field distance measurement in many fields [34]. The most common method is measuring the time-of-flight of acoustic signals. The system usually has at least one speaker and one receiver. The speaker sends a modulated signal, and the receiver collects the bounced-back signal and applies a matched filter between the reference signal and the received signal. The window of the matched filter slides along the timeline and generates a time and correlation plot. In the user localization domain, we are more interested in the user as a moving target; therefore, direct derivative style Hilbert transform is widely used. This signal is labeled as "acoustic wave" in the bottom signal chain in Fig. 2.

C. Transformer

Transformer was introduced in [28]. It provides a clear and concise way to decompose vectors into three components: Q for query, K for key, and V for value. The attention function is defined as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

It has shown great success in large language model and many other applications. This architecture provides a standard way to estimate the cross correlation between samples from the same data stream, therefore paves the way for more advanced model developments in sensing technologies. The "Wi-Fi radar Encoder", "Ultrasound Encoder" and the "Distance Decoder" in Fig. 3 all follow this architecture.

D. Convolution Neural Network

Neural Network was introduced in [35], coupled with back-propagation proposed in [36] have shown great success in computer vision and many other fields. It is one of the most important supervised learning methods. The model starts with a network of sigmoid functions as shown in Eq. 4, where x is the input vector, and w is the unknown weights. The w is usually initialized to random values. A loss function is usually defined based on the difference between output and the ground truth. The training data is then fed to the network, and partial derivatives with respect to the weights are calculated using the inference results and ground truth. Those derivatives are back-propagated from the top layer down to the bottom layers. This approach enables the model to learn the pattern from the data

automatically and systematically. The "Wi-Fi radar CNN" and the "Ultrasound CNN" block both follow this architecture.

$$\sigma(x) = \frac{1}{1 + e^{-w^T x}} \tag{4}$$

E. Range-Based Trilateration

The range-based trilateration algorithm, derived from the spherical-intersection formulation introduced by [37], has been widely used in many applications. It considers a common setup of N beacon nodes. Their locations (x_i, y_i) are known on the same 2D plane. The goal is to find the unknown node (x, y). Each beacon node reports the estimated distance d_i between itself and the unknown node. Then the unknown location (x, y) can be derived by Eq. 5.

$$X = (A^T A)^{-1} A^T B \tag{5}$$

Where

$$A = 2 \begin{bmatrix} (x_1 - x_2) & (y_1 - y_2) \\ (x_1 - x_3) & (y_1 - y_3) \\ \vdots & \vdots \\ (x_1 - x_N) & (y_1 - y_N) \end{bmatrix} \tag{6}$$

$$B = \begin{bmatrix} d_2^2 - d_1^2 - (x_2^2 + y_2^2) + (x_1^2 + y_1^2) \\ d_3^2 - d_1^2 - (x_3^2 + y_3^2) + (x_1^2 + y_1^2) \\ \vdots \\ d_N^2 - d_1^2 - (x_N^2 + y_N^2) + (x_1^2 + y_1^2) \end{bmatrix} \tag{7}$$

$$X = \begin{bmatrix} x \\ y \end{bmatrix} \tag{8}$$

This is an inexpensive and robust way to combine range estimates from different devices and has seen application in both acoustic and wireless localization related publications. This method is used in the post range estimate in our proposal as shown in Fig. 2.

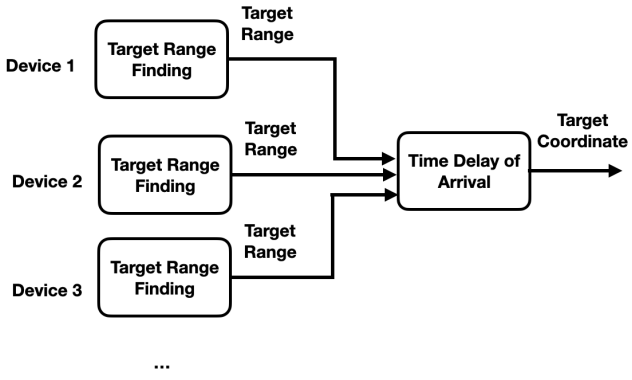


Fig. 2: General range based user localization

III. Method

In this section, we propose our fusion model for user localization. The architecture is a convolution-augmented transformer model. There are two major inputs to the model: ultrasound and Wi-Fi radar. Each of them has its own convolution network for feature extraction and dimension reduction. The output tensors are then passed through a transformer model to compute the fusion-based distance. There have been a few applications for similar architecture [38]; to our knowledge, this is the first time it is applied and optimized for our problem statement.

A. Problem Statement

Consider a room with N devices with Wi-Fi radar and ultrasound sensors. The locations $x_i, y_i, i = 1...N$ of the devices are known. With both sensors operating under separate microcontrollers streaming the data, assume there is only one person in the field of view; solve for the location of the person x, y. The main challenge here is that each sensing modality has its pros and cons. When traveling in air, the signal strength of the ultrasound wave gets attenuated heavily, depending on humidity and many other factors [39]. The radio waves have low signal-to-noise ratio because of the bandwidth limitation and digital-domain calculations. Therefore, using only one modality is not sufficient. Classical fusion approaches, such as wiener filter, uses inverse-variance weighting to combine the two range estimates, but they require an explicit noise model for each sensor and a hand-tuned handoff between the near-field (where ultrasound dominates) and the far-field (where radar dominates). These noise models depend on humidity, room layout, and device placement, none of which are known at deployment. A data-driven model learns the range-dependent weighting directly from paired samples and adapts to the local clutter and device geometry without hand-tuning.

B. Proposed Model Architecture

The overall system architecture is shown in Fig. 2. Each device measures the distance to the major moving target and the results are combined using Eq. 5. The detailed architecture for each device is shown in Fig. 3. There are three major algorithm blocks: 1. Wi-Fi radar pipeline with convolution neural network and transformer encoder, 2. ultrasound CNN and transformer encoder, 3. transformer decoder for range estimation.

Ultrasound and Wi-Fi radar have separate processing pipelines when it comes to convolution neural network. This is due to the fact that they are fundamentally two different types of signals: one is radio wave, the other one is mechanical vibration. Each reflects very differently depending on the surface and ambient environment [40], [41]. The input to the Wi-Fi radar branch is the Fourier transformed chirp. In a classic signal processing pipeline, moving target indication (MTI) is normally used for range estimation. It is usually prone to noise, especially in cluttered environments. In our model, a machine learning centered method is applied. The first step is the mean over burst in the same frame. This is because those bursts are repeated measurements of the same signal. There are no learnable parameters in this step. In the spectral convolution in Fig. 5, two conv layers slide along the frequency axis to discover spatial features on the range bins. The first layer takes

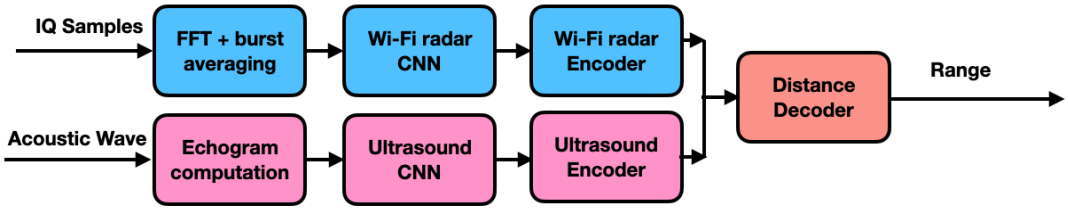


Fig. 3: Device level range estimation model architecture

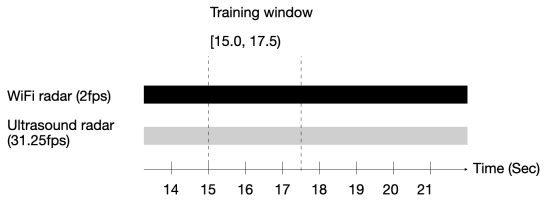
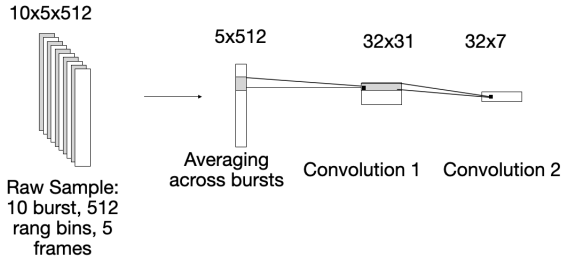
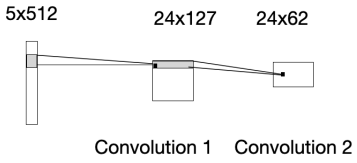


Fig. 4: Training Sample and Sensor Data



(a) Wi-Fi radar



(b) Ultrasound

Fig. 5: Detailed CNN architecture

512 bins as input and uses 32 distinct filters to extract 31 patches by sliding a window on the frequency bins; the second layer further reduces to 7 patches. The main purpose of this step is feature extraction. The Wi-Fi radar encoder layer has 2 heads and a 64-wide feedforward. Self-attention lets each range token condition on all the others, which is what allows the model to weigh a candidate target peak against leakage and wall reflections elsewhere in the profile.

The ultrasound branch follows a very similar design pattern. First the convolution neural network is used as shown in Fig. 5, then the transformer encoder is applied. The major difference is the input. Raw mic data stream goes through matched filter and Hilbert envelope [42]. The ultrasound encoder layer has 2 heads and a 48-wide feedforward. Self-attention operates over a much finer partition of the range axis than the radar branch, so it can resolve closely spaced echoes and suppress the residual direct-path and wall returns that survive the echogram's clutter removal.

The last major block is the transformer decoder. Its inputs are the two outputs from the Wi-Fi branch and the ultrasound branch. The two embeddings are concatenated together to form a fused embedding. This embedding is then fed to a one-layer 2-head decoder to generate a 1 by 32 tensor. The tensor is then passed to a linear transform to become the predicted distance. Since the embedding is a combination of two modalities, this model guarantees a completely data-driven fusion approach. One thing about those two sensors is they have different range coverage. Ultrasound has shorter range because acoustic waves are heavily attenuated in air. With this model, we do not have to manually look for the range coverage; the model learns the weighting from the data.

The Wi-Fi radar updates at 2 FPS and ultrasound updates at 31.25 FPS. The training sample is made up of 2.5 seconds of data: five Wi-Fi radar frames and about 78 ultrasound frames. The ultrasound echogram is then averaged from 78 samples to 5 samples to align with the Wi-Fi radar data. Then the two modalities are appended to create a new training sample as shown in Fig. 4.

C. Training and Loss function

Mean squared error (MSE) is the loss function for the training process as shown in Eq. 9.

$$L = \frac{1}{n}\Sigma_{i=1}^{n}(\tilde{d}_i - d_i)^2 \tag{9}$$

Where N is the number of samples considered, d_i is the ground truth distance, and $\tilde{d}_i$ is the predicted distance. This loss function is intentionally left simple, because we want the data to guide the search process about how to weight the features from the sensors, or which sensor to trust more in general.

All initial weights are set to random values. The training process uses the loss function explained in Eq. 9 as the objective function and the gradients are backpropagated from decoder, back to encoders then to the CNNs. During training,

batch normalization and layer normalization are applied. ReLU is applied at each layer. The learning rate is a cosine decay [43].

D. Baseline Methods

To show the efficacy of our proposed model, we also evaluated the following algorithms as a comparison: 1. End-to-end transformer. 2. Classic peak finding algorithms 3. Wiener filter based fusion.

1) End-to-End Transformer: One way to apply transformer architecture is to directly implement it on the data pipeline end-to-end, where the raw signal is the input and the user co-ordinate is the output. The model contains the same branches at the beginning: Wi-Fi radar and ultrasound branches. The outputs of those two branches are sent to a localization transformer. This transformer also accepts device position as input. The advantage of this approach is it is completely data driven and model based. The model can discover and correlate all features from all sensors on all of the devices. However, this is also its disadvantage: the correlation between devices will change from room-to-room, therefore, weights learnt from the training samples will not generalize very well for the validation dataset.

2) Peak Finding Algorithms: In order to show the benefit of the model over a classic signal processing pipeline, we developed a pipeline with basic trilateration and peak finding algorithms. The IQ samples were fed to the peak finding algorithms to search for the receiver-transmitter leakage, then the distance between the 2nd peak and the 0 range bin is computed using the FMCW parameters. The resulting distance is then sent to the trilateration algorithm to find the target location.

3) Wiener Filter Based Fusion: The Wiener filter is a powerful tool for sensor fusion and noise reduction, offering a closed-form optimal solution in both the time and frequency domains. Its effectiveness, however, depends on an accurate characterization of each sensor's noise statistics. In consumer electronics this is difficult to obtain: because devices are manufactured at very high volume, unit-to-unit variation is large, and the noise characteristics of any individual device may depart substantially from the population mean. Kalman filter is not considered in this paper because it was mainly for dynamic tracking, a different type of problem.

IV. Experiments and Results

In order to show the efficacy of our proposed model and its robustness to various environments, we conducted a large-scale data collection. The results are compared with classic signal processing pipelines mentioned in Sec. III-D. In all scenarios, our model shows great improvements over previous methods.

A. Data Collection Procedure

Two actors were recruited and three different room arrangements were considered. They were both compensated for their time spent in the data collection. They were instructed to do normal activities such as walking, watching TV, or standing up/sitting down. The furniture and the devices were rearranged into three different layouts. There are five types of devices, all of them have Wi-Fi radar and ultrasound capability. They all have different antenna, speaker, and microphone designs. This is to ensure that the data is not specific to any of the hardware configurations. It is known that antenna placement has a huge impact on the signal quality [44] and speaker design also contributes to acoustic characteristics [45].

The three rooms mentioned above all had ordinary furniture layouts. One room was around 4.00m x 4.00m, the second one was around 4.27m x 5.58m, the third one was 4.27m x 4.8m. The first room was a living room with a three-cushion sofa and a TV stand with TV on it. There were five devices placed around the room. The second and third room had similar layout with two functional areas: a living room area and a dining room area. The living room area contained a TV stand, an L-shaped sofa, and a coffee table. There were one TV remote, one bag of chips, and one cup of drink on the coffee table. The dining room area had one dining table, one chair, and some cabinets. The data collection devices were placed around the room following normal use cases. In all three rooms, two devices were placed on either side of the TV while the other two were placed either behind the sofa or on the tables. One room additionally had a fifth device on the far left of the TV. During data collection, the participant was asked to first

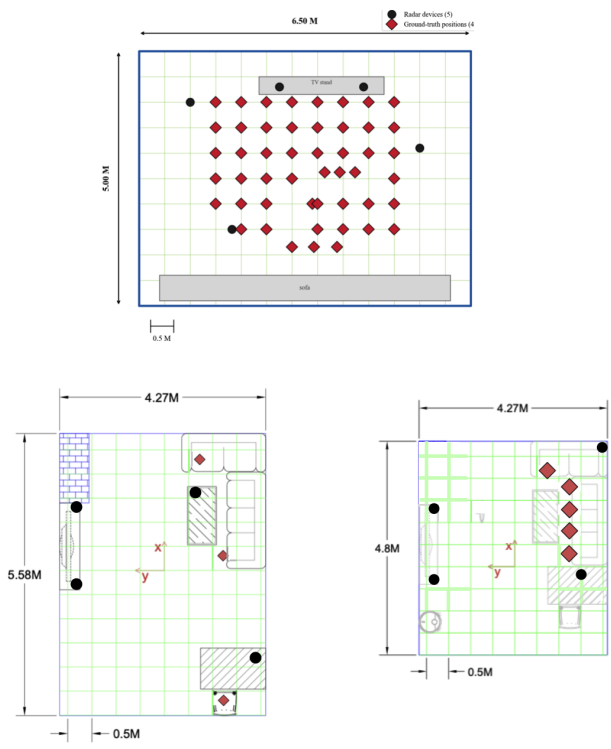


Fig. 6: Three Room setups (Black dots are the smart home devices. Red diamond are user locations.)

stand up and then sit down on the sofa. Then he/she was instructed to watch TV and/or enjoy snacks and drinks at their own pace. This is to make sure their behavior stays as

normal as possible so that the data collected is a close match to an ordinary smart home environment. Each capture lasts 120 seconds. This process repeated until all positions were collected. All devices were controlled by one single computer to ensure synchronization.

B. Data Augmentation

In order to improve the data quality, a couple of augmentation techniques were used during training. First, when generating training samples, the window slides 2 frames every time, effectively having 60% overlap between consecutive frames. Secondly, zero-mean Gaussian noise is added to the post-FFT radar data and the ultrasound echogram to diversify the data. Third and more importantly, the Wi-Fi radar and ultrasound signals have very different dynamic ranges; a normalization is applied on both before being used for inference/training.

C. Model Validation Results

The model was trained on a dataset with 12 hours of sensor data recordings. The dataset was broken down into training and validation sets based on location. Data from the same location will not show up in both training and validation sets. This is to avoid data leakage and overfitting because signals from the same location are generally similar. The validation error reduced steadily during training and the performance is at the same level as the training dataset, proof that the model is generalizing the pattern well. Also, in Fig. 7, noise-based data augmentation is turned off to show that the final values of training and validation converge to the same level.

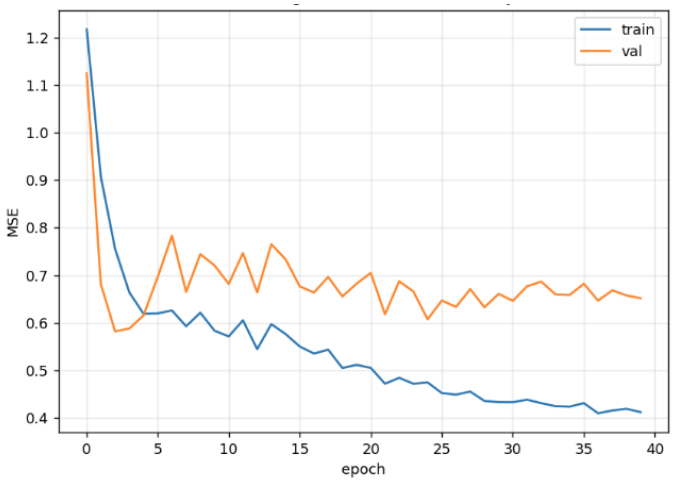


Fig. 7: Training and validation error vs. training epochs for proposed model without noise data augmentation

In Fig. 8, accuracy error comparison between four algorithms is presented. The first boxplot is the Wi-Fi radar plot. A basic peak finding algorithm mentioned in Sec. III-D2 is then applied. The second box is ultrasound signal based and a similar peak finding algorithm is applied on the echogram. The third box is a Wiener filter-based fusion using prior knowledge about the sensors [46]. The last box is our proposed range model. Our approach clearly performs better than all the other comparisons. Classic signal processing methods can produce reasonable results, but they are not enough when it comes to discovering features and correlations in large amounts of data.

TABLE I: User Location Accuracy Comparison Summary

method	MAE (m)	median (m)	$< 1m\%$
Wi-Fi radar peak	2.045	2.214	0.0
USPS peak	1.803	1.674	0.0
Wiener fusion (range)	1.742	1.668	0.0
ML distance model	1.225	1.128	49.9

TABLE II: Training and Testing Performance Comparison for End-to-End Transformer Model

Method	Training (Rm1)	Testing MAE (Rm2)
End-to-End Transformer	1.05m	2.50m

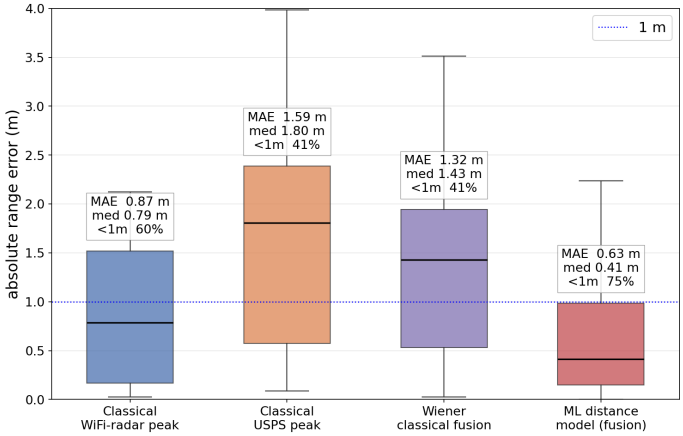


Fig. 8: Range estimation error comparison between different methods

If we pipe the range estimation results from Fig. 8 through the trilateration method mentioned in Sec. II-E, we can derive the overall performance of Fig. 2 shown in Tab. I. The errors follow the same trend as the range estimation; this is intuitive because for the trilateration algorithm we applied, the device map is ground truth, so the only thing that influences the user localization accuracy is the range estimation error.

The baseline model mentioned in Sec. III-D1 is another obvious choice for machine learning-based methods; we built a similar-sized model based on it and summarized the results in Tab. II. This model fits the training data reasonably well but failed to generalize the pattern to other rooms. The main reason is the model has to learn the relationship between the signal and device location from scratch, rather than using the physics-based multilateration algorithms. This is too difficult to learn given the model size. The model memorized room-specific information rather than deriving features from the received signals.

D. Hyper Parameter Tuning

One of the most important trade-offs in model design is complexity/power consumption vs. accuracy, either in vision transformers [47] or in speech recognition [48]. In this section,

TABLE III: Model Configuration Summary

Config	Params	Embedding	heads	layers	Fusion emb. dim.	Cross-attention	Train	Val	Gap	MAE	Median
small (proposal)	46,409	32/24	2	1	32	1	0.7031	0.5220	1.35	0.8289	0.6782
small-2layer	59,825	32/24	2	2	32	1	0.6399	0.6159	1.04	0.8406	0.6240
medium	93,105	48/32	4	1	48	1	0.3294	0.5935	0.55	0.7175	0.7051
large	269,345	64/48	4	2	64	2	0.3735	0.6242	0.60	0.7614	0.6078

Params is the total trainable parameter count. Embedding is the attention width inside encoder of that branch, the two values are given as radar/ultrasound. Heads is the number of attention heads. Layers is the depth of each per-modality encoder. Fusion emb. dim. is the common width of the tokens, which those two branches are projected to. Cross-attention is the number of decoder layers.

five hyperparameters are examined: embedding dimensions, number of attention heads, number of layers of encoders, fusion embedding width, and cross-attention depth. The main performance index is the mean absolute error during training and validation. The results are summarized in Tab. III. The medium configuration achieves the lowest MAE, while the large configuration begins to overfit the training data despite having roughly three times more parameters. The small (proposal) configuration used throughout this paper offers a favorable trade-off between accuracy and model size, which matters for on-device deployment. It is an application-dependent decision about which set of parameters is the best. Therefore, we focus on presenting some typical configurations and leave it for the downstream applications to determine the optimal configuration.

V. Conclusion and Future Work

In this paper, we investigated fusion-based localization problems in challenging environments. A novel transformer-based machine learning model is proposed. Its accuracy is presented and compared with traditional methods. Also, hyperparameters are tuned to optimize the trade-off between accuracy and model complexity. The model showed great improvements over past methods. Note that all the data was collected from Wi-Fi radar, microphone, and speaker on smart speakers, a set of zero-cost sensing solutions. This system architecture is a great choice for the consumer electronics market, given that most customers are cost-sensitive.

Future work will focus on generalizing the results to more challenging environments such as non-line-of-sight or multi-user scenarios. The quality of the signal will degrade in those settings; however, this architecture can be extended by combining more sensor data. Also, tuning the loss function of the model is another interesting direction.

References

[1] Hu, Xuhui, et al. "Stereopilot: A wearable target location system for blind and visually impaired using spatial audio rendering." IEEE Transactions on Neural Systems and Rehabilitation Engineering 30 (2022): 1621-1630.

[2] Barber, Jarred, Yifeng Fan, and Tao Zhang. "End-to-end alexa device arbitration." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[3] Boulos, Maged N. Kamel, et al. "CAALYX: a new generation of location-based services in healthcare." International journal of health geographics 6.1 (2007): 9.

[4] Hasan, Samiul, and Satish V. Ukkusuri. "Urban activity pattern classification using topic models from online geo-location data." Transportation Research Part C: Emerging Technologies 44 (2014): 363-381.

[5] Chen, Shin-Hon, and Shyh-Kang Jeng. "An SBR/image approach for radio wave propagation in indoor environments with metallic furniture." IEEE transactions on antennas and propagation 45.1 (1997): 98-106.

[6] Olive, Sean E., and Floyd E. Toole. "The detection of reflections in typical rooms." J. Audio Eng. Soc 37.7/8 (1989): 539-553.

[7] Ho, Quang-Dung, Daniel Tweed, and Tho Le-Ngoc. "Requirements and regulations in the 5 ghz unlicensed spectrum." Long term evolution in unlicensed bands. Cham: Springer International Publishing, 2016. 11-20.

[8] Babbitt, Callie W., et al. "Disassembly-based bill of materials data for consumer electronic products." Scientific data 7.1 (2020): 251.

[9] K Ngamakeur, S Yongchareon, J Yu, S. U. Rehman "A survey on device-free indoor localization and tracking in the multi-resident environment," ACM Computing Surveys (CSUR). 2020 Jul 11;53(4):1-29.

[10] D Burghal, AT Ravi, V Rao, AA Alghafis, AF Molisch. "A comprehensive survey of machine learning based localization with wireless signals," arXiv preprint arXiv:2012.11171. 2020 Dec 21.

[11] S Yiu , M Dashti, H Claussen, F Perez-Cruz . "Wireless RSSI fingerprinting localization. Signal Processing," 2017 Feb 1;131:235-44.

[12] S Sadowski, P Spachos. "Rssi-based indoor localization with the internet of things," IEEE access. 2018 Jun 4;6:30149-61.

[13] J Xu, W Liu, F Lang, Y Zhang, C Wang. "Distance measurement model based on RSSI in WSN. Wireless Sensor Network," 2010 Aug 1;2(8):606.

[14] K Heurtefeux, F Valois. "Is RSSI a good choice for localization in wireless sensor network?" In2012 IEEE 26th international conference on advanced information networking and applications 2012 Mar 26 (pp. 732-739). IEEE.

[15] Li, Wenda, et al. "A taxonomy of WiFi sensing: CSI vs passive WiFi radar." 2020 IEEE Globecom Workshops (GC Wkshps). IEEE, 2020.

[16] T Grydeland, FD Lind, PJ Erickson, JM Holt. "Software radar signal processing," InAnnales Geophysicae 2005 Jan 31 (Vol. 23, No. 1, pp. 109-121). Göttingen, Germany: Copernicus Publications.

[17] Nakatani, Sara, Mondher Bouazizi, and Tomoaki Ohtsuki. "A Novel MIMO FMCW Radar-Based Approach for Heart Rate Estimation Using Positional Feature Selection." ICC 2025 - IEEE International Conference on Communications, 2025.

[18] W Li, RJ Piechocki, K Woodbridge, C Tang, K Chetty. "Passive WiFi radar for human sensing using a stand-alone access point," IEEE Transactions on Geoscience and Remote Sensing. 2020 Jul 14;59(3):1986-98.

[19] Cobos, Maximo, et al. "A survey of sound source localization methods in wireless acoustic sensor networks." Wireless Communications and Mobile Computing 2017.1 (2017): 3956282.

[20] Liu, Manni, et al. "Indoor acoustic localization: A survey." Human-centric Computing and Information Sciences 10.1 (2020): 2.

[21] Hassan, Farrukh, et al. "State-of-the-art review on the acoustic emission source localization techniques." Ieee Access 9 (2021): 101246-101266.

[22] Couture, Olivier, et al. "Ultrasound localization microscopy and super-resolution: A state of the art." IEEE transactions on ultrasonics, ferroelectrics, and frequency control 65.8 (2018): 1304-1320.

[23] Errico, Claudia, et al. "Ultrafast ultrasound localization microscopy for deep super-resolution vascular imaging." Nature 527.7579 (2015): 499-502.

[24] Misra, Prasant, Sanjay Jha, and Diethelm Ostry. "Improving the coverage range of ultrasound-based localization systems." 2011 IEEE Wireless Communications and Networking Conference. IEEE, 2011.

[25] Yang, Lu, et al. "Large-scale person detection and localization using overhead fisheye cameras." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[26] Guzov, Vladimir, et al. "Human poseitioning system (hps): 3d human pose estimation and self-localization in large scenes from body-mounted

sensors." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[27] Huang, Zexia, et al. "Application of multi-sensor fusion localization algorithm based on recurrent neural networks." Scientific reports 15.1 (2025): 8195.
[28] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[29] Q Motger, A Miaschi, F Dell'Orletta, X Franch, J Marco. "T-frex: A transformer-based feature extraction method from mobile app reviews," In2024 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER) 2024 Mar 12 (pp. 227-238). IEEE.
[30] R Hadsell, S Chopra, Y LeCun. "Dimensionality reduction by learning an invariant mapping," In2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06) 2006 Jun 17 (Vol. 2, pp. 1735-1742). IEEE.
[31] Tian, Jingyi, Wenjun Yang, and Lin Cai. "Attention-Based Spatiotemporal Model for RTT Prediction in LEO Satellite Networks." ICC 2025 - IEEE International Conference on Communications, 2025, pp. 5059-5063.
[32] B. Park, H. Do and N. Lee, "Transformer-Based Nonlinear Transform Coding for Multi-Rate CSI Compression in MIMO-OFDM Systems," ICC 2025 - IEEE International Conference on Communications, Montreal, QC, Canada, 2025, pp. 2327-2333, doi: 10.1109/ICC52391.2025.11161672.
[33] X. Xu, H. Xu, H. Yu, Y. Liu and M. Chen, "Fluid Antenna System (FAS)-Assisted 3D UAV Positioning Performance Optimization," ICC 2025 - IEEE International Conference on Communications, Montreal, QC, Canada, 2025, pp. 2260-2265, doi: 10.1109/ICC52391.2025.11161642.
[34] Hornung, Mark R., and Oliver Brand. "Proximity sensor." Micromachined Ultrasound-Based Proximity Sensors. Boston, MA: Springer US, 1999. 83-105.
[35] Fukushima, Kunihiko. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." Biological cybernetics 36.4 (1980): 193-202.
[36] LeCun, Yann, et al. "Backpropagation applied to handwritten zip code recognition." Neural computation 1.4 (1989): 541-551.
[37] H. C. Schau and A. Z. Robinson. "Passive source localization employing intersecting spherical surfaces from time-of-arrival difference" 1987 IEEE Transactions on acoustics, speech and signal processing. Vol. ASSP-35 No. 8.
[38] Khan, Asifullah, et al. "A survey of the vision transformers and their CNN-transformer based variants." Artificial Intelligence Review 56.Suppl 3 (2023): 2917-2970.
[39] Lawrence, Beatrice D., and James A. Simmons. "Measurements of atmospheric attenuation at ultrasonic frequencies and the significance for echolocation by bats." The Journal of the Acoustical Society of America 71.3 (1982): 585-590.
[40] Bamber, J. C. "Attenuation and absorption." Physical principles of medical ultrasonics 2 (1986): 93-166.
[41] Guan, Hongtao, and D. D. L. Chung. "Absorption-dominant radio-wave attenuation loss of metals and graphite." Journal of Materials Science 56.13 (2021): 8037-8047.
[42] Min, Se Dong, et al. "Noncontact respiration rate measurement system using an ultrasonic proximity sensor." IEEE sensors journal 10.11 (2010): 1732-1739.
[43] Gotmare, Akhilesh, et al. "A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation." arXiv preprint arXiv:1810.13243 (2018).
[44] Badi, Mahmoud, et al. "Experimentally analyzing diverse antenna placements and orientations for UAV communications." IEEE Transactions on Vehicular Technology 69.12 (2020): 14989-15004.
[45] Awan, Shaheen N., et al. "The effect of microphone frequency response on spectral and cepstral measures of voice: An examination of low-cost electret headset microphones." American Journal of Speech-Language Pathology 31.2 (2022): 959-973.
[46] Didier, Paul, et al. "Distributed multichannel wiener filtering for wireless acoustic sensor networks." IEEE Transactions on Audio, Speech and Language Processing (2026).
[47] Naseeb Singh, V.K. Tewari, P.K. Biswas, "Vision transformers for cotton boll segmentation: Hyperparameters optimization and comparison with convolutional neural networks", Industrial Crops and Products, Volume 223, 2025, 120241, ISSN 0926-6690,
[48] Gumelar, Agustinus Bimo, et al. "Transformer-CNN automatic hyperparameter tuning for speech emotion recognition." 2022 IEEE international conference on imaging systems and techniques (IST). IEEE, 2022.