Beyond Convolutions: Transformer Networks for Improved UWB CIR-based Fingerprinting

Dieter Coppens, Adnan Shahid, and Eli De Poorter
INTEC-IDLab, Ghent University-IMEC Ghent, Belgium
Email: dieter.coppens@ugent.be

***Abstract*—Indoor positioning using UWB has gained popularity due to its low cost while still providing centimeter-level accuracy. Currently, Convolutional Neural Network (CNN)-based approaches are often proposed for NLOS detection, error correction, etc. to make these UWB positioning systems more accurate. Transformer (TF) networks have shown to be a more capable alternative in several other domains, but have not been used for UWB fingerprinting. We present two novel TF-based approaches: one that processes channel impulse responses (CIR) directly, and a second one that uses cross-attention to incorporate anchor position information to improve geometric understanding. Moreover, we propose a second innovation by combining fingerprinting with a time calibration method that synchronizes the CIR data using a TDOA-based setup. This second innovation can be used with our novel approach, or with other state-of-the-art fingerprinting methods. The proposed models are evaluated in an industrial environment and outperform previous state-of-the-art CNNs in both LOS and NLOS situations, reaching accuracies with errors as low as 3cm in real-life conditions, while having lower complexity and requiring fewer samples.**

I. Introduction

Indoor positioning has gained significant interest due to its potential in several Internet of Things (IoT) applications such as assistive healthcare systems, sports tracking, smart logistics, etc. This positioning is often realized with Wi-Fi and Bluetooth Low Energy (BLE) as they are widely available. They use received signal strength indicators (RSSIs) to estimate the position, resulting in a low accuracy of 2-3 meters, mostly due to multipath fading and non-line of sight (NLOS) [1].

Ultra-wideband (UWB) technology has become an interesting alternative. This is due to (1) its integration into multiple smartphones and consumer products and (2) its centimeter-level accuracy, achieved through the high temporal resolution caused by the high bandwidth, a significant advantage over competing technologies [2]. UWB systems can accurately estimate the time-of-arrival (TOA) of a signal enabling multiple techniques for position estimation such as Two-Way Ranging (TWR), Time Difference of Arrival (TDoA), and Channel Impulse Response (CIR)-based fingerprinting. Each technique has its advantages and disadvantages. TWR ensures accurate localization without clock synchronization between tag and anchor but requires three packets sent between each tag-anchor pair, limiting scalability and update rates. In TDOA, the tag only needs to send one packet which all anchors receive, however, this requires clock synchronization between the anchors, and positioning error increases when NLOS conditions are present [3]. Fingerprinting does not require synchronization and only relies on the CIR information received at the anchors, thereby retaining the small energy overhead of sending only a single packet (like TDoA) but removing the complexity of multi-hop TDoA synchronization. Deep learning has been shown to enable accurate UWB fingerprinting (and by extension TWR and TDOA), with convolutional neural networks (CNN) being the most effective and widely used architecture. Understandably so, with the proven track record as the go-to architecture for time-series models.

Influenced by the exceptional results of the transformer (TF) architecture in Natural Language Processing (NLP), computer vision, and various other domains, we present a novel CIR-based fingerprinting architecture based on the TF-encoder and show the importance of relative time synchronization of the different anchors using calibration resulting in a hybrid fingerprinting-TDOA system with improved accuracy and less required samples.

The main contributions of this work are:

- We propose the design of a transformer-based UWB CIR fingerprinting method that significantly outperforms commonly used CNN approaches. Furthermore, we improve the design by introducing a cross-attention architecture for the fusion of anchor and CIR information for better geometric understanding.
- We demonstrate that, similar to TDOA, the loss of time synchronization between anchor nodes lowers the accuracy of fingerprinting solutions and we introduce a low-complexity time-calibration method for fingerprinting to solve this problem.
- Finally, we evaluate the system with real UWB data in a realistic NLOS environment, focusing on accuracy, complexity, and sample requirements.

The remainder of this paper is structured as follows: Section II provides an overview of the related work, Section III presents the system model, Section IV discusses the datasets, Section V presents the proposed architectures, experimental results are analyzed in SectionVI, followed by a conclusion in Section VII.

II. Related work

The related work for UWB fingerprinting is summarized in Table I. In [4], TWR and fingerprinting techniques are combined, this has the downside that many packets are required

Part of this research has received funding from the FWO project PESSO with grant number G018522N

TABLE I: Related work for UWB fingerprinting showing the differences in input data, outputs, and machine learning models

Paper	Input	Output	Model	Time Cal.	LOS Acc.	NLOS Acc.
[4]	Ranges	2D	kNN		6-23 cm	/
[5]	CIR	Room	DNN		99%	99%
[6]	CIR	2D	DNN		100 cm	100cm
[7]	CIR	2D	CNN		20.9 cm	80.7cm
[8]	CIR	2D	CNN	✓	17 cm	29.2 cm
This	**CIR**	**2D**	**TF**	✓	3 cm	16cm

to estimate one 2D position, limiting scalability. The authors of [5] use a deep neural network (DNN) to predict in which room the user is present, instead of actual 2D positions, while also requiring ranges and thus many transmissions. [6] was one of the first scientific publications to perform CIR-based UWB fingerprinting with a Neural Network and can achieve an MAE of around 1m. At the time of writing this paper, recent state-of-the-art UWB fingerprinting papers that use the CIR directly all employ the CNN architecture. The authors of [7] propose a UWB fingerprinting positioning method using multiple CNNs trained on CIRs of anchor subset combinations to handle cases where not all CIRs arrive at the anchors (in NLOS situations). The method reaches a MAE of 21-87 cm but requires separate models for all anchor subsets, our approach handles it by zero-padding in the fingerprint. Similar to the work presented in this research, [8] uses CIRs for UWB fingerprinting in a TDoA-based system to calibrate the CIRs in the fingerprint. The research evaluates several different and well-known CNN architectures and achieves a MAE of 12-36 cm. None of the previous works have incorporated the TF architecture for using raw CIR signals from UWB devices to predict 2D positions. Transformer-based architectures have been shown to excel in capturing long-range dependencies and global context. Furthermore, they have been shown to perform well in using CIR information for multipath component (MPC) delay estimation [9]. These factors indicate that utilizing a TF-based approach could improve the current state-of-the-art in UWB fingerprinting.

III. System model

A. Pure fingerprinting

A UWB fingerprinting system includes N fixed anchors a_n for $n \in [1 \dots N]$ and a mobile tag t. In a fingerprinting setup, only one UWB packet is broadcasted from the tag and the received CIRs are used to get a location estimate. This setup has the most straightforward configuration, permitting the tag to enter sleep mode immediately after sending the packet. This localization technique is therefore excellent for low-power devices with years of lifetime. However, the drawback is that the approach loses timing information across the individual CIRs collected at the anchors.

B. TDOA-based time calibration system

If the positioning system requires better accuracy and allows for additional complexity then time synchronization can be introduced in the fingerprints by using a TDOA setup where each anchor a_n additionally saves a timestamp of its first path T^n_{fp} and the index of the first path in the recorded CIR i^{fp}_n both estimated by a leading edge detection algorithm (a threshold method) provided with the DW1000 UWB IC used in this research [10]. The T^n_{fp} will be translated to a common timeframe to compensate for clock drift errors between the anchors. A predefined reference anchor a_{ref} periodically sends a UWB packet containing its transmission time $T^{ref,tx}_j$, with j being the time index of the packet, to all other anchors which calculate their relative clock skew ρ:

$$\rho = \frac{T^{ref,tx}_j - T^{ref,tx}_{j-1}}{T^{ref,rx}_j - T^{ref,rx}_{j-1}}. \tag{1}$$

Where, tx indicates transmission and rx reception. This enables an anchor a_n to synchronize its timestamps with a_{ref} from the following expression:

$$T^n_{j,fp} = T^{ref,rx}_j + \rho \cdot (T^n_{j,fp} - T^{ref,rx}_{j-1}). \tag{2}$$

Finally, to compensate for the time the signal needs to propagate between a_{ref} and a_n, we determine Δ_n which is either calculated using an initial TWR protocol (if the anchor positions are unknown) or is estimated using the ground truth distance between them combined with the speed of light.

C. Channel Impulse Response (CIR)

The CIR can be defined as:

$$CIR_{a_n}(t) = \sum_{s=1}^{S} \alpha_s \delta(t - \tau_s) + n(t). \tag{3}$$

where t is the time-step for each CIR sample, S is the number of multipath components, α_s and τ_s are the amplitude and time delay of the sth multipath component, and $n(t)$ is the channel's additive noise. The multipath and NLOS conditions can cause large errors in estimated first paths and lead to severe errors in TDOA positioning. In our solution, these timings are only used for calibration.

IV. Datasets

To evaluate the proposed method, a dataset was recorded in an industrial environment, part of the IIoT testbed [11], with metal racks as obstacles. Anchors were deployed both in open and NLOS locations. The total size of this environment is approximately 30m × 10m. The measurements were performed with Wi-Pos devices [12], a platform developed for data collection with a wireless sub-GHz backbone combined with UWB based on the popular DW1000. The environment is equipped with 18 Qualisys Miqus M3 Motion Capture (MOCAP) cameras with a quantified uncertainty in the millimeter range at speeds up to 340 Hz, enabling accurate ground truth determination. In addition, a mobile robotic platform was used to drive repeatable trajectories through the lab. During the data collection, 4 different trajectories were captured with a mobile robot driving through the lab. The trajectories are visualized in Figure 1. There are three datasets in LOS conditions but the number of fingerprints for NLOS and LOS is roughly the same. The details are described in Table II.

TABLE II: Details of the 4 collected datasets. The "Racks" dataset has the most fingerprints, many being NLOS.

Name	# Anchors	# Fingerprints	Environment
Racks	16	15505	NLOS
Random	16	3268	LOS
Grid	8	6114	LOS
Tour	8	7697	LOS

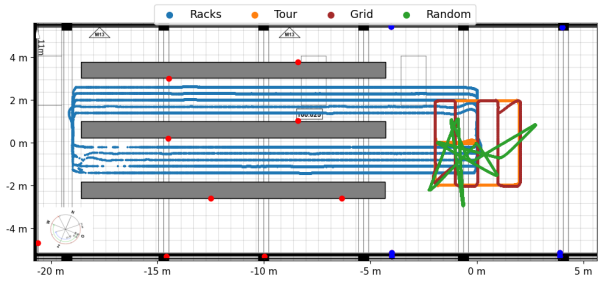


Fig. 1: The trajectories of the 4 different datasets (blue, orange, red, and green lines) for UWB fingerprinting cover both LOS and NLOS scenarios. The blue dots are anchors used in all 4 datasets, the red dots are only used in the "random" and "racks" datasets.

A. Data preprocessing

First, we process the raw CIR data. The IQ-sampled array in a complex form is converted to an amplitude array. Next, the CIR is min-max normalized. For pure fingerprinting, the CIR is now trimmed to 150 samples, 50 samples before and 100 samples after the first path i_n^{fp}.

TDOA-based time calibration uses the higher resolution of 15.65 ps of i_n^{fp} that is used internally in the DW1000. But since the CIR uses a resolution of 1.0016 ns, we apply a quadratic interpolation filter, thereby increasing the resolution by 64, to match that of i_n^{fp}. Each CIR is now trimmed by selecting a time-calibrated window starting around $T_{n,cal} = (T_{j,fp}^n - T_{j,min} + \Delta_n)$, with $T_{j,min}$ the lowest received timestamp of all anchors, due to the time-synchronization in this system this timestamp is common across all anchors. We then position and trim each CIR to the correct length relative to this. The time-calibrated window shifts the relative time of the first path $T_{j,fp}^n$ to a common timing across all anchors. Finally, the $CIR_{a_n}(t)$ is down-sampled to get the 150 samples at the original resolution.

$$CIR_{a_n}(t) = CIR_{a_n}\left[T_{n,cal} - 50 : T_{n,cal} + 100\right]. \quad (4)$$

All processed CIRs are positioned in a $N \times 150$ matrix to get to the final fingerprint. Anchors with a failing UWB link are padded with zeros. The difference between a calibrated and uncalibrated fingerprint matrix is illustrated in Figure 2, the estimated first path of all CIRs is aligned at x-position 50 and all timing is only relevant for each separate anchor. Calibrating the CIRs means that the position of the first path is related to the distance between the tag and the anchor.

V. Methodology

In this section, we first discuss the transformer concept and then introduce the two proposed transformer architectures.

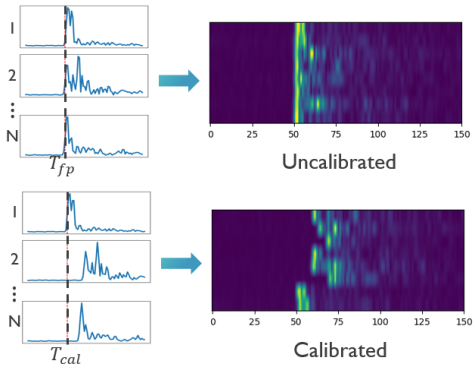


Fig. 2: Illustration of the difference between a calibrated and uncalibrated matrix of fingerprints (N anchors, 150 samples per fingerprint)

A. Transformers and attention

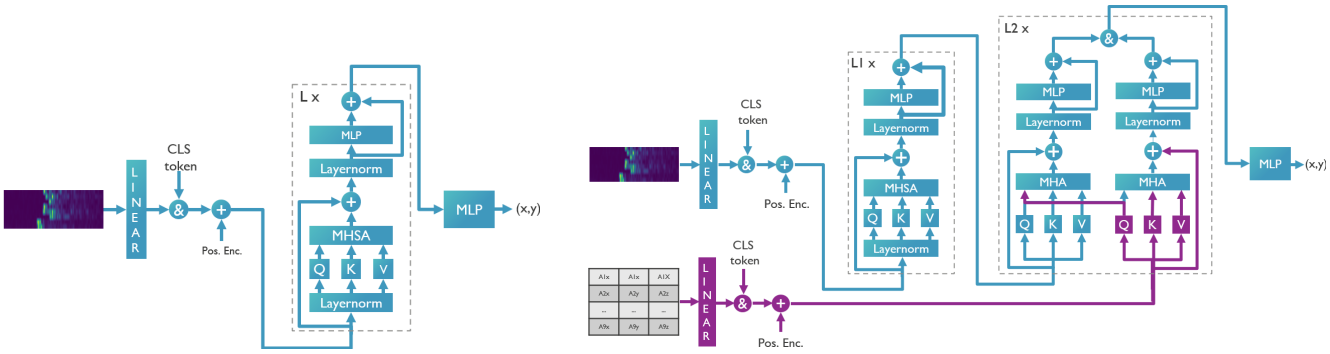
TFs have emerged as an alternative to CNNs. While CNNs effectively extract local features, TFs use a self-attention mechanism to capture long-range dependencies and global context within the input data. They allow the model to focus on parts of the input dynamically. Attention layers compute an attention score for each component of the result to each input component without the locality constraints of convolutional layers. An attention layer starts by generating the Query (Q), Key (K), and Value (V) with a linear layer of the same dimension h as the input. The attention weights A are the dot-product of Q and K followed by *scaling* and *softmax*. In a sense, this is a similarity that can be seen as a mask to be applied to V. By multiplying A and V, we get a weighted representation which can be implemented as:

$$\text{att}(Q,K,V) = \underbrace{\text{softmax}\left(\frac{QK^T}{\sqrt{h}}\right)}_{A} V. \quad (5)$$

A multi-headed attention layer a the number of heads H. Q, K, and V are split into h parts each passing through the scaled dot product attention independently. In this paper, we only use the TFs encoder [13], the original architecture uses an encoder-decoder structure. The decoder is used to construct an output sequence, since we consider positioning (a regression task) only the encoder is necessary and the decoder can be replaced with a feed-forward network to determine the 2d position. The encoder consists of a stack of identical layers, each composed of a multi-headed self-attention mechanism and a multilayer perceptron (MLP). Residual connections are added after each sub-layer and layer normalization is applied at each sub-layer output.

B. CIR-fingerprint TF architecture

Our first design is inspired by the original paper describing a TF architecture [13] which first tokenizes and embeds the input sequence. As CIR values are continuous, tokenization is not needed. However, we employ learned embeddings using a linear layer that produces a d-dimensional vector for each



(a) CIR-fingerprint TF (cir-TF)

(b) CIR-anchors cross-attention TF (cross-TF)

Fig. 3: The two proposed transformer architectures for CIR-based UWB fingerprinting: cir-TF and cross-TF.

CIR (row in the fingerprint) with $d < 150$. The embeddings can capture the information of the CIR values in a continuous vector space and reduce the number of parameters, to improve computational efficiency. Next, a class (CLS) token row is added, resulting in a $(N + 1) \times d$ tensor. This row represents the entire input sequence and after passing the complete input through the encoder, the row can be used as the accumulated representation. It contains all the information for the MLP to determine the 2d position. This has the benefit of allowing the encoder to learn a fixed size representation independent of the number of anchors N, allowing compatibility if a different number of anchors is used. Next, we add learnable positional encoding, these learnable values are summed with the values already present in the tensor. This provides the model with information about the order of the values in the CIR sequences, otherwise the relationship among the values is lost. Now, the encoding is performed by a stack of L identical encoder layers with multi-headed self-attention as discussed in V-A. After the encoder layers, the CLS token row is extracted from the output and passed through the MLP which consists of two fully connected layers, the final one with Tanh activation. This activation scales the output between [-1,1], this value is then multiplied by 20,000 mm and 5,000 mm to scale the x and y values to the environment size.

C. CIR-anchors cross-attention TF architecture

Our second architecture design is an extension of the previous one. The CIR fingerprint is processed the same. Afterward, an additional double TF encoder layer together with the anchor positions that have first been embedded, adding a CLS token row and positional encoding. This ensures that the anchor position sequences have the same dimension. In the second encoder, these sequences are processed using cross-attention. Each sequence calculates its Q, K, and V but the Q values are exchanged. The Q from the anchor positions is compared with the keys from the CIR fingerprint sequence resulting in an attention score that is used to aggregate the values from the CIR fingerprint sequence. The opposite happens in the other encoder. This approach allows each encoder to attend to different aspects of the input sequences and find dependencies and relationships between different parts of the input sequences. This allows the model to understand the relationship between the received CIRs and the positions where it was received enabling better geometric understanding, which should be most helpful in generalization.

VI. Analysis and Experimental Results

We analyze the proposed architectures using two different train/test split tests: 1) a traditional approach to show the influence of the calibration and 2) a custom split for better analyzing the generalization of the proposed architectures. Furthermore, the different architectures are compared based on their performance relative to the number of parameters. Finally, the performance based on the number of training samples is analyzed. Unless otherwise specified, the following parameters are used during the training of the architectures: the embedding dimension d is 128, the number of encoders for the CIR-specific TF is 4 (L and L1 in Figure 3) and 2 for the Cross-attention block (L2), the number of heads of all multi-headed attention layers is 4 and GELU activation [14] is used in the encoder layers. The Adam optimizer with a learning rate of 0.001 and a training batch size of 64 is used with mean squared error (MSE) loss.

A. Baselines and metrics

For performance evaluation, we will use two metrics (1) the MAE (of the Euclidean distance) as it encapsulates the performance in a single value and (2) the circular error probability (CEP), defined as the radius of a circle, centered around the ground truth, where 50%, or 95% for CEP95, of the estimated positions are within. To get a baseline positioning performance we use a TDOA-positioning algorithm using least-squares optimization to estimate the tag's position. As the baseline for CNNs, we use the GoogLeNet (G-Re-NoP) and Small-Net implementations from [8]. The reason for selecting models from this paper is that it uses large CNNs for fingerprinting in a similar setup as we describe in this paper, the authors also define a (similar to ours, but requiring additional hardware) calibration of the CIRs. The G-Re-NoP model was selected as it is the best-performing model from the paper. Small-Net was selected to compare the performance of our architectures in terms of the number of parameters with a smaller CNN that still shows

good performance. We have replicated these models and have optimized their performance to the best of our capabilities.

B. Traditional Evaluation

As a first evaluation, we use the typical method of randomly splitting the datasets into 80% training and 20% testing. We refer to the TF architecture using only the CIR fingerprint and the architecture using cross-attention with the anchor positions as **cir-tf** and **cross-tf** respectively. The three datasets in predominantly LOS conditions are grouped in the 'LOS' environment, the 'Racks' environment is the only 'NLOS' environment. Table III contains the results of the standard evaluation. Firstly, looking at the Calibrated MAE, it can be seen that the cir-tf model has the highest accuracy (lowest MAE) for all categories, followed by the cross-tf model. The CNN models from prior publications show considerably higher MAE, these results are in line with the results reported in [8]. Compared to traditional TDOA positioning, all deep learning models vastly improve the accuracy in the NLOS environment. In LOS situations only the TF-based have a distinct improvement in accuracy. The uncalibrated MAE column shows the results when the input is not calibrated as discussed in IV-A. The results show that calibration has a strong positive impact on accuracy.

TABLE III: Accuracy evaluation of our 2 proposed transformer-based approaches (**bold**) compared to state-of-the-art approaches. Calibration of the input increases is shown to increase the fingerprinting performance of both our approach and of previously proposed solutions.

Model	LOS		NLOS	
	Calibrated MAE (cm)	**Uncalibrated MAE** (cm)	**Calibrated MAE** (cm)	**Uncalibrated MAE** (cm)
TDOA	17.2		107.5	
Small-Net	14.2	27.4	31.9	34.7.5
G-Re-NoP	12.2	34.4	34.9	50.8
Cir-tf	2.6	4.1	15.7	17.0
Cross-tf	5.6	7.9	18.0	21.5

C. Generalization evaluation

While randomly sampling training data is a popular strategy, in this case, it does not provide adequate evidence of the model's generalization. This is illustrated in Figure 4a, many data points are in similar positions because random sampling spreads them out over the complete trajectory. This results in many test samples being similar to training samples. We can mitigate this effect by sampling random patches with many consecutive samples in it, as shown in Figure 4b. The results for this evaluation are given in Table IV. The errors for this evaluation are higher showing that they were overfitting in the standard evaluation. However, it paints a similar picture. The TF-based models generally outperform the CNN models in terms of MAE and CEP. Before, cir-tf was the best-performing model but here it is cross-tf. It has lower MAE and CEP50 but slightly higher CEP95 indicating that the difference with the cir-tf model is small. This could be the influence of the additional geometric understanding due to the additional anchor positions.

TABLE IV: Generalization evaluation of the proposed TF approaches (**bold**) show better performance in LOS and NLOS compared to CNNs.

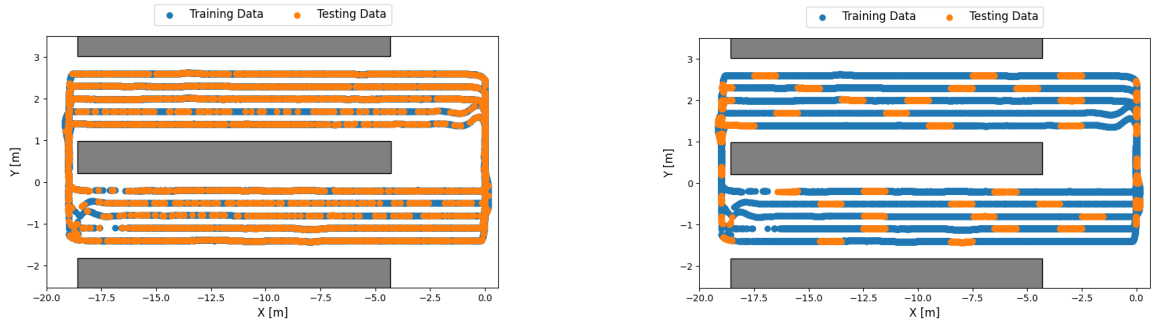
Model	LOS			NLOS		
	MAE (cm)	**CEP50** (cm)	**CEP95** (cm)	**MAE** (cm)	**CEP50** (cm)	**CEP95** (cm)
Small-Net	29.2	24.3	67.6	71.5	45.1	231.5
G-Re-NoP	32.7	29.0	75.2	69.0	39.7	223.9
Cir-tf	17.5	13.0	38.8	46.0	26.1	138.1
Cross-tf	15.1	11.6	36.2	44.2	25.9	141.8

D. Model complexity analysis

Understanding the relationship between the number of parameters in a model and its performance is valuable as it shows the trade-off between complexity and accuracy. Based on this, a model can be selected for a specific deployment. Figure 5 contains the results for the NLOS evaluation, as detailed before in the generalization section. However, for the Cross-tf and Cir-tf models, a parameter sweep was performed, including varying embedding dimension $d \in [16, 32, 64, 128]$ and the number of encoder layers for the CIR-specific TF $L, L1 \in [2, 4, 6, 8]$. The SmallNet and G-Re-NoP models are represented as single points, as they were not part of the sweep. Figure 5 shows that increasing d and $L, L1$ generally improves the performance. An exception to this is for $d = 128$, here increasing $L, L1$ above 4, seems to negatively impact the performance. This has multiple potential causes: overfitting, increased training difficulty, regularization problems, etc. The figure illustrates the differences between the cir-tf and cross-tf models. When low computational complexity is the highest priority the cir-tf model is the best option as it provides relatively good performance for the lowest number of parameters. When the generalization performance is of the highest priority, the cross-tf model can be used but at the cost of higher complexity. The G-Re-NoP and SmallNet models provide baseline performance values and show the improvement of going from CNNs to our proposed TF-based models. The TF-based models have higher accuracy for considerably fewer parameters and thus complexity. Making them the better option in all scenarios.

E. Required number of training samples

A model that achieves high accuracy with fewer samples can be deployed more quickly and at a lower cost. This allows for better adaptability to new or changing environments, and ultimately, more rapid realization of the model's benefits. In Figure 6, the achieved MAE in the number of randomly sampled training samples is displayed. The two TF-based models show the sharpest decrease in MAE for increasing samples, the CNN models converge more slowly. The G-Re-NoP converges slower than the SmallNet, probably because it is a larger model and requires more samples to reach its optimal accuracy. These results highlight that TF-based models not only achieve the lowest MAE but also require fewer samples to perform well, making them more practical for deployment.



(a) (b)

Fig. 4: Visualization of the train and test split used in the standard evaluation (a) where the samples are split randomly, showing the need for (b), the generalization evaluation where big patches are sampled as test data.

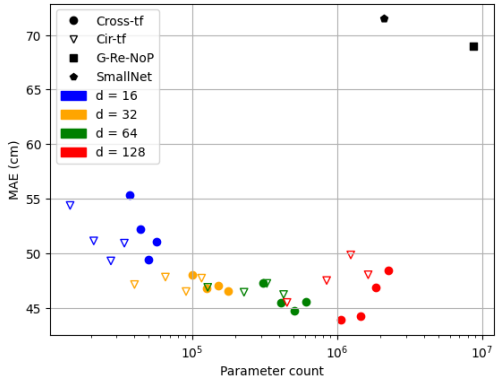


Fig. 5: The MAE vs parameter count. Our proposed solutions (circles and triangles) require significantly fewer parameters for higher accuracy. Each embedding dimension (d) has been evaluated using different numbers of encoder layers.

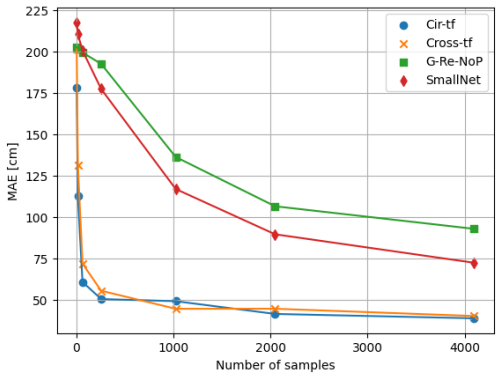


Fig. 6: The MAE in terms of the number of samples (4, 16, 64, 256, 1024, 2048, and 4096) shows that TFs outperform CNNs while requiring fewer samples.

VII. Conclusion

This paper proposes using transformer networks for CIR-based position estimation in a TDOA-based UWB system. A detailed explanation of the preprocessing is included together with an analysis showing the benefits of using a calibration step to get higher localization accuracies. The performance and efficiency of our approaches are shown in real-world environments highlighting the increased performance compared to a state-of-the-art CNN approach. These results show that TFs are well-suited for indoor localization and can provide accuracies of up to 15.1 cm in LOS and 44.2 cm in complex NLOS environments while requiring significantly fewer parameters and training samples than CNNs.

References

[1] F. Zafari, A. Gkelias, and K. K. Leung, "A survey of indoor localization systems and technologies," *IEEE Commun. Surv. Tutor.*, vol. 21, no. 3, pp. 2568–2599, 2019.

[2] D. Coppens, E. De Poorter, A. Shahid *et al.*, "An overview of uwb standards and organizations: Interoperability and research directions," *IEEE Access*, 2022.

[3] B. Van Herbruggen, J. Fontaine, and E. De Poorter, "Anchor pair selection for error correction in tdoa uwb positioning," in *Int. Conf. on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2021, pp. 1–8.

[4] S. Djosic, I. Stojanovic, M. Jovanovic *et al.*, "Fingerprinting-assisted uwb-based localization for complex indoor environments," *Expert Syst. Appl.*, vol. 167, p. 114188, 2021.

[5] M. J. Bocus, J. Paulavičius, R. McConville *et al.*, "Low cost localisation in residential environments using high resolution cir info," in *GLOBECOM 2020 - IEEE Global Commun. Conf.*, 2020, pp. 1–6.

[6] N. Al Khanbashi, N. Alsindi, S. Al-Araji *et al.*, "Performance evaluation of cir based location fingerprinting," in *IEEE Int. Symp. on Personal, Indoor and Mobile Radio Commun. (PIMRC)*. IEEE, 2012, pp. 2466–2471.

[7] J. Fontaine, B. Van Herbruggen, A. Shahid *et al.*, "Uwb localization using active cir-based fingerprinting," *IEEE Commun. Lett.*, 2023.

[8] A. Niitsoo, T. Edelhäußer, and C. Mutschler, "Cnn for position estimation in tdoa-based locating systems," in *Int. Conf. on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2018, pp. 1–8.

[9] J. Ott, M. Stahlke, S. Kram *et al.*, "Multipath delay estimation in complex environments using transformer," in *Int. Conf. on Indoor Positioning and Indoor Navigation (IPIN)*. IEEE, 2023, pp. 1–6.

[10] "Decawave - dw1000 ic," accessed May, 2024. [Online]. Available: https://www.decawave.com/product/dw1000-radio-ic/.

[11] "Industrial iot lab - idlab," accessed May, 2024. [Online]. Available: https://www.ugent.be/ea/idlab/en/research/research-infrastructure/industrial-iot-lab.htm.

[12] B. Van Herbruggen, B. Jooris, J. Rossey *et al.*, "Wi-pos: A low-cost, open source uwb hardware platform with long range sub-ghz backbone," *Sensors*, vol. 19, no. 7, p. 1548, 2019.

[13] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[14] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.