

## RESEARCH ARTICLE

# Deep Attention-Based Network Combing Geometric Information for UWB Localization in Complex Indoor Environments

KUN TANG<sup>1</sup>, BO YANG<sup>2</sup>, AND KAI DING<sup>3</sup><sup>1</sup>School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China<sup>2</sup>School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China<sup>3</sup>Science and Technology on Near-Surface Detection Laboratory, Wuxi 214035, China

Corresponding author: Bo Yang (003402@nuist.edu.cn)


This work was supported in part by Jiangsu Provincial Colleges of the Natural Science General Program under Grant 22KJB510004, in part by the National Natural Science Foundation of China under Grant 52002184 and Grant 62303230, in part by the Startup Foundation for Introducing Talent of Nanjing University of Information Science and Technology (NUIST) under Grant 2021r041, in part by Jiangsu Provincial Double-Innovation Doctor Program under Grant JSSCBS20210486, and in part by the Foundation of Science and Technology on Near-Surface Detection Laboratory under Grant 6142414211404.

**ABSTRACT** Learning-based TOA-UWB localization methods have been developed rapidly in recent years and achieve state-of-the-art localization results in complex scenes. However, they still suffer from two drawbacks: 1) biased measurements with large noise are not suppressed effectively, and 2) geometric information which is important for UWB localization is not considered. Thus, we propose two twofold strategies in this paper to overcome these issues: 1) A novel deep attention-based network is proposed. In this network, we introduce the transformer encoder to learn the weights of different ranging measurements, and thus suppress the adverse impact of the biased measurements. Meanwhile, the anchor positions including the geometric information are introduced into the network by an embedding module. 2) We present a novel learning strategy to train the proposed network. This learning strategy both considers the pre-collected ground-truth and the geometric constraints of UWB sensors. Through these two strategies, large measurement noise is further suppressed, while the geometric information and constraints are also developed for the proposed network. Therefore, the localization performance is improved. We build real-world experiments in a narrow and complex indoor scene to demonstrate the advantages of our proposed method compared to the state-of-the-art learning-based method.

**INDEX TERMS** Attention mechanism, geometric information, indoor localization, UWB sensors.

## I. INTRODUCTION

Time of arrival (TOA) UWB localization system is widely used for indoor environment [1], since it can provide high-precision tag positions based on ranging measurements and pre-calibrated anchor positions in an ideal indoor environment [2]. However, in complex environments such as some narrow or crowded scenes with many obstructions,

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed Kheir .

it is challenging for TOA-UWB localization methods to achieve accurate localization results. Because non-line-of-sight (NLOS) or multipath situations in the complex scenes will cause the reduction of ranging accuracy, and thus lead to a significant decrease in TOA-UWB localization accuracy [3].

Recently, many great efforts are afforded to improve the UWB localization performance in complex environments. Some works present robust TOA-based localization methods [4], [5], [6], [7] to achieve desirable localization results

with raw ranging measurements. Some other works propose NLOS identification or ranging correction approaches [8], [9], [10] to discriminate biased measurements with large noise, or mitigate the ranging error for the localization, and thus improve the localization accuracy.

In general, these methods can be divided into two types: 1) geometric-based approaches [4], [5], [11]. These approaches fully use the geometric information between the tag and anchors to estimate the tag positions, and 2) data-driven learning-based methods [6], [7], [12]. These methods design end-to-end neural networks to estimate the tag positions or distance information. They also need to present the learning methods to train the proposed networks. Compared to the former, the latter is more suitable for complex environments, although it is usually required to collect ground-truth for training process. Because, learning-based methods can extract high-level representative spatial or temporal features from the raw UWB measurements during the training process, and compared to the geometric-based approaches, these high-level features contribute to estimate more accurate tag positions when the ranging errors are large [12].

However, the existing learning-based localization methods also suffer from some limitations. On one hand, in a complex indoor environment, the existing deep networks cannot handle the biased ranging measurements with large noise effectively, leading to notable localization errors in a complicated scene. On the other hand, most of existing deep networks as well as their training strategies only adopt the distance between estimated and ground-truth positions as the cost function to train the network, without taking into account the UWB geometric information and constraints. These issues will affect the UWB localization performance, especially in a complex scene.

In this paper, we propose a deep attention-based UWB localization network with a novel supervised training strategy to handle these issues. To be specific, we employ a transformer encoder module containing the self-attention mechanism to learn measurement-specific weights, while extract high-level features of measurements to effectively suppress the influence of inaccurate measurement for localization. Besides, anchor locations are used in the position encoding process for introducing the geometric information into the deep network. In addition, we jointly estimate corrected ranging measurement and tag positions from the designed deep network, and develop a geometric loss which introduces the geometric constraint to the supervised learning strategy via these two estimations. With the help of these strategies, the localization performance in complex scenes is improved.

The main contributions of this work are summarized as follows:

1) The self-attention mechanism is developed for the learning-based UWB localization to further restrain the negative influence of ranging measurement with large noise.

2) The geometric information and constraints are considered for the learning-based UWB localization to further improve the localization performance in complex environments.

3) The real-world experiments in a complex indoor scene are built to illustrate the advantages of our proposed method.

## II. RELATED WORK

### A. GEOMETRIC-BASED TOA-UWB LOCALIZATION METHODS

In terms of geometric-based TOA-UWB localization approaches, the tag positions are mainly computed based on ranging measurements and pre-calibrated anchor positions. The analytical approach [13] establishes equations based on each measured distance, and the distances between each anchor position and estimated tag positions. Then, the tag position can be computed through least square approach. The accuracy of this approach is reliable in line-of-sight (LOS) environment, while decreases significantly in complex scenes caused by unreliable measurements. To improve the localization accuracy in complex scenes, the optimization-based method [4] is proposed following the analytical approach. It builds an optimal function instead of the equations considering the tag positions at multiple times. Besides, a constraint based on the maximum velocity of tag is introduced into the optimal function. This velocity constraint between continuous frames contributes to suppress the localization drifts, and thus achieves better localization results compared to the analytical approach. In addition, Kalman filter (KF) is also developed for the UWB localization method [5]. In these methods, the tag position and velocity are set as the state variable, while the measured distances are used to build the measurement equation. However, the process of KF will introduce nonlinear errors due to high nonlinearity of the measurement equation even using efficient variants of KF (e.g., extended KF [14]), and thus it is worse than the optimization-based method. Although these geometric-based approaches utilizing various geometric information to improve localization results, their performance is still undesirable in a complex environment with large measurement noise.

### B. DEEP LEARNING-BASED TOA-UWB LOCALIZATION METHODS

In terms of the deep learning-based methods, the key of these approaches is to extract representative features from the ranging measurements. Some works [6], [15] utilize the convolutional neural network (CNN) to extract spatial features including the high-level representation of the measurements, and then the localization results can be estimated through the fully-connected layers based on these spatial features. The experimental results of these works illustrate that these extracted high-level features can boost the localization accuracy, and the noise can also be suppressed partly in the extraction process. In addition to spatial features, some

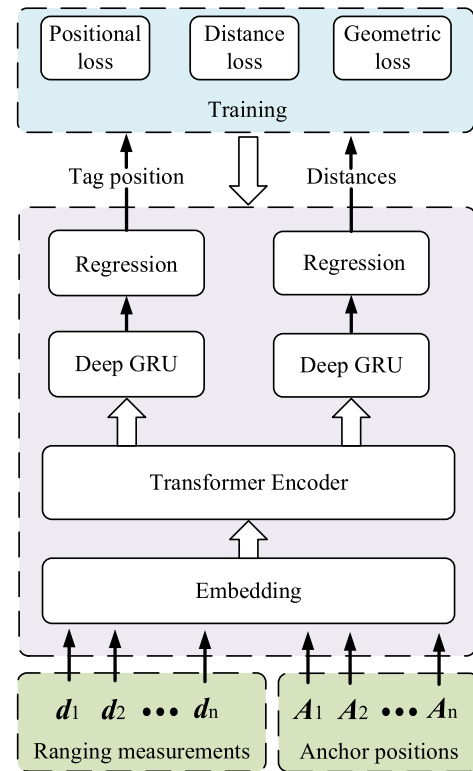
other works also consider extracting the temporal features of ranging measurements to help the UWB localization. Different types of recurrent neural network (RNN) such as long short-term memory (LSTM) [16] and gate recurrent unit (GRU) [7] are adopted to extract the temporal features from the UWB measurements. These works indicate that the temporal relationship between continuous frames is also important for localization. Thus, in our previous work [12], we combine the CNN and RNN modules to learn both the spatial and temporal features from the ranging measurements. Experimental results suggest that the spatial-temporal features can improve the UWB localization performance in a NLOS scene. However, these works ignore to reduce the weights of biased measurements with large noise which tends to severely damage the localization accuracy. Thus, in this paper, we introduce the transformer framework with attention mechanism to overcome this issue. Besides, the training strategy is also important for the deep learning-based methods. Most of the existing training strategies [6], [7], [12], [15], [16] use the distances between the estimated and ground-truth tag positions as cost function, but do not consider the structure of UWB sensors for localization. Hence, it is important to introduce the geometric information into the deep learning-based methods.

In addition, the deep learning-based methods are also developed for the NLOS identification or ranging correction [8], [9], [10], [17], [18], [19], [20]. Either CNN [8], LSTM [9] or transfer learning strategy [17] is introduced to these models. However, most of these approaches utilize channel impulse responses (CIR) waveforms or received signal strength (RSS) [20] signals as input. However, these measurements are not used by some TOA-based localization methods and also unavailable for some UWB devices. Therefore, these signals are not considered in this paper. In this paper, we can recover high-accuracy distance information with raw ranging measurement by the proposed deep network and training strategy.

### III. PROPOSED METHODS

#### A. SYSTEM OVERVIEW

Figure 1 illustrates the flowchart of our proposed deep attention-based UWB localization network and the developed training strategy accordingly. Firstly, we design an embedding module to project raw ranging measurements into a vector where the positional embedding is also contained based on the anchor positions encoding the geometric information. Then, the embedded measurements are fed into a transformer encoder module to learn their high-level spatial features. In this process, the weights of different measurements are considered through the attention mechanism. Next, the extracted spatial features are delivered into independent deep gate recurrent unit (GRU) [21] modules to learn the temporal features of the measurement sequence efficiently. After that, we design two independent regression layers to estimate the tag positions and distance information respectively based



**FIGURE 1.** Flowchart of the proposed deep attention-based UWB localization network and the designed training strategy: transformer encoder module is introduced to consider the weights of different measurements in localization and extract high-level spatial features, while deep GRU modules are utilized to extract the temporal features of continuous measurements. Besides, positional encoding with anchor positions is contained in the embedding process. Finally, in the training process, three losses are designed and minimized jointly.

on the output of the deep GRU module. Finally, a supervised learning strategy combing the pre-collected ground-truth and the geometric constrains are designed to train the proposed network. In the following parts, we will elaborate on the architecture of the proposed network and the designed training strategy, respectively.

#### B. MODEL ARCHITECTURE

##### 1) INPUT

The input of the proposed network is defined as  $Input = (d_1, d_2, \dots, d_M)$ , where  $d_i$  includes 10 continuous measured distances between the tag and  $i$ -th anchor, while  $M$  is the number of anchors. We utilize 10 continuous ranging measurements, since they can contain enough useful information for estimating positions and distances. In addition, the length of the input sequence is set as 16, due to the utilization of deep GRU module.

##### 2) EMBEDDING

The process of the embedding module is shown in Figure 2. For the input  $d_i$ , it is mapped to  $D$  dimensions ( $FeaI_i$ ) though a linear layer, while the position of  $i$ -th anchor  $A_i$  is also mapped to  $D$  dimensions ( $FeaA_i$ ) by another

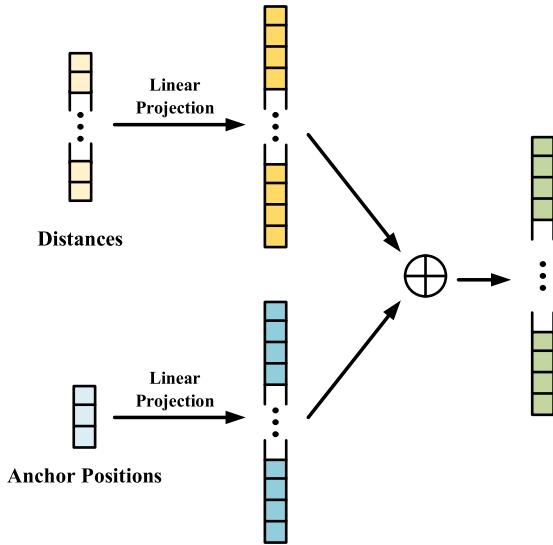


FIGURE 2. The process of the embedding.

linear layer as follows:

$$FeaI_i = FC_1(d_i) \quad i \in [1, 2, \dots, M] \quad (1)$$

$$FeaA_i = FC_1(A_i) \quad i \in [1, 2, \dots, M] \quad (2)$$

where  $FC_1(\bullet)$  represents one fully-connected layer with ReLU function.

Then, these two mapped vectors are added, yielding the output of the embedding modules ( $E_i$ ) as follows:

$$E_i = FeaI_i + FeaA_i \quad i \in [1, 2, \dots, M] \quad (3)$$

Through this process, the features of input signals are extracted coarsely through the linear layer. Furthermore, the geometric information contained in the anchor positions is embedded into the input signals.

### 3) TRANSFORMER ENCODER

In this work, we utilize four transformer encoder layers [22] to extract deep spatial features of the embedded measurements as follows:

$$\mathbf{x}_T = TRFME_4(E_1, E_2, \dots, E_M) \quad (4)$$

where  $TRFME_4(\bullet)$  represents four transformer encoder layers,  $\times \mathbf{x}_T \in \mathbb{R}^{MD}$  is the extracted spatial features of input signals.

The architecture of one transformer encoder layer is shown in Figure 3. The core of the transformer encoder is the multi-head attention (green block in the figure). Through this mechanism, the weights of different embedded measurements are calculated based on the relationship between different inputs, while the input signals are fused based on these calculated weights [22]. Thus, the weights of different UWB measurements can be considered by the transformer encoder layer.

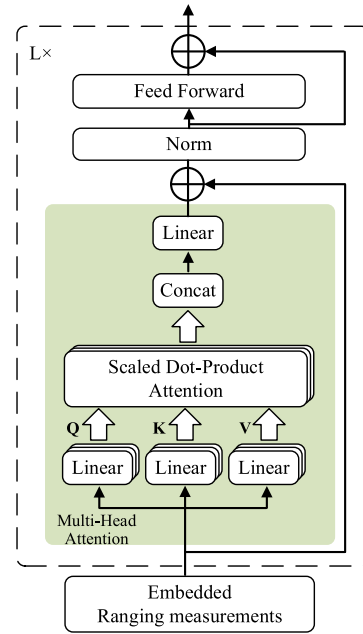


FIGURE 3. The architecture of a transformer encoder layer. The core of this module is the multi-head attention process which assigns different weights to different measurements.

Besides, the addition and normalization layer, feed forward layer and residual connection strategy are utilized in the transformer encoder to handle the vanishing gradient and ensure the generalization ability of the network.

### 4) DEEP GRU AND REGRESSION LAYERS

We build deep GRU modules to learn temporal features of the measurements efficiently, since GRU obtains promising accuracy with lower computational complexity compared to the other RNN structures [21]. After that, we design the regression layers to estimate the distances and tag positions, respectively.

For regression of the tag positions, we first flatten  $\mathbf{x}_T$  as 1D vector, and make use of a linear layer to handle it as follows:

$$\mathbf{x}_{F,j} = FC_1(F(\mathbf{x}_{T,j})), \quad j \in [t-s+1, t-s+2, \dots, t] \quad (5)$$

where  $F(\bullet)$  represents flatten process.  $s$  is the length of the sequence of the deep GRU module, while  $t$  is the current timestep.

Then, the tag positions are estimated by the deep GRU and the regression layers as follows:

$$\mathbf{x}_{p,t} = GRU_3(\mathbf{x}_{F,t-s+1}, \mathbf{x}_{F,t-s+2}, \dots, \mathbf{x}_{F,t}) \quad (6)$$

$$\hat{p}_t = D(FC_1(\mathbf{x}_{p,t})) \quad (7)$$

where  $GRU_3(\bullet)$  represents three GRU layers.  $D(\bullet)$  represents one dense layer without active functions.  $\hat{p}_t$  is the estimated tag positions at timestep  $t$ .

For regression of the distance information, we first divide  $\mathbf{x}_T \in \mathbb{R}^{MD}$  as  $\mathbf{x}_{T,i} \in \mathbb{R}^{1D}$ , where  $i \in [1, 2, \dots, M]$ , and each  $\mathbf{x}_{T,i}$  is delivered into a deep GRU module with regression layer to

learn its own features for predicting the distance between the tag and  $i$ -th anchor:

$$\mathbf{x}_{d,i,t} = GRU_3(\mathbf{x}_{T,i,t-s+1}, \mathbf{x}_{T,i,t-s+2}, \dots, \mathbf{x}_{T,i,t}) \quad (8)$$

$$\hat{d}_{t,i} = D(FC_1(\mathbf{x}_{d,i,t})) \quad (9)$$

where  $\mathbf{x}_{d,i,t}$  is the extracted spatial-temporal features of ranging measurement between the tag and  $i$ -th anchor at timestep  $t$ .  $\hat{d}_{t,i}$  is the estimated distance information between the tag and  $i$ -th anchor at timestep  $t$ .

Finally, we discuss the situations when the number of measurements does not match the input size of the proposed network. On one hand, in the situation that the number of anchors in a scenario is different with the trained network, the model architecture does not require any changes, but the input size of the network needs to be changes as the number of anchors in that scenario, and the whole network needs to be retrained. In addition, if the anchor positions are changed in a scenario, the proposed network also needs to be retrained with the re-collected ground-truth. On the other hand, in the situation where the signals of a few anchors are lost sometimes, we impose zero padding on the input and do not change the length and order of the input signals. For the transformer encoder, the mask matrix is utilized to mask the unavailable data. For the deep GRU modules, zero padding is also imposed on the output of the transformer encoder to ensure the fixed dimensions of  $\mathbf{x}_T$ , while we also mask these padded zero vectors for the distance regression.

Through this designed network, the high-level spatial-temporal features can be extracted, while the weights of different ranging measurements can be calculated. As a result, the large measurement noise can be suppressed, while the biased measurements with large noise can be masked. In addition, the anchor positions containing geometric information is also introduced to the network in the position embedding process.

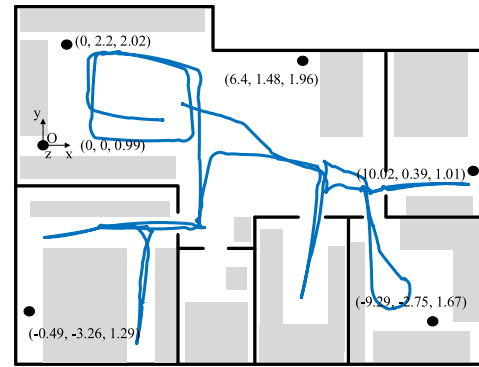
### C. TRAINING STRATEGY

In this work, we employ the supervised learning strategy combing geometric constraints of UWB sensors for model training. To this end, we design three losses, namely positional loss, distance loss and geometric loss in our cost function. The details of these losses will be provided in the following parts.

In terms of the geometric loss  $L_g$ , it is designed to introduce geometric constraints into the proposed network. Following the geometric-based localization approaches [4], we build the geometric constraint that the measured distances equal to the distances between the tag positions and each anchor position. Mathematically, the geometric loss is formulated as follows:

$$L_g = \sum_{i=1}^M \left| \hat{d}_{t,i} - \|\hat{\mathbf{p}}_t - \mathbf{A}_i\|_2 \right| \quad (10)$$

where  $\mathbf{A}_i$  is known and fixed position of  $i$ -th anchor, while  $\hat{\mathbf{p}}_t$  is the estimated tag positions at timestep  $t$  produced from



**FIGURE 4.** The structure map of the experimental field. The black lines indicate the walls, while the grey rectangles illustrate the furniture. The black points represent the anchor positions. The blue line illustrates the trajectory of sequence 3.

the proposed network. Meanwhile,  $\hat{d}_{t,i}$  denotes the estimated distance between the tag and  $i$ -th anchor at timestep  $t$ , and  $M$  is the number of anchors.

Besides, since the estimated tag location and distances are simultaneously generated from our network, the ground-truth positions as well as the ground-truth distances need to be pre-collected, such that the positional loss  $L_p$  and distance loss  $L_d$  are established as follows:

$$L_p = \|\mathbf{p}_t - \hat{\mathbf{p}}_t\|_2 \quad (11)$$

$$L_d = \sum_{i=1}^M |d_{t,i} - \hat{d}_{t,i}| \quad (12)$$

where  $\mathbf{p}_t$  and  $d_{t,i}$  are the ground-truth position and distance of frame  $t$ , respectively.

In this study, we minimize these three losses together, leading to the complete cost function defined as follows:

$$L = \gamma_1 L_g + \gamma_2 L_p + \gamma_3 L_d \quad (13)$$

where  $\gamma_1$ ,  $\gamma_2$  and  $\gamma_3$  are weight parameters of these three losses.

Through this cost function, we can leverage the conventional supervised learning strategy for training our network with the pre-collected ground-truth positions and distances. In addition, the geometric relationship between the estimated tag location and distance information is characterized by the geometric loss. Thus, the geometric constraint of the TOA-UWB localization is introduced into the proposed network, which significantly benefits the training accuracy and the localization performance.

## IV. EVALUATION

### A. EXPERIMENTAL SETTINGS

In order to evaluate the performance of the proposed methods in a complex indoor environment, we carry out real-world experiments in a 10m × 9m narrow and complex apartment scenes. Figure 4 illustrates the structure map of the

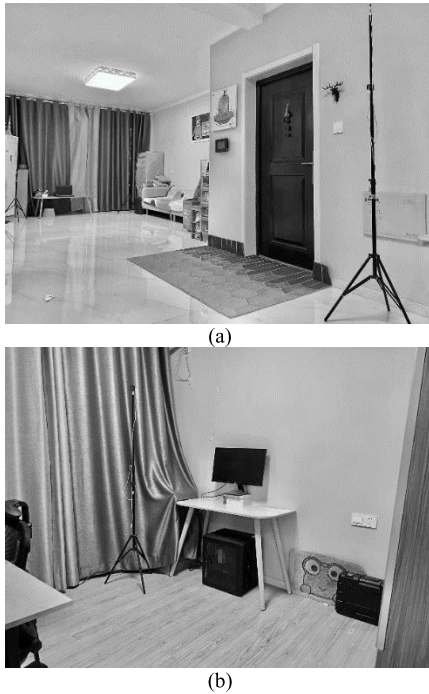


FIGURE 5. Two images of the apartment environment.

experimental field, while Figure 5 shows two images of the environment. It can be observed that this is a narrow indoor scene with walls and different furniture. Besides, people move in this environment sporadically during the experiments. Thus, the signals between tag and anchors will be blocked (NLOS situations), while the multipath effect will also affect the accuracy of ranging measurements. In addition, in the Figure 4, we also give the trajectory of sequence 3 (blue line) as an example, it shows that at any point on this trajectory, there are always several anchors in NLOS condition.

The UWB sensor involved in this experiment is the LinckTrack-S which can measure the distances between a tag and an anchor within 80m with a measurement frequency at 100Hz. The UWB tag mounted on a mobile robot, and we control the robot move randomly to collect experimental data during the experiments. Besides, a 32-line LiDAR (RS-Helios-5515) is also mounted on the robot which is used to provide ground-truth positions with stable and cm-level localization method LOAM [23]. In addition, for the ground-truth distances, it can be calculated based on the ground-truth positions and the pre-calibrated anchor positions.

In terms of the competing method, CNN-LSTM method with pure distance information [12] is involved in our comparative studies, because it achieves state-of-the-art localization results in the complex environments, while it also learns the spatial and temporal features of the ranging measurements similar to the proposed method. In addition, ablation studies are also conducted to further explore individual modules in our proposed method. We do not choose other competing

TABLE 1. Situation of each sequence in the experiments.

Sequence	Type	Number of samples	Average measurement error
Seq 0	Train	8883	0.654
Seq 1	Test	4302	0.689
Seq 2	Test	6847	0.612
Seq 3	Test	6130	0.632
Seq 4	Test	6943	0.617
Seq 5	Test	6109	0.663
Seq 6	Test	4047	0.674
Seq 7	Test	3351	0.533
Seq 8	Test	2859	0.705
Seq 9	Test	3466	0.806

methods, since even the state-of-the-art geometric-based methods cannot achieve reliable localization results in this environment, while the performance of other learning-based methods is significantly worse than the selected CNN-LSTM method.

For performance measure, the commonly used absolute trajectory error (ATE) [24] of the positions are used for the metric to evaluate the localization accuracy.

In implementation, the proposed method is achieved with Pytorch. The Adam algorithm [25] is used as the optimizer. The weight-decay of the Adam is set as 0.0001, and the learning rate is set as 0.0001. In addition, the batch size is set as 1024, while the  $\gamma_1$ ,  $\gamma_2$  and  $\gamma_3$  are all empirically set as 1.0. Besides, the whole network is trained on a NVIDIA GeForce RTX 3060, and it is executed on NVIDIA Jetson Orin NANO in real-time for testing.

## B. RESULTS

In our experiments, we totally collect 10 sequences with one for training and remaining for testing. The situations of the training and testing sequences are given in Table 1. It shows that the average measurement errors of most sequences are larger than 0.6m suggesting substantial challenges in the experiments. In these situations, the classical geometric-based approaches cannot achieve reliable localization results. Regarding the learning-based methods, the localization results of the CNN-LSTM method and our proposed method are given in Table 2.

In Table 2, the ATE of each axis, the total ATE of every sequence and the percentage of total ATE improvement are all presented. It shows that our proposed method improves performance by over 40% in most sequences compared to the state-of-the-art CNN-LSTM method, although CNN-LSTM method can also extract the high-level spatial-temporal features from the input ranging measurements. To be specific, the accuracy of  $z$ -axis achieved by the two methods are comparable since the trajectories have little displacement on  $z$ -axis. For the other two axes, our proposed method can effectively improve the localization accuracy, and thus obtains better localization results.

Besides, Figure 6 and Figure 7 illustrate the trajectory error curves of the two methods of sequence 2 and sequence

TABLE 2. ATE of testing sequences of CNN-LSTM method and our proposed method(m).

Seq	CNN-LSTM [12]				Our				Percentage of improvement
	x	y	z	total	x	y	z	total	
1	0.240	0.361	0.029	0.434	0.143	0.177	0.012	0.228	47.46%
2	0.225	0.414	0.026	0.472	0.190	0.265	0.018	0.327	30.72%
3	0.259	0.408	0.031	0.484	0.222	0.169	0.016	0.281	41.94%
4	0.316	0.534	0.027	0.621	0.196	0.291	0.017	0.336	45.89%
5	0.298	0.403	0.033	0.502	0.192	0.215	0.038	0.290	42.23%
6	0.230	0.372	0.031	0.438	0.161	0.139	0.014	0.213	51.37%
7	0.277	0.375	0.030	0.467	0.189	0.152	0.013	0.244	47.75%
8	0.212	0.516	0.025	0.558	0.199	0.203	0.013	0.284	49.10%
9	0.255	0.682	0.033	0.729	0.176	0.577	0.015	0.603	17.28%

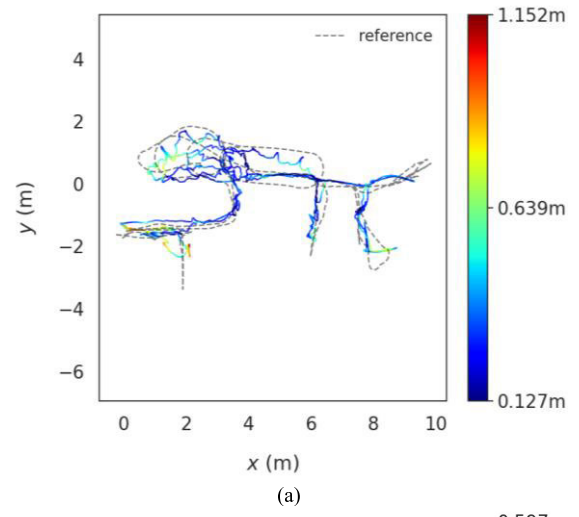
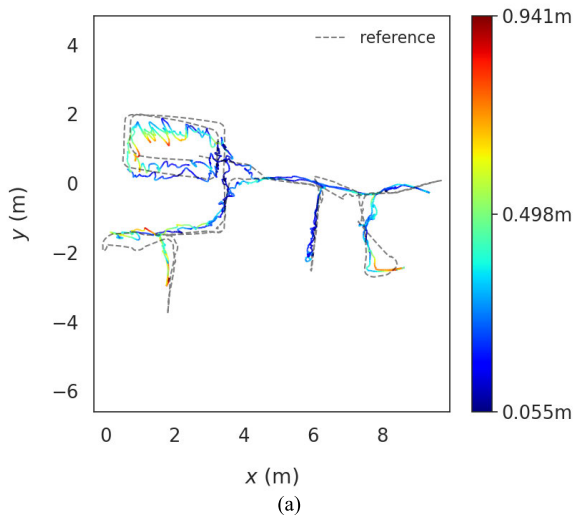


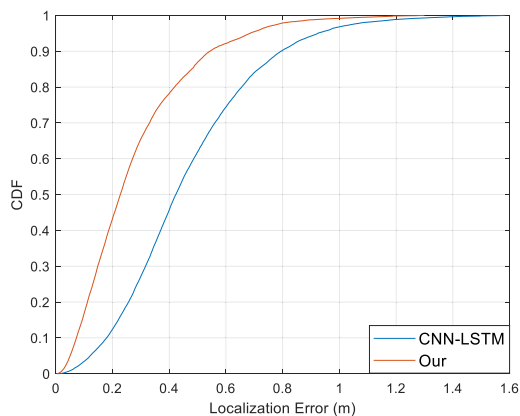
FIGURE 6. Trajectory error curves of (a) CNN-LSTM method and (b) our proposed method of sequence 2.

FIGURE 7. Trajectory error curves of (a) CNN-LSTM method and (b) our proposed method of sequence 4.

4 in x-y-plane. In these figures, it shows that CNN-LSTM method cannot achieve reliable localization results in most trajectories caused by large measurement noise and lack of geometric information, while our proposed method obtains better accuracy in almost all trajectories. Figure 8 shows the cumulative distribution functions (CDF) of localization errors

of whole testing sequences. It also indicates that our proposed method beats the CNN-LSTM method significantly.

We attribute this improvement to the proposed two strategies: the introduction of self-attention mechanism to suppress the measurement noise and mask the biased measurements, along with the introduction of geometric information and constraints to consider the structure of UWB sensors. In the



**FIGURE 8.** The cumulative distribution functions (CDF) of localization errors of sequence 7: our proposed method beats the CNN-LSTM method significantly.

next subsection, we will discuss in details the advantages of these strategies through the ablation studies.

In addition, our proposed method can correct the measured distance which is unavailable for other learning-based TOA-UWB localization method. In Table 3, we give the average error of raw measurements and the corrected distance of our method. It indicates that our proposed method can effectively correct the ranging measurements, and the correction accuracy is related to the localization accuracy. Higher localization accuracy implies higher distance correction accuracy.

### C. ABLATION STUDIES

#### 1) THE ADVANTAGE OF ATTENTION MECHANISM

In this part, we build a competing experiment to suggest the efficiency of transformer framework. To be specific, we first remove the positional embedding process for the transformer encoder in our proposed network as well as the geometric loss in the cost function. Thus, it actually replaces the CNN module in CNN-LSTM method with the transformer encoder as our method.

We give the results of CNN-LSTM method and our proposed method excluding the geometric loss and positional embedding in Table 4. It shows that although there is no geometric information in our proposed method, it still beats the CNN-LSTM method significantly since the attention mechanism in the transformer encoder can encode the weights of different ranging measurements for localization, and thus suppress the measurement noise globally for further improving the localization performance. These localization results indicate the advantage of attention mechanism for UWB localization.

#### 2) THE ADVANTAGE OF GEOMETRIC LOSS

In order to suggest the advantage of geometric loss, we compare the localization performance of our proposed method containing or discarding the geometric loss in the cost function. The localization results are given in Table 5.

**TABLE 3.** Average distance error(m).

Sequence	Raw measurements	Our
Seq 1	0.689	0.129
Seq 2	0.612	0.163
Seq 3	0.632	0.161
Seq 4	0.617	0.185
Seq 5	0.663	0.149
Seq 6	0.674	0.115
Seq 7	0.533	0.141
Seq 8	0.705	0.171
Seq 9	0.806	0.234

In the table, our proposed method with geometric loss obtains better results than that without geometric loss in  $x$  and  $y$  axis, and thus improves the localization performance overall. The geometric loss can provide geometric constraints for the learning-based TOA-UWB localization approaches which depend on the geometric relationship between the tag and anchors, and thus estimate the tag positions more accurately. This improvement is also revealed by the experimental results.

#### 3) THE ADVANTAGE OF GEOMETRIC LOSS

In order to suggest the advantage of geometric loss, we compare the localization performance of our proposed method containing or discarding the geometric loss in the cost function. The localization results are given in Table 5.

In the table, our proposed method with geometric loss obtains better results than that without geometric loss in  $x$  and  $y$  axis, and thus improves the localization performance overall. The geometric loss can provide geometric constraints for the learning-based TOA-UWB localization approaches which depend on the geometric relationship between the tag and anchors, and thus estimate the tag positions more accurately. This improvement is also revealed by the experimental results.

To sum up, we exhibit the advantages of attention mechanism, geometric loss and positional embedding process in our ablation studies. With the help of these designed strategies, our proposed method can improve the UWB localization performance significantly in a complex environment compared to the existing state-of-the-art method.

### D. COMPUTATIONAL COMPLEXITY

Finally, we discuss the computational cost of the proposed method. For the testing process, the whole network is run on NVIDIA Jetson Orin NANO. The average time cost of the whole testing sequences is about 0.004s/sample, while the sampling frequency is 10Hz. Hence, it can be executed in real-time for testing.

In addition, compare to the traditional methods, the deep learning-based methods require another training process. In terms of our proposed method, the whole network is trained on one NVIDIA GeForce RTX 3060. The average training time is about 1.1s/epoch with 8883 training samples, and



**TABLE 4.** ATE of testing sequences of CNN-LSTM method and our proposed method without the geometric loss and positional embedding (m).

Seq	CNN-LSTM [12]				Our without geometric loss and positional embedding				Percentage of improvement
	x	y	z	total	x	y	z	total	
1	0.240	0.361	0.029	0.434	0.153	0.253	0.015	0.296	31.80%
2	0.225	0.414	0.026	0.472	0.237	0.332	0.013	0.408	13.56%
3	0.259	0.408	0.031	0.484	0.252	0.268	0.016	0.368	23.97%
4	0.316	0.534	0.027	0.621	0.255	0.350	0.018	0.417	32.85%
5	0.298	0.403	0.033	0.502	0.239	0.275	0.030	0.365	27.29%
6	0.230	0.372	0.031	0.438	0.201	0.212	0.020	0.293	33.11%
7	0.277	0.375	0.030	0.467	0.222	0.208	0.019	0.305	34.69%
8	0.212	0.516	0.025	0.558	0.211	0.293	0.016	0.362	35.13%
9	0.255	0.682	0.033	0.729	0.281	0.655	0.018	0.713	2.19%

**TABLE 5.** ATE of testing sequences of our proposed method with and without the geometric loss (m).

Seq	Our				Our without geometric loss				Percentage of improvement
	x	y	z	total	x	y	z	total	
1	0.143	0.177	0.012	0.228	0.166	0.201	0.008	0.268	17.54%
2	0.190	0.265	0.018	0.327	0.225	0.301	0.012	0.371	13.46%
3	0.222	0.169	0.016	0.281	0.267	0.187	0.010	0.327	16.37%
4	0.196	0.291	0.017	0.336	0.237	0.333	0.011	0.393	16.96%
5	0.192	0.215	0.038	0.290	0.220	0.246	0.034	0.332	14.48%
6	0.161	0.139	0.014	0.213	0.199	0.146	0.009	0.248	16.43%
7	0.189	0.152	0.013	0.244	0.237	0.163	0.009	0.288	18.03%
8	0.199	0.203	0.013	0.284	0.241	0.205	0.007	0.317	11.62%
9	0.176	0.577	0.015	0.603	0.209	0.616	0.011	0.651	7.96%

**TABLE 6.** ATE of testing sequences of our proposed method with and without the positional embedding(m).

Seq	Our				Our without positional embedding				Percentage of improvement
	x	y	z	total	x	y	z	total	
1	0.143	0.177	0.012	0.228	0.141	0.220	0.014	0.261	14.47%
2	0.190	0.265	0.018	0.327	0.228	0.277	0.017	0.369	12.84%
3	0.222	0.169	0.016	0.281	0.243	0.213	0.017	0.323	14.95%
4	0.196	0.291	0.017	0.336	0.261	0.304	0.016	0.381	13.39%
5	0.192	0.215	0.038	0.290	0.195	0.240	0.037	0.312	7.59%
6	0.161	0.139	0.014	0.213	0.188	0.171	0.014	0.254	19.25%
7	0.189	0.152	0.013	0.244	0.215	0.175	0.013	0.278	13.93%
8	0.199	0.203	0.013	0.284	0.183	0.273	0.012	0.329	15.85%
9	0.176	0.577	0.015	0.603	0.170	0.601	0.015	0.625	3.65%

we train a total of 2000 epochs in the experiment. Thus, the whole training time is about 37minutes. Consider that for an application environment, the training process only needs to be executed once. This training cost can be accepted in practical applications.

## V. CONCLUSION

In this paper, we propose a self-supervised deep location and ranging correction method improving the localization accuracy compared to state-of-the-art classical approaches with desirable computational complexity. In this method, we fuse the classical approach with the self-supervised learning-based method through the estimation of the location and ranging corrections with designed deep network. The greatest advantage of our method is the self-supervised learning strategy based on the topological structure of UWB sensors. This strategy removes the requirement of ground-truth collection, and thus our method can use the deep measurement features

without the ground-truth. Consequently, our approach facilitates real-world applications and can achieve a desirable localization result. The advantages of the proposed method are also suggested through the real-world experiments with different UWB devices in different environments.

Although the proposed method improves the localization performance in complex environments, it requires the process of retraining in different scenes with different number of anchors. Thus, in the future work, we consider to further design more transferable model architecture which can adapt to different number of anchors. In addition, we also consider to propose the self-supervised training method to make the training process does not require to collect the ground-truth, and thus the retraining process will become more convenient.

## REFERENCES

- [1] P. S. Farahsari, A. Farahzadi, J. Rezazadeh, and A. Bagheri, "A survey on indoor positioning systems for IoT-based applications," *IEEE Internet Things J.*, vol. 9, no. 10, pp. 7680–7699, May 2022.

- [2] A. Alarifi, A. Al-Salman, M. Alsaleh, A. Alnafessah, S. Al-Hadhrami, M. Al-Ammar, and H. Al-Khalifa, "Ultra wideband indoor positioning technologies: Analysis and recent advances," *Sensors*, vol. 16, no. 5, p. 707, May 2016.
- [3] S. D. Iacono, V. Paciello, and P. Sommella, "Distance measurement characterization for ultra wide band indoor localization systems," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf. (I2MTC)*, May 2022, pp. 1–6.
- [4] X. Fang, C. Wang, T.-M. Nguyen, and L. Xie, "Graph optimization approach to range-based localization," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 11, pp. 6830–6841, Nov. 2021.
- [5] Y. Xu, Y. S. Shmaliy, C. K. Ahn, G. Tian, and X. Chen, "Robust and accurate UWB-based indoor robot localisation using integrated EKF/EFIR filtering," *IET Radar, Sonar Navigat.*, vol. 12, no. 7, pp. 750–756, Jul. 2018.
- [6] J. Joung, S. Jung, S. Chung, and E. Jeong, "CNN-based Tx–Rx distance estimation for UWB system localisation," *Electron. Lett.*, vol. 55, no. 17, pp. 938–940, Aug. 2019.
- [7] D. T. A. Nguyen, J. Joung, and X. Kang, "Deep gated recurrent unit-based 3D localization for UWB systems," *IEEE Access*, vol. 9, pp. 68798–68813, 2021.
- [8] C. Jiang, S. Chen, Y. Chen, D. Liu, and Y. Bo, "An UWB channel impulse response de-noising method for NLOS/LOS classification boosting," *IEEE Commun. Lett.*, vol. 24, no. 11, pp. 2513–2517, Nov. 2020.
- [9] C. Jiang, J. Shen, S. Chen, Y. Chen, D. Liu, and Y. Bo, "UWB NLOS/LOS classification using deep learning method," *IEEE Commun. Lett.*, vol. 24, no. 10, pp. 2226–2230, Oct. 2020.
- [10] X. Yang, "NLOS mitigation for UWB localization based on sparse pseudo-input Gaussian process," *IEEE Sensors J.*, vol. 18, no. 10, pp. 4311–4316, May 2018.
- [11] W. Wang, D. Marelli, and M. Fu, "Multiple-vehicle localization using maximum likelihood Kalman filtering and ultra-wideband signals," *IEEE Sensors J.*, vol. 21, no. 4, pp. 4949–4956, Feb. 2021.
- [12] B. Yang, J. Li, Z. Shao, and H. Zhang, "Robust UWB indoor localization for NLOS scenes via learning spatial-temporal features," *IEEE Sensors J.*, vol. 22, no. 8, pp. 7990–8000, Apr. 2022.
- [13] S. Gezici, Z. Tian, G. B. Giannakis, H. Kobayashi, A. F. Molisch, H. V. Poor, and Z. Sahinoglu, "Localization via ultra-wideband radios: A look at positioning aspects for future sensor networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 70–84, Jul. 2005.
- [14] Y. Hua, Z. Zhu, G. Zhou, and G. Shen, "Chain state monitoring for a heavy scraper conveyor using UWB-based extended Kalman filter technique with range constraint selection method," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022.
- [15] D. T. A. Nguyen, H.-G. Lee, E.-R. Jeong, H. L. Lee, and J. Joung, "Deep learning-based localization for UWB systems," *Electronics*, vol. 9, no. 10, p. 1712, Oct. 2020.
- [16] A. Poulou and D. S. Han, "UWB indoor localization using deep learning LSTM networks," *Appl. Sci.*, vol. 10, no. 18, p. 6290, Sep. 2020.
- [17] J. Park, S. Nam, H. Choi, Y. Ko, and Y.-B. Ko, "Improving deep learning-based UWB LOS/NLOS identification with transfer learning: An empirical approach," *Electronics*, vol. 9, no. 10, p. 1714, Oct. 2020.
- [18] F. Che, Q. Z. Ahmed, J. Fontaine, B. Van Herbruggen, A. Shahid, E. De Poorter, and P. I. Lazaridis, "Feature-based generalized Gaussian distribution method for NLoS detection in ultra-wideband (UWB) indoor positioning system," *IEEE Sensors J.*, vol. 22, no. 19, pp. 18726–18739, Oct. 2022.
- [19] T. Wang, K. Hu, Z. Li, K. Lin, J. Wang, and Y. Shen, "A semi-supervised learning approach for UWB ranging error mitigation," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 688–691, Mar. 2021.
- [20] M. Katwe, P. Ghare, P. K. Sharma, and A. Kothari, "NLOS error mitigation in hybrid RSS-TOA-based localization through semi-definite relaxation," *IEEE Commun. Lett.*, vol. 24, no. 12, pp. 2761–2765, Dec. 2020.
- [21] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.
- [22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.
- [23] J. Zhang and S. Singh, "Low-drift and real-time LiDAR odometry and mapping," *Auto. Robots*, vol. 41, no. 2, pp. 401–416, Feb. 2017.
- [24] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Oct. 2012, pp. 573–580.
- [25] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.



**KUN TANG** received the Ph.D. degree in transportation engineering from Southeast University, Nanjing, China, in 2019.

He is currently a Lecturer with the School of Automation, Nanjing University of Science and Technology, Nanjing. His research interests include traffic data analysis, dynamic traffic modeling, driving behavior analysis, and traffic safety. He was a recipient of the High-Level Personnel Project of Jiangsu.



**BO YANG** received the B.S. degree in electrical engineering from Nanjing University of Information Science and Technology, China, in 2014, and the Ph.D. degree in navigation, guidance and control from Southeast University, China, in 2020, under the supervision of Prof. Xiaosu Xu.

From 2017 to 2018, he held a visiting position with the Department of Computing Science, University of Alberta, Canada, under the supervision of Prof. Hong Zhang. He is currently a Lecturer with the School of Artificial Intelligence, Nanjing University of Information Science and Technology. His main research interests include SLAM, indoor localization, multi-sensors fusion, and deep learning.



**KAI DING** received the Ph.D. degree in artillery, automatic gun and ammunition engineering from the Army Engineering University of PLA, Nanjing, China, in 2013.

He is currently an Engineer with the Science and Technology on Near-Surface Detection Laboratory, Wuxi, China. His research interests include integrated localization and navigation, and target recognition.

...