Arxiv Version

3D Indoor Localization Using Partial-Frequency CSI Fingerprints with Mask Transformer Encoder

Xiping Wang, *Student Member, IEEE,*

***Abstract*—Fingerprint-based deep learning (DL) method is a promising solution for precise 3D indoor Localization, but existing methods struggle with compatibility across various frequency bands. In this paper, we propose a mask Transformer encoder (MTE) model for 3D indoor localization using partial-frequency Channel State Information (CSI) fingerprints. We first introduced how to model CSI data into sequential data suitable for Transformer-based models, with the frequency band as the sequence length dimension. Based on this, we designed the MTE model that can achieve fingerprint localization with arbitrary sub-channel CSI data input. The fingerprint database is constructed sing ray-tracing (RT) simulations with real indoor scenarios 3D models and versatile electromagnetic (EM) coefficients. Extensive experiments demonstrate that the MTE model significantly outperforms various baselines, including CNN, LSTM, RNN, and Swin-Transformer, especially under conditions of varying available channel CSI data. Results indicate that the MTE model not only achieves high localization accuracy but also offers significant advantages in terms of training and storage costs compared to conventional models. Additionally, we explored the model's performance and robustness through various methods.**

***Index Terms*—Channel state information, fingerprint localization, mask Transformer, ray-tracing**

I. Introduction

The 6th generation (6G) wireless communication is anticipated to achieve ubiquitous connectivity, massive communication, and integrated sensing and communication (ISAC) [1] [2]. These capabilities will revolutionize various industries while enabling numerous applications [3]. Three-dimensional (3D) wireless indoor localization plays an increasingly vital role in 6G. Many 6G technologies in indoor environments, such as beamforming and reconfigurable intelligent surfaces (RISs), depend heavily on the precise locations of devices. Localization is also one of the significant challenges for mobile systems [4], such as unmanned aerial vehicles (UAVs), to be reliably employed in 3D indoor environments [5]. Furthermore, the demand for location-based services in large indoor venues, such as passenger navigation and luggage tracking, is growing rapidly. Overall, wireless localization technologies hold extensive research prospects in the 6G era.

To achieve high-precision and robust indoor localization systems, extensive research has been conducted on various wireless signals, including UWB, RFID, Wi-Fi, Bluetooth, and LoRa [6]. Many indoor localization techniques, such as trilateration, triangulation, and fingerprint methods [7], have been developed. Traditional techniques like trilateration and triangulation are susceptible to multi-path effects and often perform poorly in Non-Line-of-Sight (NLOS) environments [8], whereas fingerprint methods have shown high precision in such challenging settings. Recently, the advancement of artificial intelligence (AI), particularly deep learning (DL), has garnered significant attention in fingerprint localization. DL excels in modeling complex, non-linear relationships and capturing intricate patterns in high-dimensional data [9], enabling the efficient handling of large volumes of wireless fingerprint data, such as channel state information (CSI), with higher precision compared to traditional methods. The DL-based fingerprint localization typically involves constructing a fingerprint database and training a DL model to map wireless fingerprints to estimated locations in the offline phase. In the online phase, the trained model is deployed for real-time localization. Researchers have explored DL models, including convolutional neural networks (CNNs) [10], attention mechanisms [11], Transformers [12], and reinforcement learning [13], to enhance localization precision and speed. Various DL techniques, such as broad learning [14], extreme learning [15], and ensemble learning [16], have also been applied. Additionally, methods like fingerprint crowd-sourcing [17] for data collection and federated learning [18] for model training have been widely studied.

Despite many advancements in the research of DL-based fingerprint localization, there is still limited research focusing on localization with frequency-variable channel fingerprints. Current studies assume that the receiver (Rx) operates on a fixed frequency band, which holds true for most wireless communication systems. Even with the evolution of new communication standards exploiting larger bandwidths (such as IEEE 802.11n and 5G New Radio) and incorporating new frequency bands (such as the 6 GHz band in IEEE 802.11be and 6G), backward compatibility in next generation devices facilitates communication with previous generation devices. Additionally, there are extensive research on spectrum sensing [19], aggregation [20], and sharing [21], achieving intelligent spectrum management. However, DL-based fingerprint localization methods relying heavily on fingerprint database specificity. They struggle with frequency-variable generalization using variable bandwidth or different frequency band CSI data. The variable frequency bands alter the dimensions of the input CSI data, posing challenges for DL models, particularly CNNs. Furthermore, variable frequency bands require the model to approximate a surjective function. Although DL models have successfully approximated surjective functions in computer vision tasks (e.g., image classification), applying this

This work is supported(*Corresponding author:*)

to wireless fingerprint localization presents some difficulties (as discussed later in this paper). To the best of our knowledge, literature addressing this challenge is scarce. For instance, [22] achieved data completion for missing CSI frames, enhancing robustness in sensing tasks. However, this article only addresses the completion of data for specific frequency points and does not achieve wireless sensing in the absence of data for certain frequency bands. As a result, developing new DL models for frequency-variable fingerprint localization is imperative. First, if a DL model could be designed to extract channel information from different frequency bands during the offline phase and learn the correlations between data from different bands, it would enhance fingerprint localization during the online phase. In scenarios where the receiver can only receive channel data from certain frequency bands, the DL model would leverage the learned correlations from the offline phase to infer additional channel information, thereby improving localization accuracy. Second, training separate DL models for many frequency band combinations would lead to substantial training costs and storage requirements, which significantly restrains the practical application of DL-based fingerprint localization.

With the advent of the Transformer model, DL research in natural language processing (NLP) entered a new era. Transformers, with their self-attention mechanism and ability to capture long-range dependencies, are theoretically capable of handling variable-length inputs [23]. Additionally, Transformers possess a larger number of parameters, often outperforming CNNs in tasks with larger datasets. The suitability of Transformer-based models for processing CSI data is also widely acknowledged in wireless communication community [24]. Therefore, further exploring how to leverage the variable-length input capability of Transformers to achieve partial-frequency (several sub-channels within a large frequency band) CSI fingerprint localization holds considerable potential. Therefore, this study explores Transformer-based models for handling partial-frequency CSI data in fingerprint localization. The contributions include:

1) Detailed 3D modeling of two real indoor scenarios with five set of material electromagnetic (EM) coefficients. Following IEEE 802.11be standards, ray-tracing (RT) simulations were conducted to generate 160 MHz bandwidth CSI data received by four Rx, establishing a versatile CSI fingerprint database.
2) We proposes data transformation and pre-processing procedures to convert partial-frequency CSI data into sequential data according to the minimum bandwidth of a sub-channel. Building upon this, utilizing the mask Transformer encoder (MTE) as the backbone model, we achieve 3D indoor fingerprint localization using random partial-frequency CSI data as input. We provide a detailed description of the MTE model's architecture, loss function, training configuration, and many other details.
3) Extensive experiments were conducted to prove the proposed MTE model's superiority when using random partial-frequency CSI data as input. We compared our model with state-of-the-art (SOTA) baselines from recent literatures as well as classical time-series models such as LSTM and RNN. In addition, we compared the proposed MTE with alternative Transformer methods. We also conducted a comprehensive performance evaluation of the proposed model.

In the remainder of the paper, related works are reviewed in Section II. In Section III, we present the overall model architecture, followed by the problem definition and data preprocessing procedures. Sections IV introduces the RT simulation and how we constructed CSI datasets. The experimental setups and evaluation results are presented in Section V. Finally, our work is summarized in Section VI.

II. Related Works

A. DL-based CSI Fingerprint Localization

CSI fingerprinting is a prominent technique for wireless localization, leveraging the unique characteristics of wireless channel to identify precise locations [25]. CSI-based localization captures detailed channel characteristics such as amplitude, phase, or angle of arrival at multiple sub-channels and antennas, providing a high-resolution spatial signature that can be used to distinguish between different locations within an environment.

Recently, many DL-based methods are proposed for CSI localization. CNN has significantly advanced CSI fingerprint localization. For instance, literature [26] applied deep CNNs to CSI data, demonstrating improved localization precision by leveraging the spatial correlations within the CSI data. Literature [27] proposes a novel 3D CNN-enabled method for localizing mobile terminals in massive multi-input-multi-output (MIMO) orthogonal frequency division multiplexing systems using angle-delay channel power matrix fingerprints, achieving higher precision and robustness compared to traditional methods.

Transformers, with their attention mechanisms, offer another promising approach to CSI fingerprint localization. Contrary to CNNs, Transformers can capture global dependencies within the data, providing a more comprehensive understanding of the spatial characteristics [28]. Literature [11] proposes a novel attention-augmented residual CNN for CSI fingerprint indoor localization, enhancing both local information and global context utilization. Literature [12] presents Swin-Loc, a Swin Transformer-based CSI data-driven indoor localization framework, which enhances CSI features and improves localization precision, achieving superior precision and reduced storage overhead. Besides, broad learning [14] and extreme learning [15] are also implemented in DL-based CSI localization.

Despite the advantages of DL-based CSI localization, it still faces challenges such as dealing with multi-path fading and noise in localization environments. Moreover, it's difficult to develop high-quality CSI dataset for training DL models. Further research is needed to address these issues and improve the robustness and efficiency of the localization systems.

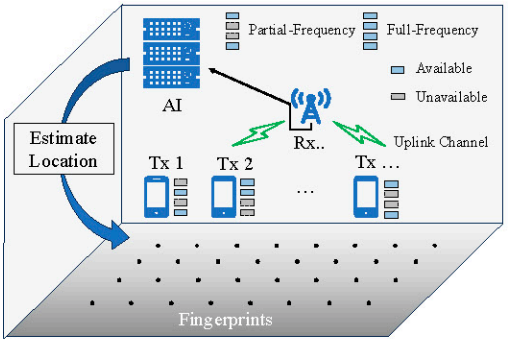


Fig. 1: An illustration of Fingerprint Localization of multiple transmitters (Tx) with partial-frequency CSI data

B. The Mask Mechanism in Transformer

The mask mechanism in Transformers originated from the need to handle sequential data effectively, particularly in the context of natural language processing (NLP). The concept was introduced with the original Transformer architecture in the groundbreaking paper [23]. This architecture leveraged the self-attention mechanism to process input sequences in parallel, making it highly efficient for tasks like machine translation. The mask mechanism in Transformers serves two key functions. First, preventing information leakage. During training for tasks like language modeling, a causal mask is applied to ensure that predictions for a given position depend only on preceding positions, preventing the model from seeing future tokens. Second, handling padding tokens. Padding masks are used to ignore padding tokens added to sequences of varying lengths, ensuring they do not influence the model's predictions during attention calculations.

Mask Transformers are widely used in AI applications including computer vision and NLP. For example, in image segmentation, Mask2Former [29] handles panoptic, instance, and semantic segmentation by utilizing masked attention to focus on relevant regions, significantly improving performance across multiple benchmark datasets. The mask Transformer in literature [30] efficiently model long-range interactions for large hole in-painting, ensuring high-resolution image recovery with high fidelity and diversity. MaskGIT [31] uses a bidirectional Transformer decoder to predict and refine masked tokens, enabling faster and more efficient image generation compared to traditional models. Similar ideas are also implemented in literature [32] where the famous model bidirectional encoder representations from Transformers (BERT) is proposed. Techniques like mask-piloted training in MP-Former [33] address inconsistencies in mask predictions, boosting performance and speeding up training in segmentation tasks. Moreover, the mask mechanism in Transformer has also attracted attentions from wireless communication researchers. Recent work has already tried to use masked Transformer auto-encoder to recover the data lost during wireless transmission [34]. More research results are expected in the field of wireless communication utilizing the mask mechanism of Transformers.

III. Model Architecture

A. Problem Definition

For indoor 3D wireless localization, our objective is to DL models to estimate the location of one or more user equipment (UE) based on the up-link partial CSI data received by M Rx. The entire frequency range is divided into K sub-channels, each with a bandwidth of Δf. Not all sub-channels could be actively transmitting data at a given time (unavailable) hence, a subset of these sub-channels, denoted by $\mathcal{K} \subset \{1, 2, \ldots, K\}$, is used. The full frequency CSI data is denoted by I_C, and the partial CSI data for Rx m and available sub-channel k (if k belongs to $\mathcal{K}$) is represented as $I_C^{m,k}$. An illustration is provided in Fig. 1. For each Rx m in the set $M = \{1, \ldots, M\}$ and each sub-channel k in $\mathcal{K}$, the partial-frequency CSI data $I_C^{m,k}$ can be defined by the frequencies it includes:

$$\mathcal{F}_k = [f_{k,\min}, f_{k,\max}] \tag{1}$$

$$I_C^{m,k} = \{I_C(f) \mid f \in \mathcal{F}_k\} \tag{2}$$

We define a localization mapping function L where the input is the collection of partial-frequency CSI data from the available sub-channels for each Rx, and the output is the estimated position $\mathbf{l}$. The objective is to ensure that the estimated position $\mathbf{l}$ closely approximates the true 3D position $\hat{\mathbf{l}}$. This can be mathematically expressed as:

$$\mathbf{l} = L(\{I_C^{m,k} \mid m \in M; k \in \mathcal{K}\}) \tag{3}$$

The error function E is to quantify the closeness between the estimated and the true positions. The optimization objective for the localization model is to minimize this error:

$$\min_L \mathbb{E}[E(\mathbf{l}, \hat{\mathbf{l}}) \mid \{I_C^{m,k} \mid m \in M; k \in \mathcal{K}\}] \tag{4}$$

Combining these elements, the overall goal is to develop robust models for fingerprint localization that can effectively handle partial-frequency CSI data to achieve accurate indoor 3D localization, ensuring that each component performs effectively under the constraints of limited bandwidths of radio frequency units.

B. Partial Frequency CSI Fingerprints Data Transformation and Pre-processing

In 3D indoor wireless localization, CSI fingerprint data is critical for achieving accurate localization. The CSI data in this paper is structured as an $M \times (K \times \Delta f)$ matrix, where M represents the number of Rx, and $K \times \Delta f$ denotes the total number of frequency points. The minimum spacing between each coordinate in space significantly exceeds the wavelength distance, resulting in substantial phase variations. Therefore, only the amplitude of the CSI is utilized as fingerprinting information, disregarding the phase. Consequently, the CSI data set is expressed as:

$$\mathcal{D} = \{I_C : x \in [1, X], y \in [1, Y], z \in [1, Z]\} \tag{5}$$

where X, Y, and Z define the dimensions of the 3D localization space along the respective axles, forming an $X \times Y \times Z$ spatial grid.

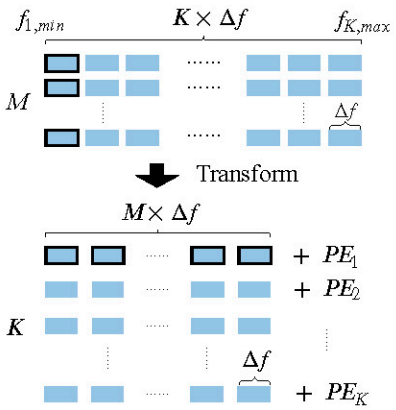


Fig. 2: The transform process and positional encoding for CSI data

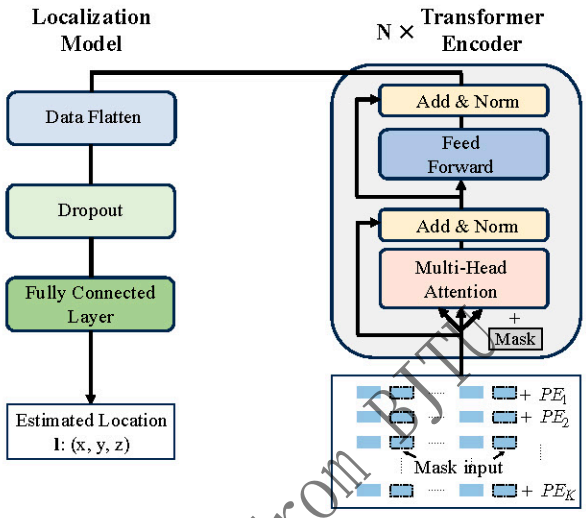


Fig. 3: The proposed mask Transformer encoder architecture for partial-frequency CSI fingerprint localization.

The transformation process of the CSI fingerprint data, as illustrated in Fig. 2, involves reshaping the original matrix $M \times (K \times \Delta f)$ into a format suitable for sequential input into time series DL models. Specifically, the original matrix is divided into $M \times \Delta f$ segments, each with K rows. This reshaping allows each K row segment along the $M \times \Delta f$ dimension to be input into time-series models sequentially. Physically, this transformation results in a 3D tensor where each row corresponds to the CSI data received by M Rx for each sub-channel. For each sub-channel, we have an M-dimensional vector representing the received CSI data across all Rx. Consequently, the time-series DL models such as Transformer, LSTM, or RNN can process these sequences, maintaining the spatial and frequency domain relationships inherent in the CSI data.

The next part will introduce the CSI pre-processing procedures. Considering that only partial-frequency CSI data may be available, these pre-processing steps are applied separately to each of the K sub-channels of M Rx.

• **Sliding Average Filtering**: To mitigate additive noise, a sliding average filter is employed along the frequency dimension for each Rx:

$$I_{\text{filt}}^{m,k}(f) = \frac{1}{W} \sum_{w=-W/2}^{W/2} I_C^{m,k}(f+w) \tag{6}$$

Here, W represents the window size of the filter, k is the sub-channel index, and m is the Rx index. Post-filtering, zero-padding is applied to maintain consistent data dimensions.

• **Standardization**: This process normalizes the filtered data per Rx to a mean of zero and a variance of one:

$$I_{\text{std}}^{m,k}(f) = \frac{I_{\text{filt}}^{m,k}(f) - \mu^{m,k}}{\sigma^{m,k}} \tag{7}$$

where $\mu^{m,k}$ and $\sigma^{m,k}$ are the mean and standard deviation computed from $I_{\text{filt}}^{m,k}$ across all frequency points f within sub-channel k.

• **Normalization**: The standardized data is scaled to the range $[-1, 1]$ to facilitate uniform processing by machine learning models:

$$I_{\text{norm}}^{m,k}(f) = 2\left(\frac{I_{\text{std}}^{m,k}(f) - \min(I_{\text{std}}^{m,k})}{\max(I_{\text{std}}^{m,k}) - \min(I_{\text{std}}^{m,k})}\right) - 1 \tag{8}$$

These pre-processing steps are performed for each sub-channel independently, ensuring that the partial-frequency data is thoroughly prepared for subsequent localization tasks.

C. Mask Transformer Encoder for CSI Fingerprint Localization

Transformers are known for their large number of parameters, making them well-suited to effectively process the complex and voluminous nature of CSI data. Moreover, since this paper investigates localization with partial-frequency CSI data, the chosen model must be capable of handling incomplete information. Therefore, Transformer-based models are highly appropriate for the problem at hand.

The data received by the Rx is first processed using the data pre-processing procedures introduced in Section III. B. This includes sliding average filtering, standardization, and normalization. Based on the known missing sub-channels, a mask is generated for the Transformer encoder. Missing sub-channel data is either set to zero or a minimal number (1e-10 in this paper). Positional encoding (PE) is then added to the pre-processed CSI data to help the model capture positional relationships within the sequence. The prepared data, along with the mask and positional encoding, is input into the proposed MTE. The model processes input CSI data to produce either the estimated 3D coordinates or the recovered full-frequency CSI data. The overall model architecture of our proposed MTE is demonstrated in Fig. 3.

To enhance the model's understanding of positional information within the sequence, we incorporate PE similar to that used in Transformer models. Given that we have K unique sub-channels corresponding to the K rows, the positional encoding PE is added to the CSI data. For each sub-channel, the positional information is applied across the $M \times \Delta f$

TABLE I: RT simulation configuration

Scenario	**Siyuan Hall (SY)**	**Xuehuo 3F (XH)**
Frequency band	5.49 - 5.65 GHz (160 MHz) [35]	
Frequency resolution	78.125 kHz	
Tx antenna	Dipole antenna	
Rx antenna	Omni-directional antenna	
Area	19 × 6 × 8 (m)	22 × 6 × 8 (m)
Tx locations	The ceiling and surrounding walls	
minimum intervals of Rx in x, y, z axle	0.25, 0.25, 0.50 (m)	0.275, 0.25, 0.25 (m)
Propagation mechanism	Direct path, reflection, scattering, penetration	

B. RT Simulation Configuration and CSI Datasets

RT simulation is conducted by the CloudRT simulator [36] [37]. This RT simulator is configured according to the IEEE 802.11be standards [35], as listed in TABLE I. In the simulation, the channel 114, covering 5.49 GHz to 5.65 GHz (160 MHz), is chosen as the simulation frequency band. The frequency resolution is 78.125 kHz, which closely aligns with the sub-carrier space of orthogonal frequency-division multiple access (OFDMA). Therefore, there are a total of 2048 frequency points. The simulation employs an omni-directional vertical polarization antenna as the antenna pattern of Rx, and dipole antenna pattern for Tx. Direct propagation, reflection, scattering, and penetration are simulated in this work. We set 4 Rx distributed at the ceiling and surrounding walls, while the positions of Tx (fingerprints) are distributed at around 0.25 m intervals in 3D space. Due to the length limit, the Tx and Rx locations are not introduced in detail.

The electromagnetic (EM) coefficients of materials can be input as parameters into the RT simulator used in this paper. These coefficients influence the results of reflection, scattering, and transmission in multipath propagation. In the RT simulation, a total of 12 materials are involved, with corresponding coefficients refereed from ITU-R P-2040. We name this set of material electromagnetic coefficients selected from ITU-R P-2040 as em_0. It is noteworthy that in the RT simulation, we do not only use the EM coefficients from ITU-R P-2040 but also simulate using four additional sets of coefficients for each material. Among the latter four sets, two sets involve only a 20% variation in the material coefficients, while the other two sets involve swapping different material coefficients. The former is to simulate slight EM coefficient changes due to temperature, humidity, and material degradation, referred to as em_1 and em_2 in this paper. The latter is to simulate the situation of inputting incorrect types of material EM coefficients, referred to as em_3 and em_4 in this paper. Therefore, this paper uses a total of five different sets of material parameters for evaluation, thus constructing a versatile fingerprint dataset. It should be noted that for em_1 - em_4, we only performed simulations and constructed the dataset based on the minimum interval of 1 meter along the x, y, and z axles

The CSI database contains data from two scenarios where, after the Tx sends a signal at each location, four Rx capture the CSI data. In this paper, we use RT simulation to generate the CTF and consider it as the CSI. The magnitude of the CSI data received by each Rx is taken as fingerprint data. There are four Rx and 2048 frequency points in full-frequency CSI data, as introduced before. According to the IEEE 802.11 standard, the smallest communication band is 20 MHz, corresponding to 256 frequency points. Consequently, the full-frequency CSI data can be divided into 8 sub-channels of CSI data.

V. Experiments and Analysis

A. Experiment Configurations

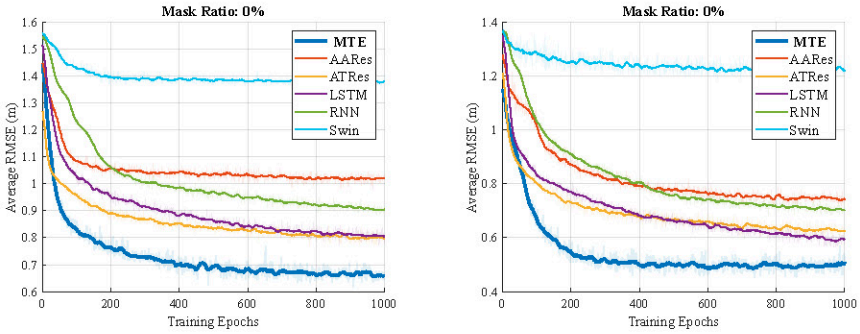
In this paper, MTE and all the baseline model training are performed using PyTorch 1.10.2 on a workstation with 1 NVIDIA H100 GPU, 1 Intel Core i9-12900K CPU and 256 GB DDR4 RAM. The MTE model used in this paper consists of 6 layers of Transformer encoders, with each layer having 8 heads, a feed-forward layer size of 2048, and a dropout rate of 0.1. The remaining settings follow the default configurations from [23] as closely as possible. If there is no special clarification, the default training configuration for every model is trained for 1000 epochs. The learning rate is set as 0.0001. Adam optimizer is used for gradient descent. Batchsize is set as 50.

We extracted fingerprint data of size $21 \times 21 \times 10$ from the central locations of the two scenarios to serve as datasets for training and evaluating the model. The two scenarios, referred to as Siyuan Hall and Xuehuo 3F, have dimensions of $5\,\text{m} \times 5\,\text{m} \times 5\,\text{m}$ and $5\,\text{m} \times 5\,\text{m} \times 2.5\,\text{m}$, respectively. All data were randomly splitted into training and evaluation sets in an 8:2 ratio. Additionally, in our study of partial-frequency CSI fingerprint localization, we defined four cases of partial-frequency conditions. There are a total of 8 sub-channels, with 0, 2, 4, and 6 sub-channels missing, corresponding to 0%, 25%, 50%, and 75% of Mask Ratio, respectively. Unless otherwise specified, a Gaussian noise with a signal-to-noise ratio (SNR) of 40 dB is added to every sampling in all experiments to simulate real-world conditions. For MTE, Transformer, and LSTM and RNN models, unless otherwise specified, the input data is randomly selected according to the Mask Ratio for the available sub-channel CSI data.

B. Comparisons of MTE and Baseline Localization Methods

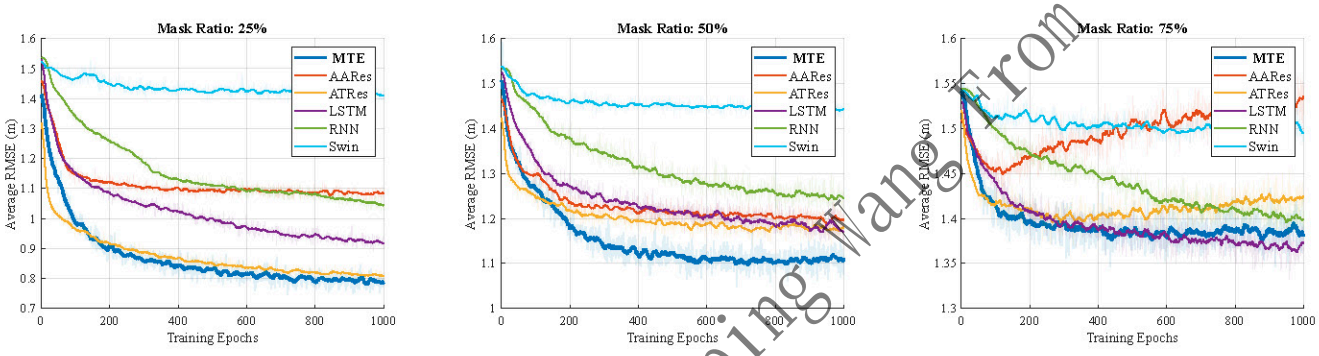
To compare and evaluate the proposed MTE method in this paper, we conducted comparative experiments with SOTA fingerprint localization baselines. These baselines encompass mainstream DL models, including residual neural network (ResNet), attention mechanism, Swin-Transformer, as well as classic sequential neural networks such as long short-term memory (LSTM) and recurrent neural network (RNN):

- AARes [11]: Literature [11] proposes a novel high-accuracy and generalizable indoor localization system that utilizes an attention-augmented residual CNN to comprehensively leverage both local and global information in CSI. In this paper, we employ a similar attention-augmented residual CNN to achieve fingerprint localization. The input data is modeled as an $M \times (K \times \Delta f)$ matrix in Fig. 2, since the experimental data in this paper is not the structure of the massive MIMO communication. The



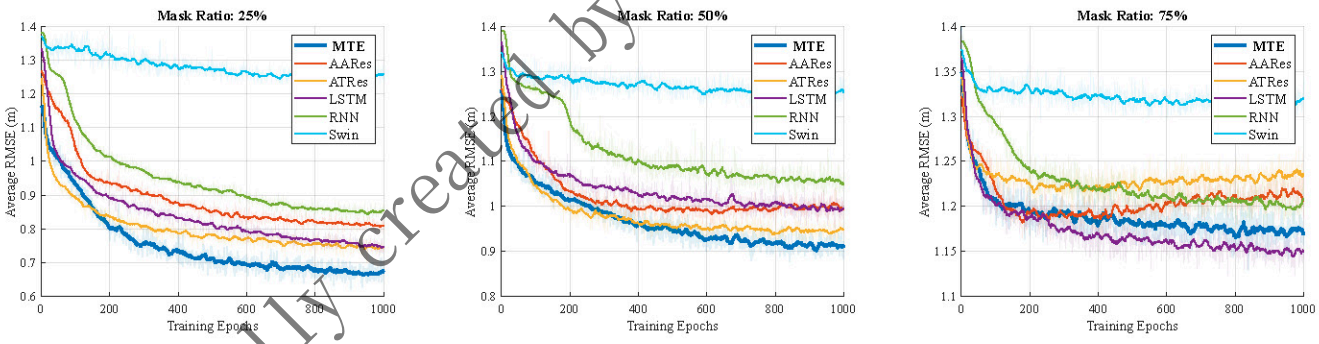
(a) Siyuan Hall (b) Xuehuo 3F

Fig. 5: The evaluation average RMSE training curve of Mask Ratio 0% at two scenarios.



(a) Mask Ratio 25% (b) Mask Ratio 50% (c) Mask Ratio 75%

Fig. 6: The evaluation average RMSE training curve of Mask Ratio 25% - 75% at **Siyuan Hall**.

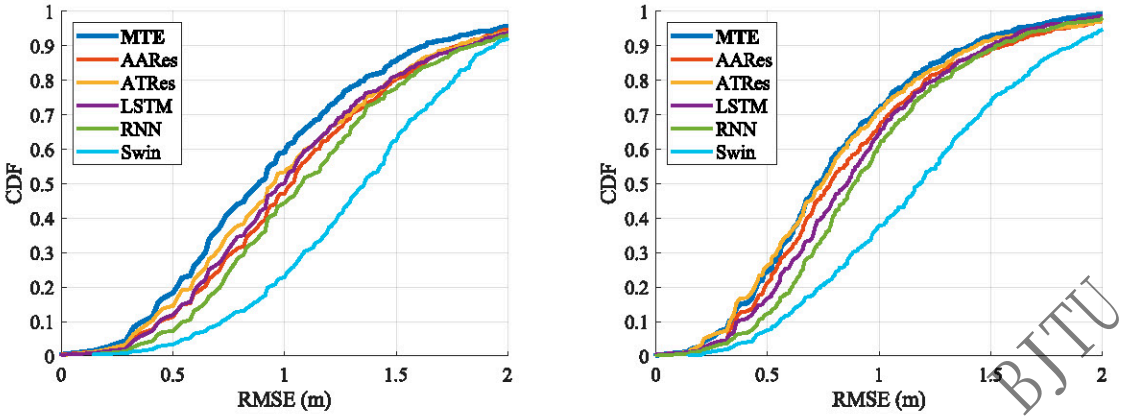


(a) Mask Ratio 25% (b) Mask Ratio 50% (c) Mask Ratio 75%

Fig. 7: The evaluation average RMSE training curve of Mask Ratio 25% - 75% at **Xuehuo 3F**.

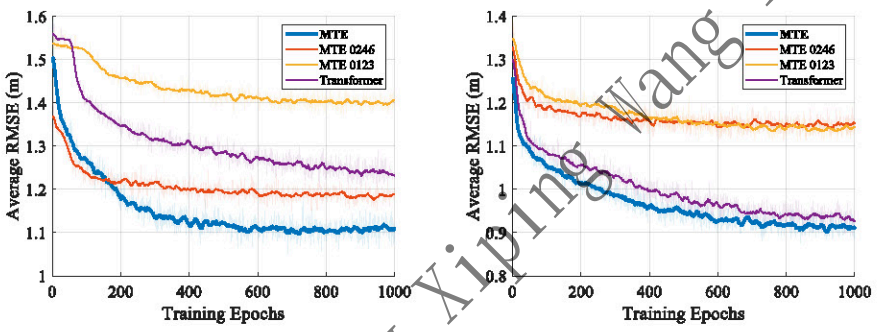
model has been fine-tuned for optimal performance. This baseline can be regarded as representative of attention-based DL models.

- ATRes [38]: ATRes is a residual CNN capable of achieving 3D fingerprint localization. In this paper, we tested the similar model of ATRes in [38] and performed fine-tuning. The input data structure is the same as in the AARes experiments. ATRes can be regarded as representative of ResNet-type models.
- LSTM: We use the LSTM layer to process the input sequentially across the $K\times(M\times\Delta f)$ sequence data, where K is the sequence dimension, producing outputs at each step. The input data structure is identical to that in MTE training, which also illustrated in Fig. 2. The final time step's output is then fed into two fully connected layers for further processing and prediction.
- RNN: Similar to LSTM introduced above, we developed an RNN model to process the sequential input data of $K\times(M\times\Delta f)$ size and realize the localization (regression) via two FC layers.
- Swin: Literature [12] introduced an Swin-Transformer [39] based models to realize massive MIMO CSI localization, solving the CSI fingerprint distortion while reducing resource consumption. We also developed a Swin-Transformer based model for evaluation, trying the best to follow every details in [12]. As our training data is not massive MIMO type, we employed the shift window in the processed high-dimensional feature. Specifi-



(a) Siyuan Hall (b) Xuehuo 3F

Fig. 8: The CDF of evaluation RMSE from MTE and baselines at Mask Ratio 50%.



(a) Siyuan Hall (b) Xuehuo 3F

Fig. 9: The evaluation RMSE curve of our MTE and alternative Transformer methods at Mask Ratio 50%.

TABLE II: The minimum evaluation RMSE (m) of MTE and baselines at Mask Ratio 0% - 75%. Best results in bold face.

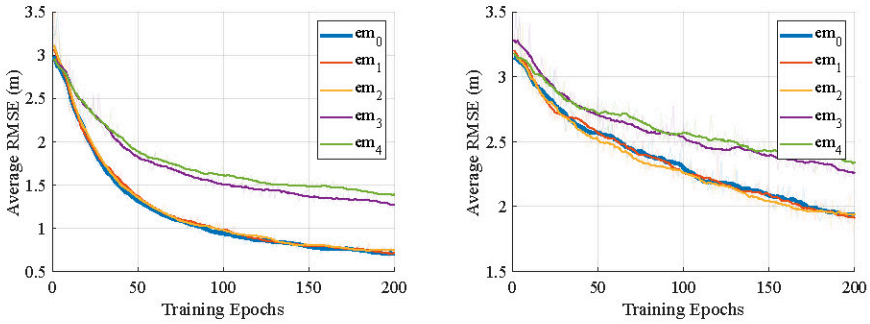
Scenario	Siyuan Hall				Xuehuo 3F			
Mask Ratio	0	25%	50%	75%	0	25%	50%	75%
MTE (ours)	**0.63**	**0.75**	**1.11**	1.36	**0.45**	**0.62**	**0.88**	1.15
AARes	0.99	1.07	1.19	1.42	0.74	0.80	0.96	1.17
ATRes	0.77	0.80	1.17	1.39	0.61	0.42	0.93	1.21
LSTM	0.79	0.89	1.16	**1.34**	0.58	0.73	0.97	**1.13**
RNN	0.90	1.02	1.26	1.39	0.71	0.84	1.04	1.19
Swin	1.38	1.41	1.45	1.48	1.21	1.24	1.24	1.30

TABLE III: Comparison of model sizes, training times, and inference times.

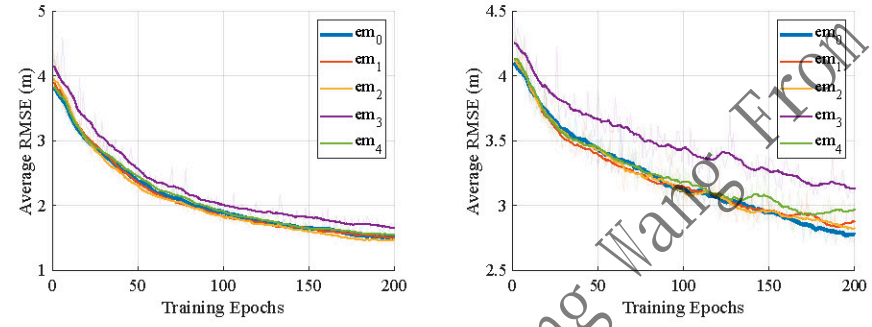
Models	MTE	AARes	ATRes	LSTM	RNN	Swin
Size	224 MB	18.5 MB	21.7 MB	4.07 MB	2.23 MB	188 MB
Training Time (1000 Epochs)	142 min	38 min	29 min	15 min	14 min	168 min
Inference Time	0.6 ms	0.2 ms	0.3 ms	0.2 ms	0.2 ms	0.8 ms

cally, we transform the $K \times (M \times \Delta f)$ data into multi-dimensional square-shaped data, resembling image data. This restructured format is then processed using the Swin-Transformer's hierarchical sliding window mechanism.

In the experiments of this section, the average root mean squared error (RMSE) refers to the average of all localization RMSEs in the evaluation set. Fig. 5 - 7 show the evaluation average RMSE training curves of our proposed MTE and other baselines under different Mask Ratio conditions for the Siyuan Hall and Xuehuo 3F scenarios. Firstly, from Fig. 5, it can be observed that when the Mask Ratio is 0%, meaning all sub-channel CSI data is available, our proposed MTE model performs significantly better than the other baselines. Only during the initial few epochs, some baselines like ATRes have a faster error reduction compared to MTE. However, in the majority of the subsequent epochs, MTE converges faster and achieves lower localization error. As the Mask Ratio increases, meaning more sub-channels become unavailable, the localization error of all models increases. This is natural because the volume of available information decreases. We present the minimum localization error achieved by all models under different Mask Ratio conditions in TABLE II. It can be



(a) Siyuan Hall: Mask Ratio 0% (b) Siyuan Hall: Mask Ratio 50%



(c) Xuehuo 3F: Mask Ratio 0% (d) Xuehuo 3F: Mask Ratio 50%

Fig. 10: The evaluation average RMSE of versatile EM coefficients configured RT simulation results used for fingerprints in evaluation set.

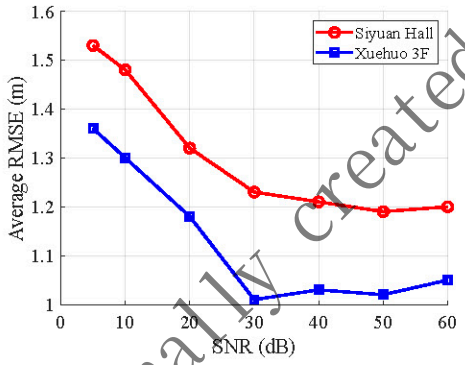


Fig. 11: The evaluation error at various SNR conditions after 200 epochs training, Mask Ratio 50%.

found that under Mask Ratio 0% - 50%, MTE out-performs all the baselines. Similarly, the training curves under different Mask Ratio conditions, as shown in Fig. 6 and Fig. 7, along with the cumulative distribution function (CDF) of evaluation RMSE in Fig. 8, also reflect the superiority of the MTE model. We estimate that the better localization ability originated from the better ability to aggregate information from other sub-channels. Under the Mask Ratio of 75%, the localization error of the MTE model is not the least one. However, all models exhibit poor localization precision under such a high Mask Ratio, making comparative evaluation of limited significance.

Comparing MTE horizontally with various baselines, it can be observed that when the Mask Ratio is small (0%, 25%), the residual connection-based CNN, ATRes model, also achieves high localization precision. However, as the Mask Ratio increases and more sub-channel information becomes unavailable, the performance of CNN models (AARes and ATRes) deteriorates. Simply adding attention layers (AARes) does not effectively improve localization performance. It's explainable because CNNs do not possess the ability to aggregate information from other sub-channels. For non time-series models like AARes and ATRes, we choose to input only the CSI data from the available sub-channels and adjust the input layer parameters accordingly. This approach is necessary because, if we follow the data concatenation procedures in Eq. 14, the training results would be unsatisfying, sometimes even failing to converge. This is exactly the difficulty in achieving surjection, as introduced in the Section I. Additionally, for time-series DL models such as LSTM and RNN, the performance is average when the Mask Ratio is low, but as the ratio increases, the decline in localization precision is not as pronounced as in CNN models. As for the Swin-Transformer, despite its advantages in computer vision, its performance is relatively poor for the data and tasks in this paper. It could be more appropriate to show its strengths under MIMO conditions.

To be noted that due to differences in datasets and the unavailability of code for the models cited in the aforementioned literature, our analysis and summary only focus on comparing different categories of models, rather than comparing with the specific models proposed in any single paper.

C. Comparisons of Alternative Mask Modeks

In addition to the proposed MTE in this paper, there are several Transformer-based alternative methods that can be used for 3D localization with partial-frequency CSI data as input:

- Masked Auto Encoder (MAE): Literature [40] proposed the original concept of MAE, successfully applying the BERT-like mask mechanism from NLP to computer vision. In fact, the model in this paper is quite similar to MAE, both using multi-layer Transformer encoders. However, there is a fundamental difference: MAE inputs only the available data, whereas in this paper, the unavailable sub-channel CSI data is set to a minimal value and combined with the CSI data received by the Rx before being input into the proposed MTE model. In the experiments of this paper, if the training method of MAE from [40] is adopted, i.e., randomly inputting partial CSI sub-channel data for training, the model almost cannot converge. Therefore, in the MAE experiment group, we use fixed sub-channel data as input. For example, the total set of sub-channels K is {0, 1, 2, 3, 4, 5, 6, 7, 8}, and the input set K' can be {0, 1, 2, 3}, noted as MTE 0123, or K' can be {0, 2, 4, 6}, noted as MTE 0246.
- Transformer: Transformer includes multiple layers of encoders and decoders [23], whereas the MTE used in this paper contains only multiple layers of encoders. In the experiments of this paper, the Transformer's decoder is fed with both the encoder's input and output, closely mimicking the configuration of a machine translation task. The Transformer uses 6 layers for both the encoder and decoder, and the remaining configurations are kept as consistent as possible with those used in MTE.

From Fig. 9, it can be seen that whether using fixed input sub-channel CSI data or using a Transformer, the performance is inferior to our proposed MTE. Considering the application of the model to real localization systems, methods such as MTE 0246 and MTE 0123, as well as CNNs like AARes and ATRes, require training a total of $C_8^4 = 70$ models for each sub-channel combination. The training costs and storage space will be significantly greater than what is required for MTE.

D. Analysis of Model Performances

In the above two subsections, we introduced the comparison of localization precision between our proposed MTE and various baselines and alternative methods. In this section, we will analyze in depth from the perspective of DL model performance. We provide the storage space required for model parameters, training time, and inference time for MTE and baseline models in Table III. It can be concluded that the training time and parameter storage space required for the MTE model are higher than most other models. However, considering the CNN models (ATRes and AARes) that can achieve high accuracy, as mentioned in Section V. III, it is necessary to train and store several models for different sub-channel combinations. In contrast, MTE only needs to train one model to achieve localization for any sub-channel combination. This gives MTE a significant advantage in terms of low-cost training and model storage.

We trained and analyzed the localization performance of the models under different signal-to-noise ratio (SNR) conditions, as shown in Fig. 11. It can be seen that the localization accuracy becomes consistent when the SNR is greater than 30 dB. The 40 dB SNR used in this paper is within the required SNR range for wireless communication systems to operate and fully leverages the localization performance of the models, making it reasonable.

As introduced in Section IV. B, we also used four different sets of EM coefficients, making a total of five sets, to generate fingerprint data using the RT simulator. Among them, em_0 is the set of EM coefficients used in other experiments of this paper, while em_1 - em_4 are the adjusted EM coefficients. We used the fingerprint data generated with the em_0 as the training set and the data from em_1 - em_4 as the evaluation set to achieve data isolation between training and testing. The training curves are provided in Fig. 10. It can be seen that eventhough there exist EM coefficient variations, the localization performance difference is at most about 20%. For the cases of em_1 and em_2, the training curves show very little difference compared to the results when using em_0 as the evaluation set. Even when swapping EM coefficients in RT simulations, which means setting the EM coefficients incorrectly as in em_3 and em_4, the performance difference is at most 20%. This indicates that the MTE model has strong robustness, and it also suggests that the channel data generated by RT simulation has the potential to replace real measurement data for fingerprinting localization. In the future, we will further explore this topic.

VI. Conclusion

In this study, we introduced a novel mask Transformer encoder model for 3D indoor localization utilizing partial-frequency CSI fingerprints. Through comprehensive experiments, we showed that the MTE model consistently outperforms traditional CNN, LSTM, RNN, and Swin-Transformer models, particularly when dealing with varying Mask Ratios. We further validated the robustness of our model under different SNR conditions and across multiple sets of electromagnetic parameters, reinforcing the practical applicability of our approach. The MTE model's ability to provide high localization accuracy with reduced training and storage requirements underscores its potential for real-world deployment in wireless communication systems. Future work will focus on extending the MTE model to handle more complex scenarios and exploring its integration with other advanced localization techniques.

References

[1] B. Yang, X. Liang, S. Liu, Z. Jiang, J. Zhu, and X. She, "Intelligent 6g wireless network with multi-dimensional information perception," *ZTE Communications*, vol. 21, no. 2, p. 3, 2023.

[2] R. Liu, M. Hua, K. Guan, X. Wang, L. Zhang, T. Mao, D. Zhang, Q. Wu, and A. Jamalipour, "6g enabled advanced transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–17, 2024.

[3] P. Gao, L. Lian, and J. Yu, "Cooperative isac with direct localization and rate-splitting multiple access communication: A pareto optimization framework," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 5, pp. 1496–1515, 2023.

[4] W. Fang, Y. Wang, H. Zhang, and N. Meng, "Optimized communication resource allocation in vehicular networks based on multi-agent deep reinforcement learning," *Beijing Jiaotong University*, vol. 46, no. 02, pp. 64–72, 2022.

[5] B. Hua, H. Ni, Q. Zhu, C.-X. Wang, T. Zhou, K. Mao, J. Bao, and X. Zhang, "Channel modeling for uav-to-ground communications with posture variation and fuselage scattering effect," *IEEE Transactions on Communications*, 2023.

[6] X. Yang, Y. Zhuang, F. Gu, M. Shi, X. Cao, Y. Li, B. Zhou, and L. Chen, "Deepwipos: A deep learning-based wireless positioning framework to address fingerprint instability," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 6, pp. 8018–8034, 2023.

[7] S.-W. Ko, H. Chae, K. Han, S. Lee, D.-W. Seo, and K. Huang, "V2x-based vehicular positioning: Opportunities, challenges, and future directions," *IEEE Wireless Communications*, vol. 28, no. 2, pp. 144–151, 2021.

[8] X. Wang, K. Guan, D. He, A. Hrovat, R. Liu, Z. Zhong, A. Al-Dulaimi, and K. Yu, "Graph neural network enabled propagation graph method for channel modeling," *IEEE Transactions on Vehicular Technology*, pp. 1–11, 2024.

[9] D. He, Z. Xu, H. Can, Y. Yin, L. Wu, and K. Guan, "Path loss prediction based on machine learning and satellite image," *Chinese journal of radio science*, vol. 37, no. 3, p. 8, 2022.

[10] C. Wu, X. Yi, W. Wang, L. You, Q. Huang, X. Gao, and Q. Liu, "Learning to localize: A 3d cnn approach to user positioning in massive mimo-ofdm systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4556–4570, 2021.

[11] B. Zhang, H. Sifaou, and G. Y. Li, "Csi-fingerprinting indoor localization via attention-augmented residual convolutional neural network," *IEEE Transactions on Wireless Communications*, 2023.

[12] X. Xu, F. Zhu, S. Han, Z. Yu, H. Zhao, B. Wang, and P. Zhang, "Swinloc: Transformer-based csi fingerprinting indoor localization with mimo isac system," *IEEE Transactions on Vehicular Technology*, 2024.

[13] Y. Li, X. Hu, Y. Zhuang, Z. Gao, P. Zhang, and N. El-Sheimy, "Deep reinforcement learning (drl): Another perspective for unsupervised wireless localization," *ieee internet of things journal*, vol. 7, no. 7, pp. 6279–6287, 2019.

[14] C. Yu, J.-P. Sheu, and Y.-C. Kuo, "Broad learning system for indoor csi fingerprint localization," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2023, pp. 1–6.

[15] J. Yan, C. Ma, B. Kang, X. Wu, and H. Liu, "Extreme learning machine and adaboost-based localization using csi and rssi," *IEEE Communications Letters*, vol. 25, no. 6, pp. 1906–1910, 2021.

[16] G. Tian, I. Yaman, M. Sandra, X. Cai, L. Liu, and F. Tufvesson, "Deep-learning based high-precision localization with massive mimo," *IEEE Transactions on Machine Learning in Communications and Networking*, 2023.

[17] S. A. Junoh and J.-Y. Pyun, "Enhancing indoor localization with semi-crowdsourced fingerprinting and gan-based data augmentation," *IEEE Internet of Things Journal*, 2023.

[18] Y. Etiabi, W. Njima, and E. M. Amhoud, "Federated learning based hierarchical 3d indoor localization," in *2023 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2023, pp. 1–6.

[19] A. Gharib, W. Ejaz, and M. Ibnkahla, "Distributed spectrum sensing for iot networks: Architecture, challenges, and learning," *IEEE Internet of Things Magazine*, vol. 4, no. 2, pp. 66–73, 2021.

[20] Y. Li, W. Zhang, C.-X. Wang, J. Sun, and Y. Liu, "Deep reinforcement learning for dynamic spectrum sensing and aggregation in multi-channel wireless networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 6, no. 2, pp. 464–475, 2020.

[21] S. Bhattarai, J.-M. J. Park, B. Gao, K. Bian, and W. Lehr, "An overview of dynamic spectrum sharing: Ongoing initiatives, challenges, and a roadmap for future research," *IEEE Transactions on Cognitive Communications and Networking*, vol. 2, no. 2, pp. 110–128, 2016.

[22] Z. Zhao, T. Chen, F. Meng, H. Li, X. Li, and G. Zhu, "Finding the missing data: A bert-inspired approach against package loss in wireless sensing," *arXiv preprint arXiv:2403.12400*, 2024.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] A. Salihu, M. Rupp, and S. Schwarz, "Self-supervised and invariant representations for wireless localization," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2024.

[25] Z. Gao, L. Li, H. Wu, X. Tu, and B. Han, "A unified deep learning method for csi feedback in massive mimo systems," *ZTE COMMUNICATIONS*, vol. 20, no. 4, 2023.

[26] X. Wang, L. Gao, and S. Mao, "Csi phase fingerprinting for indoor localization with a deep learning approach," *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 1113–1123, 2016.

[27] C. Wu, X. Yi, W. Wang, L. You, Q. Huang, X. Gao, and Q. Liu, "Learning to localize: A 3d cnn approach to user positioning in massive mimo-ofdm systems," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, pp. 4556–4570, 2021.

[28] W. Zhang, Y. Wang, X. Chen, Z. Cai, and Z. Tian, "Spectrum transformer: An attention-based wideband spectrum detector," *IEEE Transactions on Wireless Communications*, 2024.

[29] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[30] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 758–10 768.

[31] H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "Maskgit: Masked generative image transformer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 315–11 325.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[33] H. Zhang, F. Li, H. Xu, S. Huang, S. Liu, L. M. Ni, and L. Zhang, "Mp-former: Mask-piloted transformer for image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 074–18 083.

[34] A. Zayat, M. A. Hasabelnaby, M. Obeed, and A. Chaaban, "Transformer masked autoencoders for next-generation wireless communications: Architecture and opportunities," *IEEE Communications Magazine*, 2023.

[35] "Ieee draft standard for information technology–telecommunications and information exchange between systems local and metropolitan area networks–specific requirements - part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications amendment: Enhancements for extremely high throughput (eht)," *IEEE P802.11be/D3.0, January 2023*, pp. 1–999, 2023.

[36] D. He, K. Guan, D. Yan, H. Yi, Z. Zhang, X. Wang, Z. Zhong, and N. Zorba, "Physics and ai-based digital twin of multi-spectrum propagation characteristics for communication and sensing in 6g and beyond," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 11, pp. 3461–3473, 2023.

[37] D. He, B. Ai, K. Guan, L. Wang, Z. Zhong, and T. Kürner, "The design and applications of high-performance ray-tracing simulation platform for 5g and beyond wireless communications: A tutorial," *IEEE Communications Surveys Tutorials*, vol. 21, no. 1, pp. 10–27, 2019.

[38] R. Liu, Z. Wan, J. Wang, H. Zhou, J. Li, D. Wang, and X. You, "Fingerprint-based 3d hierarchical localization for cell-free massive mimo systems," *IEEE Transactions on Vehicular Technology*, 2024.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[40] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.