Learning-Based Tracking-before-Detect for RF-Based Unconstrained Indoor Human Tracking

Zhi Wu[1] , **Dongheng Zhang**[1] , **Zixin Shang**[1] , **Yuqin Yuan**[1] , **Hanqin Gong**[1] , **Binquan Wang**[1] , **Zhi Lu**[1] , **Yadong Li**[1] , **Yang Hu**[1] , **Qibin Sun**[1,2] and **Yan Chen**[1,2*]

[1] School of Cyber Science and Technology, University of Science and Technology of China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
wzwyyx@mail.ustc.edu.cn, dongheng@ustc.edu.cn, {zxshang, yuanyuqin, hanqin_gong}@mail.ustc.edu.cn, {wbq0556, zhilu}@ustc.edu.cn, yadongli@mail.ustc.edu.cn, {eeyhu, qibinsun, eecyan}@ustc.edu.cn

Abstract

Existing efforts on human tracking using wireless signal are primarily focused on constrained scenarios with only a few individuals in empty spaces. However, in practical unconstrained scenarios with severe interference and attenuation, accurate multi-person tracking has been intractable. In this paper, we propose NeuralTBD, utilizing the capability of deep models and advancement of Tracking-Before-Detect (TBD) methodology to achieve accurate human tracking. TBD is a classical tracking methodology from signal processing accumulating measurement in time domain to distinguish target traces from interference, which however relies on handcrafted shape/motion models, impeding efficacy in complex indoor scenarios. To tackle this challenge, we build an end-to-end learning-based TBD framework leverages the advanced modeling capabilities of deep models to significantly enhance the performance of TBD. To evaluate NeuralTBD, we collect an RF-based tracking dataset in unconstrained scenarios, which encompasses 4 million annotated radar frames with up to 19 individuals acting in 6 different scenarios. NeuralTBD realizes a 70% improvement in performance compared to conventional TBD methods. To our knowledge, this is the first attempt dealing with RF-based unconstrained human tracking. The code and dataset will be released.

1 Introduction

Passive human tracking which predicts multiple target trajectories without requiring body-worn sensors, is one of the most fundamental yet challenging topics in wireless sensing [Thormann *et al.*, 2018; Zhang *et al.*, 2019a; Zhang *et al.*, 2019b; He *et al.*, 2020]. Previous efforts employ thresholding on signal measurement to maintain high tracking accuracy. However, hard-thresholding presents severe performance degradation encountering complex indoor scenarios. Higher threshold leads to increased number of false alarms, referring to false positive tracks, while lower threshold leads to the loss of true positives. To resolve this problem, researchers adopted classical tracking-before-detect (TBD) method to aggregate information of data sequence to distinguish targets from interference [Grossi *et al.*, 2013; Jiang *et al.*, 2017; Zhou *et al.*, 2019].

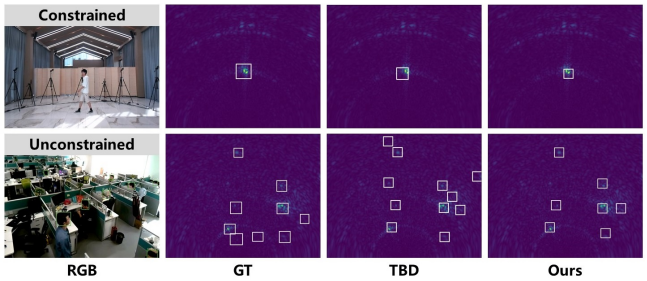

Figure 1: Comparisons of TBD method and our proposed NeuralTBD. Each column depicts the RGB view, RF heatmap, TBD output, and NeuralTBD output sequentially. The top row illustrates tracking results in constrained scenes (a single person in an obstacle-free space), where both TBD and our proposed NeuralTBD perform well. The bottom row illustrates tracking results in unconstrained scenes (multiple individuals acting in an obstacle-rich space), where TBD demonstrates significant performance degradation and NeuralTBD maintains consistent accuracy as constrained scenarios.

TBD accumulates trace-level measurement from trace proposals, and detects true target trajectory according to handcrafted shape/motion model. With deliberately designed model, TBD can accurately distinguish target trajectories from false alarms. As depicted in Figure 1, in constrained scenarios involving a few individuals moving in empty spaces, TBD achieves satisfying detection/tracking accuracy. However, in practical unconstrained scenarios involving significant human targets acting randomly in the space with various obstacles, TBD suffers from performance degradation. This is due to the fact that TBD relies on handcrafted signal model, which is difficult to handle time-varying distribution of target reflections in complex indoor scenarios. To address this problem, we build an end-to-end learning based TBD model which utilizes strong modeling capability of deep models to further unleash the power of TBD. This involves resolving

*The corresponding author is Yan Chen

two key challenges.

The primary challenge is how to achieve TBD through deep models. To aggregate temporal information, our NeuralTBD learn to accumulate inter-frame information with the help of trace center supervision, and then predicts trace offsets at each position to inference trace proposals. Specifically, NeuralTBD first extracts frame-wise deep features to adapt to the time-varying reflection distribution and motion pattern of targets. After that, it predicts per-pixel probabilities for trace centers and temporal offset sequences to generate trace proposals. Final traces are then extracted by filtering out false positives from these proposals.

The second challenge lies in the absence of a public dataset for unconstrained indoor human tracking. The main reason is that RF signal is not human readable and hard to annotate, especially in scenes with severe interference. To address this, we present the RF-UNIT (**RF**-based **U**nconstrained **I**ndoor **T**racking) dataset. It contains 4,030,880 radar heatmaps (about 56 hours) collected under 6 different office scenes with at most 19 individuals performing daily life activities. To alleviate the workload of manual labeling, we design an annotation algorithm, utilizing well-developed vision-based methods to annotate RF heatmaps. This allows us to generate full tracking annotations for the RF-UNIT dataset. We believe that the release of RF-UNIT would encourage more innovations on RF-based sensing. The main contributions of this paper can be summarized as follows:

1. To the best of our knowledge, we are the first to tackle the RF-based human tracking problem under unconstrained indoor scenarios. By leveraging the advancement of deep models and TBD, we expand the scope of RF-based tracking and provide novel insights of aggregating target information for real-world applications.
2. We propose NeuralTBD, which is a brand-new learning-based tracking-before-detect framework, utilizing deep models to adapt to time-varying target reflection distribution while preserving the temporal information aggregation capabilities of TBD. NeuralTBD realizes a 70% improvement in performance compared to conventional TBD methods.
3. We present the RF-UNIT dataset, which encompasses million-level radar heatmaps of at most 19 individuals in multiple different scenarios. To our knowledge, RF-UNIT is the first fully-annotated large-scale RF dataset for indoor human tracking in unconstrained indoor scenarios.

2 Related Works

2.1 Tracking-Before-Detect

In contrast to detect-then-track techniques that typically impose frame-wise detection on input data to determine the presence of target and then perform tracking. Track-Before-Detect (TBD) methods engage with either raw data or minimally processed data and output target trajectories. TBD methods can be classified into single-frame recursive TBD (SFR-TBD) method and multi-frame TBD (MF-TBD) method.

SFR-TBD methods estimate target states at each time step by sequentially predicting and updating intermediate parameters which are then used to construct target trajectories. Typical algorithms include particle filters [Garcia-Fernandez *et al.*, 2013], histogram-PMHT [Davey, 2014], and random finite set algorithms [Hoseinnezhad *et al.*, 2012]. MF-TBD [Jiang *et al.*, 2017; Zhou *et al.*, 2019] methods accumulate target energy along physically feasible trajectories between multiple consecutive frames achieving superior tracking performance under interference. Typical MF-TBD approaches include the Hough transform [Moyer *et al.*, 2011], maximum likelihood probabilistic data association (ML-PDA) [Ciuonzo *et al.*, 2014], velocity matched filtering [Zhou and Wang, 2019], and dynamic programming (DP) [Zhang *et al.*, 2021].

Despite yielding promising results, TBD methods still lack robustness in handling complex scenarios due to handcrafted shape/motion model priors, resulting in degraded performance in unconstrained indoor tracking scenarios.

2.2 Learning-Based Tracking

Existing learning-based tracking algorithms can be classified into separate detection and embedding (SDE) style and joint detection and embedding (JDE) style. SDE algorithms [Wojke *et al.*, 2017; Wojke and Bewley, 2018] separates the tasks into detection and embedding, while JDE models [Zhang *et al.*, 2020b] learning detection and embedding simultaneously within a shared neural network. Building upon high-accuracy detectors, such as [Ren *et al.*, 2017; Redmon *et al.*, 2016; Zhu *et al.*, 2020], SDE style algorithms dominate learning-based tracking. Both SDE and JDE style tracking methods fundamentally follow the tracking by detection pipeline which highly relies on the performance of detectors.

Although remarkable results have been achieved by learning-based tracking methods, directly adopt them into our situation is infeasible. Specifically, in unconstrained human tracking scenarios using RF signals, severe interference lead to significant false alarms which pose challenges in frame-wise detection of targets. Therefore, we proposed to utilize the advancement of TBD and the strong modeling capability of deep models to aggregate temporal information and achieve accurate tracking in unconstrained scenarios.

2.3 RF-Based Dataset

To advance the development of learning-based approaches in wireless sensing, efforts has been made in building RF dataset of various applications. Examples include object detection [Caesar *et al.*, 2020; Wang *et al.*, 2021b], action recognition [Singh *et al.*, 2019; Sengupta *et al.*, 2020], pose/mesh prediction [Sengupta *et al.*, 2020; Xue *et al.*, 2021; Chen *et al.*, 2022; Wu *et al.*, 2022], gait [Meng *et al.*, 2020], gesture [Palipana *et al.*, 2021], and rehabilitation [An and Ogras, 2021; An *et al.*, 2022]. Existing RF dataset focused on constrained scenarios involving limited number of individuals performing scheduled actions in an empty space without obstacles. However, things are different in practical unconstrained indoor sensing situation. The presence of dozens of individuals and obstacles poses great challenges in sensing tasks.

To facilitate the development of learning-based methods in unconstrained scenarios, we collect RF-UNIT, which as we

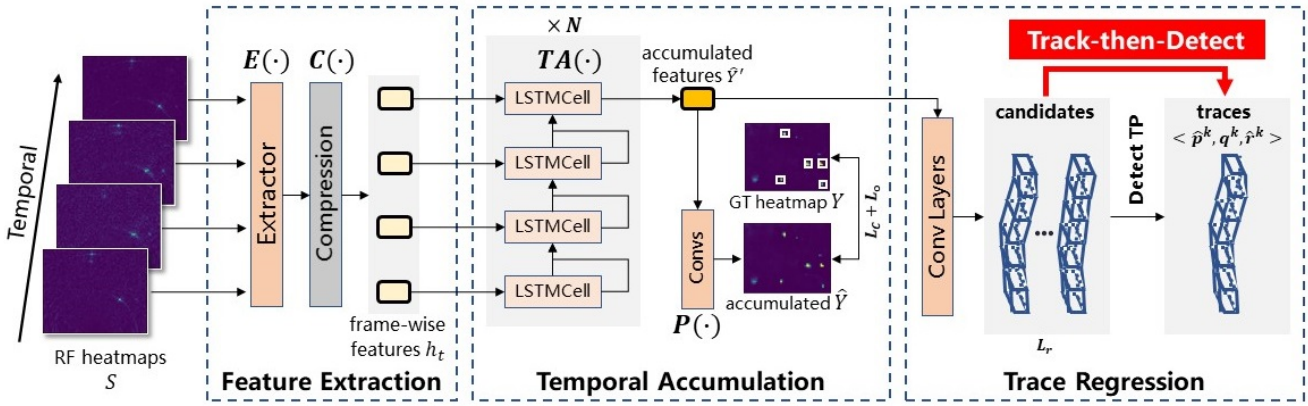

Figure 2: The architecture of NeuralTBD. NeuralTBD consists of three components: feature extraction module, temporal accumulation module and trace regression module. Firstly, the feature extraction module extracts frame-wise features from RF heatmaps to fit targets' shape/motion model. Secondly, the temporal accumulation module fuses temporal information to generate temporal accumulated features. Finally, the trace regression module inference trace proposals followed by thresholding to filter out false alarms and outputs final traces. As shown in this figure, our accumulated heatmap $\hat{Y}$ clearly depicts the positions of targets that are difficult to distinguish in the input heatmaps.

know is the first large-scale RF datasets for unconstrained indoor human tracking scenarios. RF-UNIT dataset comprises 4,030,880 radar heatmaps collected under six office scenes with at most 19 individuals, each paired with corresponding tracking annotations. We hope RF-UNIT can offer valuable insights for designing RF-based deep learning models.

3 NeuralTBD

As depicted in Figure 2, NeuralTBD is an end-to-end, three-stage, learning-based TBD framework that comprises a feature extraction module, a temporal accumulation module, and a trace regression module. It takes RF heatmap sequence of input and fits targets' shape/motion model with feature extraction module. And then, it utilizes temporal accumulation module to aggregate temporal information and further augment targets features. At last, it adopts trace regression model to predict trace proposals followed by thresholding to eliminate false positives and produce final tracking results. In the following subsections, we provide an in-depth exploration of each components.

3.1 Feature Extraction

The biggest problem brought by interference in wireless sensing is multi-path effect, which is a fundamental and challenging problem in RF-based sensing applications. It becomes more challenging in unconstrained scenarios (involving multiple individuals acting in an obstacle-rich space), where the reflections from multi-paths can be stronger than that from targets. There are generally two kinds of multi-path interference, static multi-paths and dynamic multi-paths, with different characteristics. Static multi-paths refer to reflections from stationary objects in the environment, whereas dynamic multi-paths are caused by the movement of individuals, as depicted in Figure 3. To alleviate static multi-paths, we leverage the distinct characteristics of reflected signals from moving targets and static objects in the time domain.

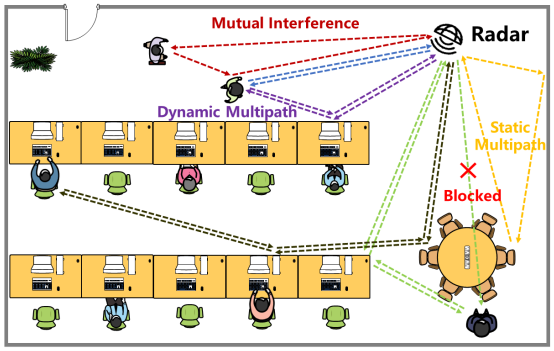

Figure 3: Illustration of the static and dynamic multi-path effect in unconstrained indoor environment. Static multi-path is caused by stationary objects in the environment, such as workstations, computers, and other items. Dynamic multi-path is caused by human activities and motion occlusions.

Concretely, the reflections from moving individuals vary over time, while the reflections of static multi-paths remain consistent. Assuming we have a heatmap sequence $S' = \{s'_t\}$, where t is the time index, we can remove static multi-paths by subtracting consecutive frames in time, as represented in the following equation:

$$S = \{s_t | s_t = s'_{t+1} - s'_t\}. \quad (1)$$

On the other hand, signals to and from targets may be occluded, refracted and scattered by other moving objects in unconstrained scenarios, leading to significant alterations in the distribution of reflections. This presents a formidable challenge for TBD methods in accurately identifying target traces due to limited modeling capability of handcrafted shape/motion priors. To overcome this challenge, we propose to utilize the modeling capability of deep models to adapt to the dynamic changes of targets' reflections.

Specifically, given an input heatmap sequence $S \in$

Dataset	Scene	n-frames	annotation	modality	max n-subs/frame	application
mRI	indoor	160k	2D/3D keypoints, categories	camera, depth, IMU, radar	1	indoor sensing
MARS	indoor	40k	2D/3D keypoints, categories	radar	1	rehabilitation
mmPose	indoor	40k	2D/3D keypoints, categories	radar	1	indoor sensing
mmBody	indoor	200k	3D keypoints	3D keypoints, mesh	1	indoor sensing
mmActivity	indoor	16k	categories	radar	1	indoor sensing
mmMesh	indoor	3k	3D keypoints, categories	camera, radar	1	indoor sensing
mmGait	indoor	1080k	2D point-wise, track id	radar	5	gait
HuPR	indoor	14k	2D/3D keypoints	radar, camera	1	indoor sensing
Radar Scenes	outdoor	23k	2D point-wise, track id, categories	radar, camera	7	autonomous driving
CRUW	outdoor	396k	heatmap, categories	radar	5	autonomous driving
HIBER	indoor	800k	2D/3D keypoints	radar, camera	2	pose
RF-UNIT(Ours)	indoor	**4030k**	2D bbox, track id	radar	**19**	indoor sensing

Table 1: RF-based Dataset Comparisons

$\mathbb{R}^{H\times W\times T}$, we adopt a multi-layer convolution network with skip connections as backbone extractor $\boldsymbol{E}$, producing frame-wise features $h_t \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times C}$ for each frame s_t, where R is the down-sampling ratio. These features are then passed through a compression network $\boldsymbol{C}$, which transforms high-dimensional but sparse features into low-dimensional and dense features to condense target information and outputs a compressed feature sequence H'. Therefore, our feature extraction module can be represented as

$$h_t = \boldsymbol{E}(s_t), \tag{2}$$

$$H' = \{h'_t | h'_t = \boldsymbol{C}(h_t), t = 0, ..., T\}. \tag{3}$$

3.2 Temporal Accumulation

Besides the dramatic changes in the target's reflections, severe dynamic multi-paths also result in significant false alarms with identical reflection distributions as targets, making them difficult to distinguish.

As the number of individuals increases, the occlusion and refraction of signal is increased due to mutually interference between moving individuals which introduces greater noise and intensifies the false-alarms issues. The main difference between true targets and false alarms is their continuity over time. Concretely, target positions change continuously due to limited moving speed. However, the positions of false alarms undergo rapid and erratic variations, influenced by the arrangement of objects within the environment.

TBD aggregates temporal information to distinguish target measurement from noise. Inspired by that, we propose a learning based temporal aggregation module to amplify target feature and suppress noise. Our temporal accumulation module comprises a temporal accumulator $\boldsymbol{TA}$ and a propose head $\boldsymbol{P}$. The temporal accumulator integrates target information over an extended time period through multiple Long-Short-Term-Memory (LSTM) layers. The propose head employs multiple convolution layers to predict a confidence map, indicating potential trace centers.

Specifically, given the compressed feature sequence H', we first pass it through $\boldsymbol{TA}$ followed by mean operation across LSTM layers to obtain temporal-fused sequence features $\hat{Y}'$. To supervise the temporal accumulation and target augmentation process of $\boldsymbol{TA}$, we introduce propose head $\boldsymbol{P}$, generating a confidence heatmap $\hat{Y} \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}}$. Each $\hat{Y}_{x,y} = 1$ corresponds to a potential trace center, while $\hat{Y}_{x,y} = 0$ represents the background. Meanwhile, we put each scaled ground-truth trace centers onto a heatmap $Y \in [0,1]^{\frac{H}{R}\times\frac{W}{R}}$ using a Gaussian kernel followed by the element-wise maximum between any overlapped gaussians.

Our temporal accumulation module can be represented as

$$\hat{Y}' = Mean(\boldsymbol{TA}(H')), \tag{4}$$

$$\hat{Y} = \boldsymbol{P}(\hat{Y}'). \tag{5}$$

To address the significant imbalance between the number of positive and negative samples, we take the focal loss, noted as L_c, as training objective of $\boldsymbol{P}$.

To recover the discretization error caused by the output stride R, we additionally predict an offset $\hat{O} \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times 2}$ for each candidate trace center. The training objective of offset prediction can be expressed as

$$L_o = \frac{1}{N}\sum_c ||O_c - \hat{O}_c||_2. \tag{6}$$

By supervising the confidence map $\hat{Y}$ generated by $\boldsymbol{TP}$ with ground-truth confidence maps Y, the temporal accumulation module will learn to distinguish regions containing target motions from those not by aggregating and augmenting target information in time domain.

3.3 Trace Regression

Build upon the temporally aggregated features, we are able to inference per-position trace proposals using trace regression module. Our trace regression module composed of multiple convolution layers, to predict the positional offset relative to trace center at each timestamp. Specifically, our trace regression head takes compressed features $\hat{Y}' \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times C}$ as input, predict positional offsets $Q \in \mathbb{R}^{\frac{H}{R}\times\frac{W}{R}\times T\times 2}$ for traces centered at each position at each timestamp, where T denotes the length of input sequence, and the last dimension indicates horizontal and vertical offsets, respectively. We adopt $L1$ loss as training objective of trace regression module, denoted as L_r. After training is finished, we threshold trace proposals according to predicted confidence scores to remove false alarms and output target traces.

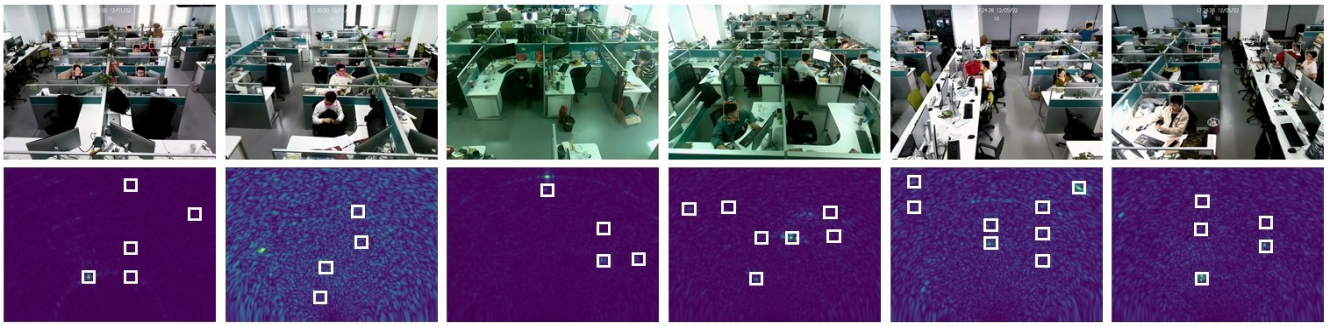

Figure 4: Data samples of our collected RF-UNIT dataset. The first row demonstrates the RGB images of six office scenarios. The second row illustrates corresponding RF heatmaps and tracking annotations. Each frame spans a spatial region of $9m \times 10m$. Our dataset consists of different number of individuals and obstacles leading to severe interference.

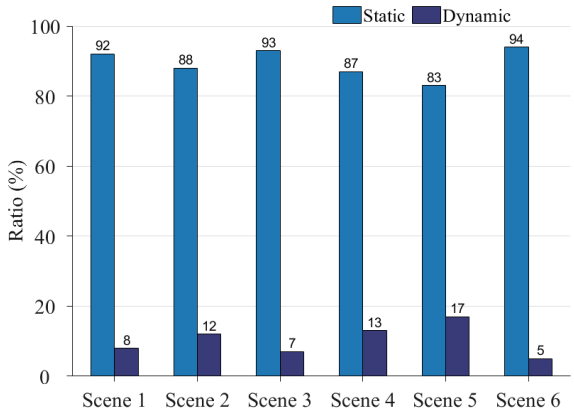

Figure 5: Statistics for the ratio of frame-level static and dynamic human targets in six environments. In practical office scenarios, individuals typically remain in one position for most of the time.

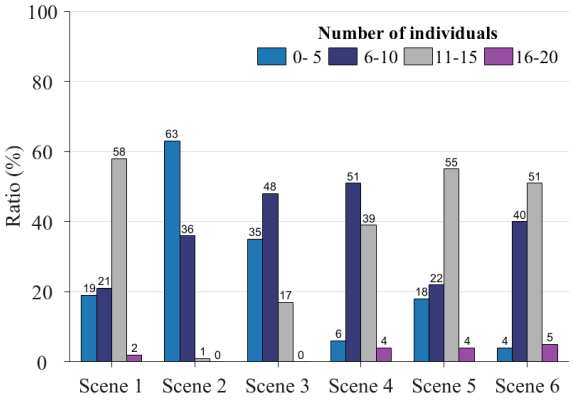

Figure 6: Statistics for the number of individuals in each scene expressed as ratios. Predominantly, the frames contains between 6 to 15 individuals. Notably, the maximum number of individuals reaches 19, surpassing the counts observed in other RF datasets.

The overall objective is formulated as follows:

$$L = L_c + L_o + L_r. \tag{7}$$

Leveraging the modeling capacities of deep model and the advancement of TBD, NeuralTBD learn to aggregate temporal information to overcome the severe interference without handcrafted efforts, and effectively predict target traces.

4 Dataset

Learning-based methods show promise to outperform traditional methods but are often limited by substantial data needs. While many RF-based datasets have been introduced , they typically cover a narrow range of scenarios with few human subjects, leaving a gap in data for unconstrained indoor environments.

Addressing this, we introduce RF-UNIT, a large-scale RF dataset for human tracking in diverse, unconstrained indoor settings. RF-UNIT comprises 4,030,880 radar frames from six office environments and includes scenes with up to 19 individuals, surpassing other public datasets that typically contains only single individual. We provide an extensive comparison with other RF datasets in Table 1. As can be seen in Figure 4, RF-UNIT is specifically tailored for real-world office environments, featuring a variety of obstacles and materials, including wood and metal, that affect signal propagation and cause notable interference. We detail individual counts per scene in Figure 6 as well as dynamic/static counts per scene in Figure 5. Most of the heatmaps in RF-UNIT contains 5-15 individuals. RF-UNIT stands as the first dataset of million-level scale for multi-person RF tracking in unconstrained scenarios.

We will introduce the hardware setup, data collection, data processing and automatic annotation methods next.

4.1 Hardware Settings

We use a TI MMWCAS-RF-EVM FMCW radar with 12 transmitters and 16 receivers offering 1.4-degree horizontal resolution, working at 77GHz and 1.23GHz bandwidth, mounted at 2.2m height. To minimize human annotation efforts, we employ an 18-node Raspberry Pi camera network, calibrated with Zhang's 2000 algorithm [Zhang, 2000], to automatically annotate data aided by vision-based techniques.

4.2 Data Collection

Deploying the well-calibrated multi-camera system in new environments can be highly time-consuming. To build a diverse dataset efficiently, we re-positioned our radar within a pre-setup scene, simulating varied RF environments and reducing the time-intensive deployment of multi-camera system. We leveraged NTP for millisecond-level synchronization between the radar and camera systems, with TCP for timestamp signal exchanges. With the cameras running at 10 fps and the radar at 20 fps, we conducted 112 data collection from six unique locations at different time of a day, each lasting 30 minutes. The total collection spans 36 days.

4.3 Data Processing

We perform signal processing algorithm to transform RF signals captured by FMCW radars into heatmap sequences. Inspired by [Zhang *et al.*, 2018; Zhang *et al.*, 2019b; Zhang *et al.*, 2021], we compensate the phase shift and combine the signals of different antennas and frequencies, which coherently superimposes the signals from specific locations while suppressing the signals from other locations. In our tracking situation, we mainly concern about signals on horizontal plane. Specifically, RF signals from a specific location (x, y) can be extracted using the following equation:

$$s'_{hor}(x, y, t) = \sum_{k=1}^{K} \sum_{m=1}^{M} a_{k,m,t} \cdot e^{j2\pi \frac{d_m(x,y)}{\lambda_k}}. \quad (8)$$

where a denotes the amplitude of signal, $s_{k,m,t}$ denotes the k-th sample of FMCW sweep on the m-th antenna at time t, λ_k is the signal wavelength of the k-th sample, $d_m(x, y)$ denotes the round-trip distance from the transmitting antenna to location (x, y) and back to the receiving antenna.

4.4 Automatic Annotation

The massive collection of over 4 million RF frames, coupled with the severe interference encountered in unconstrained environments, makes manual annotation a daunting and time-consuming task.

To deal with this problem, we developed an automatic annotation algorithm composed of three key stages: camera-specific head detection, cross-camera triangulation, and clustering. Concretely, we employed YOLOv4 [Bochkovskiy *et al.*, 2020], trained on the crowd-human dataset, for head detection in footage from each camera. The centers of these detections are projected as rays emitting from respective cameras. By solving intersection points of rays from different cameras, we are able to estimate 3D human positions. Lastly, we refine these estimates by applying mean-shift clustering to mitigate noise followed by a Kalman Filter to construct target trajectories. We conduct manual annotation corrections to uphold the precision of the tracking annotations.

As shown in Figure 4, our dataset is collected under unconstrained practical indoor office scenarios, encompassing challenging situations, such as diverse obstacles, varying numbers of people (up to 19), overlapping between targets, and complex human activities. We believe our RF-UNIT dataset will aid in the advancement of learning-based methods of wireless sensing.

Method	AP↑	MOTA↑	MOTP↓	IDF1↑	IDP↑	IDR↑
MTrack	0.378	0.031	0.715	0.112	0.333	0.068
MF-TBD	0.521	0.130	0.861	0.151	0.423	0.092
MKCF-TBD	0.333	0.031	0.867	0.071	0.410	0.039
RODNet*	0.842	0.349	0.475	0.271	0.383	0.213
Ours	**0.885**	**0.420**	**0.465**	**0.448**	**0.586**	**0.372**

* indicates this method is followed by Kalman Filter.

Table 2: Comparisons between ours and baseline methods

	NeuralTBD							
SEQ LEN	6	12	24	48	6	12	24	48
TA					✓	✓	✓	✓
AP ↑	0.447	0.458	0.455	0.445	0.852	0.871	**0.885**	0.872
MOTA ↑	0.088	0.087	0.102	0.095	0.254	0.341	**0.420**	0.398
MOTP ↓	0.427	0.429	0.430	0.431	0.509	0.475	0.465	**0.420**
IDF1 ↑	0.152	0.211	0.268	0.342	0.246	0.323	0.448	**0.533**
IDP ↑	0.348	0.427	0.493	0.582	0.342	0.459	0.586	**0.670**
IDR ↑	0.099	0.143	0.190	0.250	0.196	0.257	0.372	**0.462**

Table 3: Ablation Experiments

5 Experiments

5.1 Implementation Details

In this section, we present the evaluation results of NeuralTBD on RF-UNIT dataset. We first preform group-wise shuffle on RF-UNIT and divide data into train, validation, and test subset, following 8:1:1 ratio. All evaluations are reported on test sets. NeuralTBD is designed to deal with RF heatmap sequence within a fixed time-window. Therefore, we adopt sliding window strategy to deal with long sequence. At each step of this process, we retain trace results that exhibit a higher confidence score and/or a lower Intersection over Union (IoU) with pre-existing traces, ensuring both accuracy and minimal redundancy. We employ the Adam optimizer with initial learning rate of 1.0×10^{-2} and weight decay of 0.05. During the training process, we adopt a step-based learning rate decay strategy. All experiments are conducted on a single NVIDIA A100 GPU with a batch size of 16.

5.2 General Performance

To showcase the performance of our NeuralTBD on the RF-UNIT dataset, we conducted comparative experiments against baseline methods, including both TBD and learning based methods:

- **MF-TBD** [Grossi *et al.*, 2013]: MF-TBD associates CFAR segmentation results across frames according to IoU and shape similarity to propose traces. And then it retains those with trace energy higher than a threshold.
- **MKCF-TBD** [Zhou *et al.*, 2019]: MKCF-TBD improves MF-TBD using kernel methods which better models target distribution of reflections.
- **MTrack** [Zhang *et al.*, 2020a]: MTrack is a graph-based TBD method. It constructs a directed graph based on CFAR segmentation results and subsequently solves for the shortest path to accomplish tracking tasks.

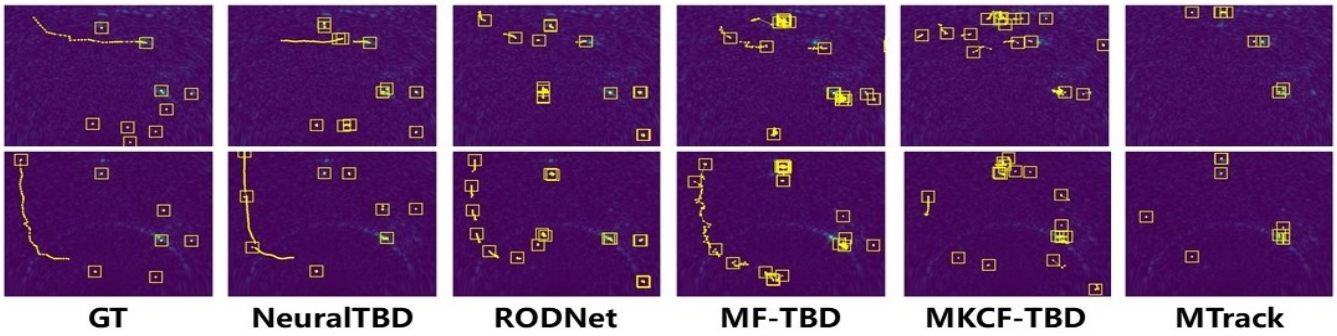

Figure 7: Qualitative results of comparisons. The boxes indicate the endpoints of traces, while the curves represent traces from the recent past. Our proposed NeuralTBD achieves higher detection/tracking precision and recall. It maintains traces for extended periods, whereas other methods mostly report trace fragments. NeuralTBD reports fewer false alarms. This exemplifies the significant improvement brought by combining learning-based and TBD.

- **RODNet** [Wang *et al.*, 2021a]: RODNet is a learning-based radar object detection method. We performs kalman filter on the detection results of RODNet to output trace predictions.

Table 2 shows that our NeuralTBD significantly outperforms traditional methods, achieving a higher AP of 0.885, compared to the substantially lower scores of MF-TBD (0.521), MKCF-TBD (0.333), MTrack (0.378), and RODNet (0.842). These results underline the limitations of conventional TBD techniques in detecting human targets indoors.

Furthermore, our method achieves a substantially higher IDF1 score of 44.8%, exceeding other methods. This signifies that NeuralTBD undergoes fewer trace mismatches and ID switches when processing long sequences. In addition, the MOTA value of 0.420 illustrates that NeuralTBD provides superior target-locating capabilities. These metrics highlight the efficacy and advancement of our learning-based TBD approach.

The qualitative results in Figure 7 further demonstrate the advantages of our method over baseline approaches. Baselines exhibit substantially more false alarms, with long target traces fragmented into shorter segments. In contrast, our approach produces fewer false alarms and successfully tracks targets over extended periods. This highlights the effectiveness of our method in leveraging deep models' robust representational capabilities to differentiate targets from noise and clutter. Moreover, it fully capitalizes on the benefits of TBD to accomplish long-term tracking in unconstrained settings.

5.3 Ablation Study

In this subsection, we provide ablation results on temporal accumulation module (denoted as TA) and on different input sequence length. Refer to Table 3 for the detailed experimental outcomes.

Temporal Accumulation Module

To assess the effectiveness of temporal accumulation module, we substituted it with CFAR segmentation followed by kalman filter to inference trace proposals, while keeping the other modules unchanged.

The experimental results demonstrate a substantial decrease in performance when the temporal accumulation module is omitted for all input sequence lengths. In particular, the AP score drops from 0.885 to 0.455, and the MOTA value decreases from 0.420 to 0.102 with an input sequence of 24 frames. This underscores the significant enhancement in trace localization achieved through temporal accumulation. Overall, these outcomes validate the effectiveness of our NeuralTBD approach.

Time Domain Analysis

Constrained by hardware limitations, our NeuralTBD accumulates information within a fixed time window, termed the accumulation window, to predict target motions. To further analyze the impact of the accumulation window length, we conduct experiments varying the input sequence duration. As shown in Table 3, NeuralTBD exhibits significant performance gains as the input sequence lengthens. For example, when the input sequence increases from 6 to 24 frames, the AP, MOTA, and IDF1 values rise from 0.852, 0.254, 0.246 to 0.885, 0.420, 0.448, respectively. This implies longer accumulation windows yield enhanced tracking accuracy.

However, NeuralTBD experiences a performance bottleneck once the sequence surpasses 24 frames. We assume this relates to the long-term modeling capacity of NeuralTBD.

6 Conclusion

This study explores human tracking of unconstrained indoor environments, introducing NeuralTBD which builds a brand-new learning-based TBD framework, utilizing strong modeling capability and advancement of TBD to aggregate temporal information and address multi-path effects. To train NeuralTBD, we also collected RF-UNIT, the first large-scale public RF dataset for unconstrained indoor human tracking, comprising 4 million heatmaps with at most 19 individuals collected from six different scenarios. NeuralTBD realizes a 70% improvement in performance compared to conventional TBD methods. RF-UNIT as well as NeuralTBD are valuable resources for propelling RF-based tracking and downstream applications, empowering the community to develop robust solutions for practical wireless settings. The dataset and code will be public.

Acknowledgements

This work was supported by National Key R&D Programmes under Grant 2022YFC2503405, National Natural Science Foundation of China under Grant 62201542, 62172381 and 62302471, the Fundamental Research Funds for the Central Universities, the fellowship of China Postdoctoral Science Foundation under Grant 2023M743401, and the Postdoctoral Fellowship Program of CPSF under Grant GZC20232565.

References

[An and Ogras, 2021] Sizhe An and Umit Y Ogras. Mars: mmwave-based assistive rehabilitation system for smart healthcare. *ACM Transactions on Embedded Computing Systems (TECS)*, 20(5s):1–22, 2021.

[An *et al.*, 2022] Sizhe An, Yin Li, and Umit Ogras. mri: Multi-modal 3d human pose estimation dataset using mmwave, rgb-d, and inertial sensors. *Advances in Neural Information Processing Systems*, 35:27414–27426, 2022.

[Bochkovskiy *et al.*, 2020] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv*, 2020.

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[Chen *et al.*, 2022] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmbody benchmark: 3d body reconstruction dataset and analysis for millimeter wave radar. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3501–3510, 2022.

[Ciuonzo *et al.*, 2014] Domenico Ciuonzo, Peter K Willett, and Yaakov Bar-Shalom. Tracking the tracker from its passive sonar ml-pda estimates. *IEEE Transactions on Aerospace and Electronic Systems*, 50(1):573–590, 2014.

[Davey, 2014] Samuel J Davey. Efficient histogram pmht via single target chip processing. *IEEE Signal Processing Letters*, 22(5):569–572, 2014.

[Garcia-Fernandez *et al.*, 2013] Angel F Garcia-Fernandez, Jesus Grajal, and Mark R Morelande. Two-layer particle filter for multiple target detection and tracking. *IEEE Transactions on Aerospace and Electronic Systems*, 49(3):1569–1588, 2013.

[Grossi *et al.*, 2013] Emanuele Grossi, Marco Lops, and Luca Venturino. A novel dynamic programming algorithm for track-before-detect in radar systems. *IEEE Transactions on Signal Processing*, 61(10):2608–2619, 2013.

[He *et al.*, 2020] Ying He, Yan Chen, Yang Hu, and Bing Zeng. Wifi vision: Sensing, recognition, and detection with commodity mimo-ofdm wifi. *IEEE Internet of Things Journal*, 7(9):8296–8317, 2020.

[Hoseinnezhad *et al.*, 2012] Reza Hoseinnezhad, Ba-Ngu Vo, and Ba-Tuong Vo. Visual tracking in background subtracted image sequences via multi-bernoulli filtering. *IEEE Transactions on Signal Processing*, 61(2):392–397, 2012.

[Jiang *et al.*, 2017] Haichao Jiang, Wei Yi, Thia Kirubarajan, Lingjiang Kong, and Xiaobo Yang. Multiframe radar detection of fluctuating targets using phase information. *IEEE Transactions on Aerospace and Electronic Systems*, 53(2):736–749, 2017.

[Meng *et al.*, 2020] Zhen Meng, Song Fu, Jie Yan, Hongyuan Liang, Anfu Zhou, Shilin Zhu, Huadong Ma, Jianhua Liu, and Ning Yang. Gait recognition for co-existing multiple people using millimeter wave sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 849–856, 2020.

[Moyer *et al.*, 2011] Lee R Moyer, Jeffrey Spak, and Peter Lamanna. A multi-dimensional hough transform-based track-before-detect technique for detecting weak targets in strong clutter backgrounds. *IEEE Transactions on Aerospace and Electronic Systems*, 47(4):3062–3068, 2011.

[Palipana *et al.*, 2021] Sameera Palipana, Dariush Salami, Luis A Leiva, and Stephan Sigg. Pantomime: Mid-air gesture recognition with sparse millimeter-wave radar point clouds. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(1):1–27, 2021.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[Ren *et al.*, 2017] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[Sengupta *et al.*, 2020] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 20(17):10032–10044, 2020.

[Singh *et al.*, 2019] Akash Deep Singh, Sandeep Singh Sandha, Luis Garcia, and Mani Srivastava. Radhar: Human activity recognition from point clouds generated through a millimeter-wave radar. In *Proceedings of the 3rd ACM Workshop on Millimeter-wave Networks and Sensing Systems*, pages 51–56, 2019.

[Thormann *et al.*, 2018] Kolja Thormann, Marcus Baum, and Jens Honer. Extended target tracking using gaussian processes with high-resolution automotive radar. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1764–1770. IEEE, 2018.

[Wang *et al.*, 2021a] Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu.

Rodnet: Radar object detection using cross-modal supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 504–513, 2021.

[Wang *et al.*, 2021b] Yizhou Wang, Gaoang Wang, Hung-Min Hsu, Hui Liu, and Jenq-Neng Hwang. Rethinking of radar's role: A camera-radar dataset and systematic annotator via coordinate alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2815–2824, June 2021.

[Wojke and Bewley, 2018] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 748–756. IEEE, 2018.

[Wojke *et al.*, 2017] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[Wu *et al.*, 2022] Zhi Wu, Dongheng Zhang, Chunyang Xie, Cong Yu, Jinbo Chen, Yang Hu, and Yan Chen. Rfmask: A simple baseline for human silhouette segmentation with radio signals. *IEEE Transactions on Multimedia*, 2022.

[Xue *et al.*, 2021] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. mmmesh: Towards 3d real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 269–282, 2021.

[Zhang *et al.*, 2018] Dongheng Zhang, Ying He, Xinyu Gong, Yang Hu, Yan Chen, and Bing Zeng. Multitarget aoa estimation using wideband lfmcw signal and two receiver antennas. *IEEE Transactions on Vehicular Technology*, 67(8):7101–7112, 2018.

[Zhang *et al.*, 2019a] Dongheng Zhang, Yang Hu, Yan Chen, and Bing Zeng. Breathtrack: Tracking indoor human breath status via commodity wifi. *IEEE Internet of Things Journal*, 6(2):3899–3911, 2019.

[Zhang *et al.*, 2019b] Dongheng Zhang, Yang Hu, Yan Chen, and Bing Zeng. Calibrating phase offsets for commodity wifi. *IEEE Systems Journal*, 14(1):661–664, 2019.

[Zhang *et al.*, 2020a] Dongheng Zhang, Yang Hu, and Yan Chen. Mtrack: Tracking multiperson moving trajectories and vital signs with radio signals. *IEEE Internet of Things Journal*, 8(5):3904–3914, 2020.

[Zhang *et al.*, 2020b] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *arXiv preprint arXiv:2004.01888*, 2020.

[Zhang *et al.*, 2021] D. Zhang, Y. Hu, and Y. Chen. Mtrack: Tracking multiperson moving trajectories and vital signs with radio signals. *IEEE Internet of Things Journal*, 8(5):3904–3914, 2021.

[Zhang, 2000] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

[Zhou and Wang, 2019] Gongjian Zhou and Liangliang Wang. Pseudo-spectrum based speed square filter for track-before-detect in range-doppler domain. *IEEE Transactions on Signal Processing*, 67(21):5596–5610, 2019.

[Zhou *et al.*, 2019] Yi Zhou, Tian Wang, Ronghua Hu, Hang Su, Yi Liu, Xiaoming Liu, Jidong Suo, and Hichem Snoussi. Multiple kernelized correlation filters (mkcf) for extended object tracking using x-band marine radar data. *IEEE Transactions on Signal Processing*, 67(14):3676–3688, 2019.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *ArXiv*, abs/2010.04159, 2020.