IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 December 2023, accepted 4 January 2024, date of publication 10 January 2024, date of current version 15 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3352022

STANDARDS

Causal Localization Network for Radar Human Localization With Micro-Doppler Signature

SUNJAE YOON, (Member, IEEE), GWANHYEONG KOO, JUN YEOP SHIM, SOOHWAN EOM, JI WOO HONG, (Member, IEEE), AND CHANG D. YOO, (Senior Member, IEEE)

School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea

Corresponding author: Chang D. Yoo (cd_yoo@kaist.ac.kr)

This work was supported in part by the Center for Applied Research in Artificial Intelligence (CARAI) funded by Defense Acquisition Program Administration (DAPA) and Agency for Defense Development (ADD) under Grant UD230017TD; in part by the Institute for Information & Communications Technology Promotion (IITP) funded by the Korean Government [Ministry of Science and ICT (MSIT)] through the Development of Causal Artificial Intelligence (AI) through Video Understanding and Reinforcement Learning, and Its Applications to Real Environments under Grant 2021-0-01381; and in part by the Brain Korea 21 FOUR (BK21 FOUR) (Connected AI Education & Research Program for Industry and Society Innovation), Korea Advanced Institute of Science and Technology Electrical Engineering (KAIST EE), under Grant 4120200113769.

ABSTRACT The Micro-Doppler (MD) signature includes unique characteristics from different-sized body parts such as arms, legs, and torso. Existing radar identification systems have attempted to classify human identification using these characteristics in MD signatures, achieving remarkable classification performance. However, we argue that radar identification systems should also be extended to perform more fine-grained tasks for greater identification flexibility. In this paper, we introduce the radar human localization (RHL) task, which involves temporally localizing human identifications within untrimmed MD signatures. To facilitate RHL, we have constructed a micro-Doppler dataset named IDRad-TBA. Additionally, we propose the Causal Localization Network (CLNet) as the baseline system for RHL, built on the IDRad-TBA dataset. CLNet utilizes a novel temporal causal prediction approach for MD signature localization. Experimental results demonstrate CLNet's effectiveness in executing the RHL task. Our project is available at: https://github.com/dbstjswo505/CLNet.

INDEX TERMS Deep learning, temporal human identification, micro-Doppler radar, information retrieval.

I. INTRODUCTION

Radar human identification (RHI) [4], [5] underpins many personal identification systems regarding security, surveillance, and other personalized services. While visual information has been a popular choice for human identification due to the distinctive features of human external appearance, it faces challenges in operating effectively under low light conditions and raises privacy concerns. In contrast, radar devices offer a viable alternative by circumventing these difficulties. Radars emit electromagnetic waves toward the target and measure a target's physical properties (e.g., distance, speed, angle) based on the waves reflected back from the target. Radars can operate under low light conditions over long distances and their ability to bend around obstacles makes them suitable for identification in obscured environments [1], [2], [3]. Moreover, radar offers a significant advantage regarding privacy concerns since the information obtained through radar is inherently difficult for individuals to interpret directly. Therefore, radar emerges as a robust sensor for human identification compared to image or video sensors.

To identify humans in radar, recent RHI systems [4], [5] have incorporated temporally recorded radar signatures, where they utilize gait as a biometric of a human in temporally recorded Doppler radar signatures. The gait can be observed from a distance and holds unique patterns of behavior made by humans' different-sized body parts. Figure 1 (a) shows micro-Doppler (MD) radar signatures [4] that contain gait patterns made by human walking. These MD signatures record variations of frequencies in moving objects, such that they effectively represent distinctive characteristics manifested by different-sized body parts during the gait, including rapid movements of the arms and legs, as well as slower movements of the torso. Leveraging these distinct

The associate editor coordinating the review of this manuscript and approving it for publication was Fu Lee Wang.

characteristics, numerous radar identification systems [4], [5] have been proposed, resulting in outstanding identification performance, as shown in Figure 1(b). Based on the advanced identification capabilities of RHI systems, this paper first proposes a new challenge of extending human identifications to more fine-grained identification as radar human localization (RHL) in MD signatures.

Our proposed radar human localization aims to localize temporal regions of a targeted individual within untrimmed micro-Doppler radar signatures, assuming that only one object (i.e., target) exists at each time. To this end, we build a radar dataset referred to as IDRad-TBA (IDRad-Temporal Boundary Annotation), which provides temporal boundary annotations of human identities on sequential MD signatures based on the IDRad [4] dataset. The IDRad dataset captures MD signatures of a person walking in 2-dimensional data according to the 256 Doppler channels and about 45 frames. Thus, a single frame contains 256 data from the Doppler channel. Since the IDRad dataset is primarily designed for the RHI task, all individual MD signatures come with annotations representing single human identities. To perform RHL, as depicted in Figure 1(c), we synthesize new MD signatures by combining multiple different-sized MD signatures of the IDRad dataset. On our synthesized MD signatures, we annotate all their identities and their temporal boundaries in the MD signatures. Therefore, as shown in Figure 2, our designed RHL task takes MD signatures and the target identity of humans as input and produces outputs of temporal boundaries (*i.e.*, start-time and end-time) pertinent to the target identity within the MD signatures.

To this end, we build a baseline system to perform the RHL task, referred to as Causal Localization Network (CLNet). CLNet is composed of three modules: (1) Patch-wise Doppler Encoding (PDE) which encodes Doppler signals into d-dimensional features by considering human gait patterns contained in Doppler patches, (2) Targeted-guided Doppler Encoding (TDE) which encodes the Doppler feature by focusing on the information related to the input target using self-attention in Transformer [21], and (3) Causal Localization Head (CLH) which predicts temporal boundaries in MD signatures pertinent to the input target under our designed causal prediction approach. As CLNet is the first work designed for the proposed RHL task, we have also made modifications to previous models used in the RHI task to incorporate localization capabilities. Experimental studies have been conducted to validate the efficiency and effectiveness of CLNet, demonstrating the potential for human localization within the MD signatures.

The contributions of this paper can be summarized as three-fold: (1) We propose a radar human localization (RHL) task to localize the temporal region of a targeted human in micro-Doppler signatures, which has never been attempted before, (2) We present synthesized micro-Doppler signatures referred to as IDRad-TBA as an experimental contribution to perform RHL task, and (3) We build baseline model referred to as Causal Localization Network (CLNet) to perform RHL task.

II. RELATED WORKS

A. MICRO-DOPPLER SIGNATURES

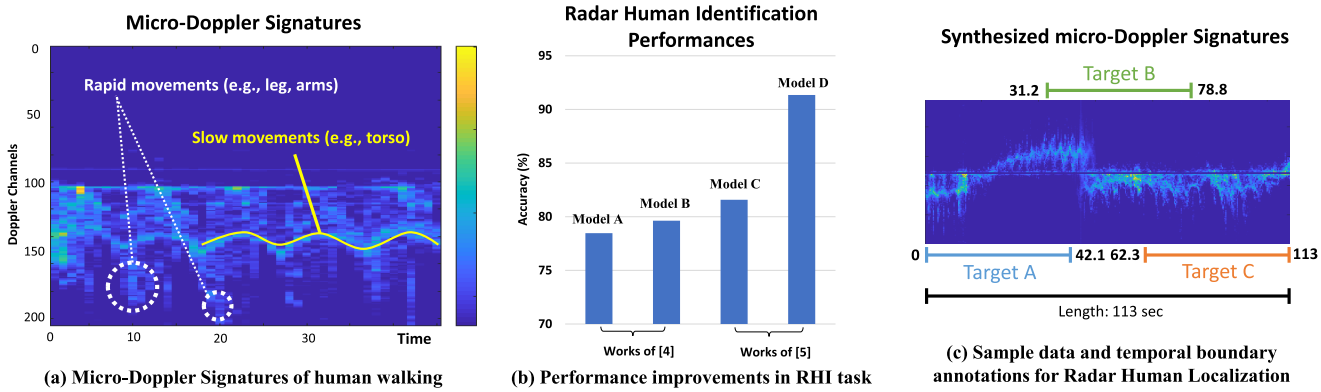
Micro-Doppler (MD) signature refers to the small variations in Doppler frequency caused by the motion of individual parts or features within a larger object (*i.e.*, the term ''micro'' indicates that we are looking at very small-scale movements or features within an object.). To obtain MD signatures, we first build a range-Doppler map that analyzes the Doppler frequency within a specific distance using a two-dimensional Fourier transform. The absolute values of the signals in the range-Doppler map are then integrated along the range dimension, which makes the frequency variations to the direction of the radar. The recording of these variations over time builds MD signatures. Therefore MD signature measures the change in frequency of single-tone radio waves reflected off a moving object to determine its velocity. In addition to velocity, the MD signature also considers the Doppler effect, which refers to the modulation of the radar signal caused by the motion of object components or features. In detail, by the Doppler effect, the observed frequency f from a moving object is shifted away from the emitted frequency f_0 such that the Doppler frequency $\Delta f = f - f_0$ is defined by subtracting f from f_0 as given below:

$$f = f_0 \times \frac{c \pm v_r}{c \pm v_s}, \tag{1}$$

where c is the velocity of waves in the medium and v_r is the velocity of the receiver relative to the medium, where it is positive if the receiver is moving towards the source and negative if in the opposite direction. v_s is the velocity of the source relative to the medium, where it is positive if the source is moving away from the receiver and negative if in the opposite direction. By analyzing the Doppler shift [6] in a more fine-grained scale (*e.g.*, micro-scale) over multi-components in an object, it is able to detect and recognize its motions [7], [8]. In this work, following the work in [4], our system is operated on 77GHz Frequency-Modulated Continuous-Wave (FMCW), which is a radar that employs a continuous transmission of a radio wave with a linearly increasing or decreasing frequency over time. The FMCW is commonly used in radar systems for various applications, including distance measurement, velocity estimation, and target tracking. The recent computational power of deep learning bridges these MD signals to many various studies, and we elaborate on this in the following section.

B. DEEP LEARNING FOR MICRO-DOPPLER SIGNATURES

As the demand for radar-based systems continues to grow, the exploration of deep learning techniques has led to the emergence of several applications [10], [11], [12], [13] leveraging micro-Doppler signatures. One notable application of deep learning with micro-Doppler signatures is in motion recognition and classification. By training deep neural networks on large datasets of micro-Doppler signatures [4], [9] containing different motions on various objects



(a) Micro-Doppler Signatures of human walking (b) Performance improvements in RHI task (c) Sample data and temporal boundary annotations for Radar Human Localization

FIGURE 1. Illustrations of micro-Doppler signatures: (a) micro-Doppler signatures of human walking. When a person walks, the arms and legs generate rapid movements (white circle) and the torso generates slow movements (yellow curve) in the signatures. (b) Previous performances of Radar Human Identification (RHI) task. (Model A: classifier with CNN on MD signatures of input 45 frames x 205 channels, Model B: classifier with CNN on MD signatures of 150 × 205 size, Model C: classifier with LSTM on MD signatures of 150 × 205 size, and Model D: classifier with attention on MD signatures of 150 × 205 size.) (c) Sample of synthesized micro-Doppler signatures for Radar Human Localization (RHL).

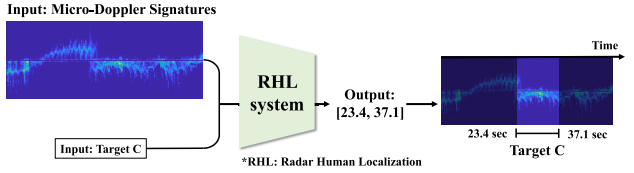


FIGURE 2. Illustration of Radar Human Localization (RHL) system. The RHL system takes inputs of micro-Doppler signatures and target information, where it predicts temporal boundaries in the signatures related to the input target.

(*e.g.*, vehicles, aircraft, or human movements), the networks are able to distinguish various objects or activities based solely on their micro-Doppler signatures. To be specific, Kim et al. [15] first employed a neural network on MD signatures for motion recognition of humans, demonstrating the potential of deep learning in radar signal analysis. Following the work, researchers have introduced various deep learning techniques on the MD signatures, including transfer learning [16] and learning-based algorithms [17]. Lin et al. [17] devise an iterative convolutional neural network (CNN) under random forests algorithm to operate on MD signatures which repetitively enhance the representations of MD signatures. Park et al. [16] introduce a deep CNN model pre-trained on ImageNet [14], which is a large-scale visual image classification dataset, to contribute the knowledge transfer about pattern recognition from image to the MD signatures. Furthermore, the deep learning approach is also applied for monitoring fine-grained human body parts [4], [18], [20]. By analyzing the micro-Doppler patterns induced by human body movements, deep learning models can understand vital sign information such as heart rate, respiration rate, and gait pattern. The non-contact monitoring approach has the potential to revolutionize healthcare and security applications. Recently, to mitigate the expensive and time-consuming process of obtaining MD signatures, augmentation [28] has been applied to raw MD signatures, where a learning-based approach [29] is also designed using the generative adversarial network [30]. Especially, MD signatures on human gait characteristics [4] are used for radar human identification (RHI), where the system processes the information of movements from fine-grained human body parts in an uncontrolled scenario where a target is allowed to walk around in a free and spontaneous way. Cao et al. [18] first applied deep CNNs on MD signatures induced by limb movements and torso motion for HI. Vandersmissen et al. [4] also used the deep CNN and released a public dataset IDRad for RHI, which contributed to subsequent research. Yoon et al. [5] applied an attention method among fast and slow movements of human body parts to enhance representations of the distinguished walking patterns. Currently, there have been many advancements in RHI. On top of these works, to enhance the sensibility of sequential radar sequences, this paper first proposes a new challenge of extending human identifications to more fine-grained identification as radar human localization (RHL) in MD signatures.

III. DATASET

Our IDRad-TBA is built on micro-Doppler signatures of IDRad dataset [4], which recorded human (*i.e.*, 5 people) free-walking in a room for 100 minutes (*i.e.*, 20 minutes per person). We synthesized the MD signatures for the radar human localization task. Therefore, a total of 5000 synthesized MD signatures are constructed, where we provide (12078/1463/1479) annotations (*i.e.*, start-end time) for (train/val/test) on top of the synthesized MD signatures. The detailed process of synthesizing MD signatures is presented in the following.

A. DATA SYNTHESIZING

Figure 3 provides a schematic process of synthesizing MD signatures from the MD signature corpus (*i.e.*, IDRad dataset). The MD signature corpus includes five types of MD

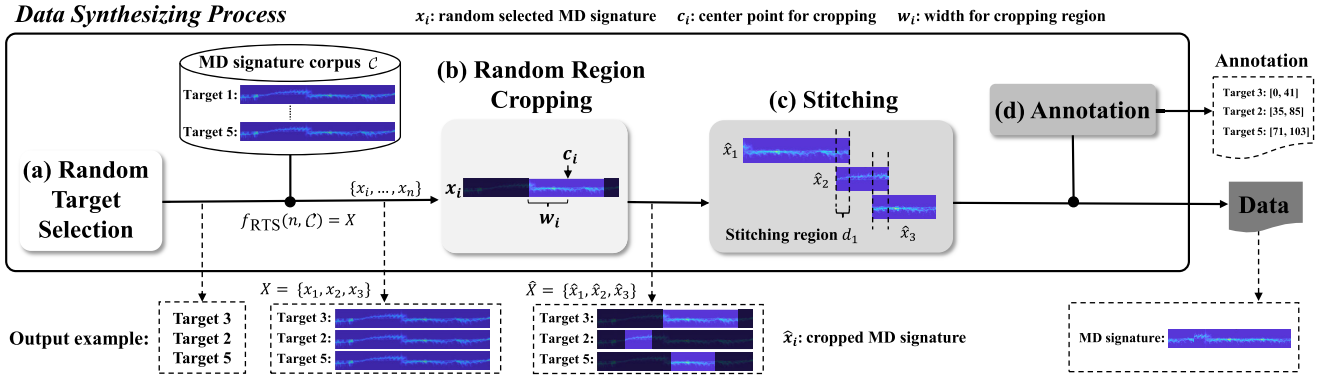


FIGURE 3. Pipeline of synthesizing MD signatures for RHL: (a) Random Target Selection: It randomly selects the targets to incorporate their MD signatures. (b) Random Region Cropping: It randomly crops the selected MD signatures to stitch together. (c) Stitching: It stitches all the cropped MD signatures to make a synthesized MD signature. (d) Annotation: It annotates all the boundary information contained in the synthesized MD signatures.

signatures from five different people (*i.e.*, target), where each MD signature is a continuous recording of 20 minutes. For preparing the dataset to perform the RHL task, we summarize the data synthesizing process into four steps: (a) Random Target Selection, (b) Random Region Cropping, (c) Stitching, and (d) Annotation. The output example of each step is also illustrated below each step process in Figure 3.

1) RANDOM TARGET SELECTION

We assume that all MD signatures include a single human walking signal.[1] Under this assumption, the random target selection (RTS) determines how many targets are included in one sample. For a total of 5 categorical types in the MD signature corpus, the RTS selects $1 \leq n \leq 5$ targets excluding redundant selection and provides MD signatures corresponding to the selections. Formally we define the RTS as below:

$$X = f_{\text{RTS}}(n, \mathcal{C}), \tag{2}$$

where the f_{RTS} is the random target selection, n is the randomly decided number of targets to select, $\mathcal{C}$ is MD signature corpus. $X = \{x_i, \cdots, x_n\}$ is the set of selected n different MD signatures. For instance, RTS randomly selects 3 different targets as $\{x_1, x_2, x_3\} = \{\text{target 3}, \text{target 2}, \text{target 5}\}$ and provides the MD signatures about the targets.

2) RANDOM REGION CROPPING

For the selected MD signatures, we assign a region for each MD signature to crop and stitch all the cropped signatures to build synthesized MD signatures. Here, we first provide the details of cropping. For each 256-channel MD signature x_i with the frame length of N_i (*i.e.*, $x_i \in \mathbb{R}^{N_i \times 256}$) in the selected set $X = \{x_i, \cdots, x_n\}$, we randomly specify the region to crop, which can be formulated as given below:

$$\hat{x}_i = x_i[c_i - w_i : c_i + w_i], \tag{3}$$

[1]We share the same assumption about MD signatures in the previous work [4] designed for radar human identification task.

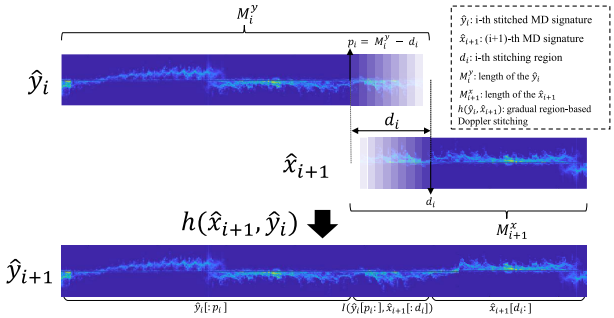


FIGURE 4. Illustration of the gradual region-based stitching process.

where $c_i = \text{Random}(w_{max}, N_i - w_{max})$ and $w_i = \text{Random}(w_{min}, w_{max})$ are random scalar indicating the center point and cropping width respectively to specify cropping region in x_i as shown in Figure 3. Random(x, y) denotes the random sampling of integer value between x and y along the Gaussian sampling distribution. $w_{min} = 40$, $w_{max} = 200$ are hyperparameters for the width and N_i is the length of the MD signatures. This makes the cropped set of MD signatures as $\hat{X} = \{\hat{x}_1, \cdots, \hat{x}_n\}$, which will be integrated in the following stitching process.

3) STITCHING

We stitch the cropped MD signatures among adjacent MD signatures in the $\hat{X}$ as shown in Figure 3 (c). If the MD signatures are simply attached together by end-to-end stitching, the seam exists in the merged image, which gives an undesirable shortcut for the network to memorize the seam. Therefore as shown in Figure 4, we apply gradual region-based stitching between the two MD signatures $\hat{y}_i$ and $\hat{x}_{i+1}$ to make $\hat{y}_{i+1} = h(x_{i+1}, y_i)$, where the $\hat{y}_i$ is the previously stitched image up to i-th MD signature. The $h(x, y)$ is our proposed gradual region-based Doppler stitching that stitches Doppler signatures x on top of the signatures y. In the process of $h(x_{i+1}, y_i)$, the stitching regions between x_{i+1} and y_i are

decided as d_i randomly between 5 and 10 frames. To be specific, the initial condition is $\hat{y}_1 = \hat{x}_1$ and the values in the stitching region are updated by an interpolation between $\hat{y}_i[M_i^y - d_i :]$ and $\hat{x}_{i+1}[: d_i]$ as given below:

$$I(\hat{y}_i[p_i :], \hat{x}_{i+1}[: d_i]) = \bigcup_{n=0}^{d_i} \{\delta \times \hat{y}_i[p_i + n] + (1 - \delta) \times \hat{x}_{i+1}[n]\}, \quad (4)$$

where $p_i = M_i^y - d_i$ is the starting point of stitching along the frame axis of $\hat{y}_i$. M_i^y is the length of the $\hat{y}_i$ and d_i is the length of stitching region. $I(\cdot, \cdot)$ is a frame-wise interpolation between $\hat{y}_i$ and $\hat{x}_{i+1}$ by scaling with $\delta = n/d_i$. The final stitching output $\hat{y}_{i+1} \in \mathbb{R}^{(M_i^y + M_{i+1}^x - d_i) \times 256}$ is concatenated[2] from input $\hat{y}_i$, $\hat{x}_{i+1}$ and the final stitched MD signatures $y_{i+1} = h(\hat{x}_{i+1}, \hat{y}_i)$ as formulary given below:

$$\hat{y}_{i+1} = \text{cat}(\hat{y}_i[: p_i], I(\hat{y}_i[p_i :], \hat{x}_{i+1}[: d_i]), \hat{x}_{i+1}[d_i :]), \quad (5)$$

where the $\text{cat}(\cdot, \cdot, \cdot)$ is the concatenation operation along the frame axis. After a gradual stitching process by the h, a single synthesized MD signature $\hat{y}_n$ is obtained, where we annotate the boundary information contained in each target.

4) ANNOTATION

To annotate the temporal boundary information in the synthesized MD signature $\hat{y}_n$, we utilize the starting point p_i of the stitching boundary of each step i. To be specific, if the n MD signatures are involved to synthesize $\hat{y}_n$, n temporal boundaries are available to annotate with their target information. Thus, for the synthesized signature $\hat{y}_n$, we provide the n annotations as $\{[p_0, p_1 + d_1], \cdots, [p_{i-1}, p_i + d_i], \cdots, [p_{n-1}, M_n^y]\}$ satisfying $p_0 = 0$, M_n^y is the length of the $\hat{y}_n$, and d_i is the inclusion of overlapping region by stitching. We also add target information t_i contained in each boundary. Therefore, to perform RHL task, a single annotation in the synthesized MD signature sample is constructed as given:

$$\{\text{MD} : \hat{y}_n, \text{Target} : t_i, \text{Boundary} : [p_{i-1}, p_i + d_i]\}, \quad (6)$$

where the inputs are $\hat{y}_n$ and t_i, and the ground-truth is $[p_{i-1}, p_i + d_i]$ for RHL systems.

B. STATISTICAL ANALYSIS

To ensure our proposed IDRad-TBA not to be biased, we investigate the statistics about the start-end times of each annotation and their distributions of the lengths. The lengths of each target are evenly distributed in the MD signatures, resulting in similar average lengths and standard deviations as shown in Figure 5 (a). Moreover, Figure 5 (b) shows the 2 dimensional histograms in terms of start times and end times of all the temporal boundary annotations, where the vertical axis denotes the start time of the temporal boundary and the horizontal axis denotes the end time. The histogram

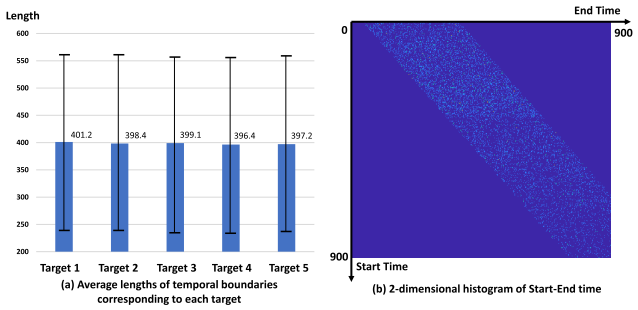


FIGURE 5. Statistical analysis on temporal boundary annotations for each target: (a) Average length of temporal boundaries corresponding to each target, (b) 2-dimensional histogram of start-end time.

shows that the annotations are evenly distributed across all the temporal regions,[3] which mitigates a bias that localizes specific temporal regions in the MD signatures.

IV. CAUSAL LOCALIZATION NETWORK

To perform the radar human localization (RHL) task, we present the simple baseline referred to as Causal Localization Network (CLNet) as shown in Figure 6. CLNet takes an MD signature and a target as input and predicts temporal boundaries pertinent to the input target. The proposed CLNet is composed of three modules: (1) Patch-wise Doppler Encoding (PDE) which embeds the MD signatures into d-dimensional Doppler features containing gait patterns by the patch-wise convolutional neural network, (2) Target-guided Doppler Encoding (TDE) which attends on the Doppler feature conditioned on the input target information, and (3) Causal Localization Head (CLH) which predicts start-end time related to the input target under our designed causal predicting approach. The details of CLNet are given in the following.

A. PREPROCESSING AND INPUT REPRESENTATION

1) PREPROCESSING

Some channels in MD signature contain information unrelated to human behavior. Therefore, we first remove these distinguished unnecessary channels in the preprocessing step. Typically, the wave frequency reflected from static objects corresponds to this unrelated information. To be specific, the y-axis of MD signatures used in this work represents the speeds ranging from -3.8 m/s to 3.8 m/s and the static objects show the velocities in the proximity of $[-0.03$ $m/s, 0.03$ $m/s]$. Thus, as a preprocessing, we remove their corresponding channels (*i.e.*, regions in 127-129 channels of the y-axis) as a biased channel due to the static object in the space. Furthermore, the top 24 channels (*i.e.*, 233-256 channels) and bottom 24 channels (*i.e.*, 1-24 channels) are also removed as they are high-speed regions that humans can not make. After the aforementioned

[2] M_{i+1}^x is the length of the $\hat{x}_{i+1}$ and M_i^y is the length of the $\hat{y}_i$. 256 is the channel for the MD signatures.

[3] Since we specified the minimum and maximum lengths (*i.e.*, minimum: 40 frames, maximum: 200 frames) of the temporal boundary, it shows that annotations do not exist in areas other than the specified regions.

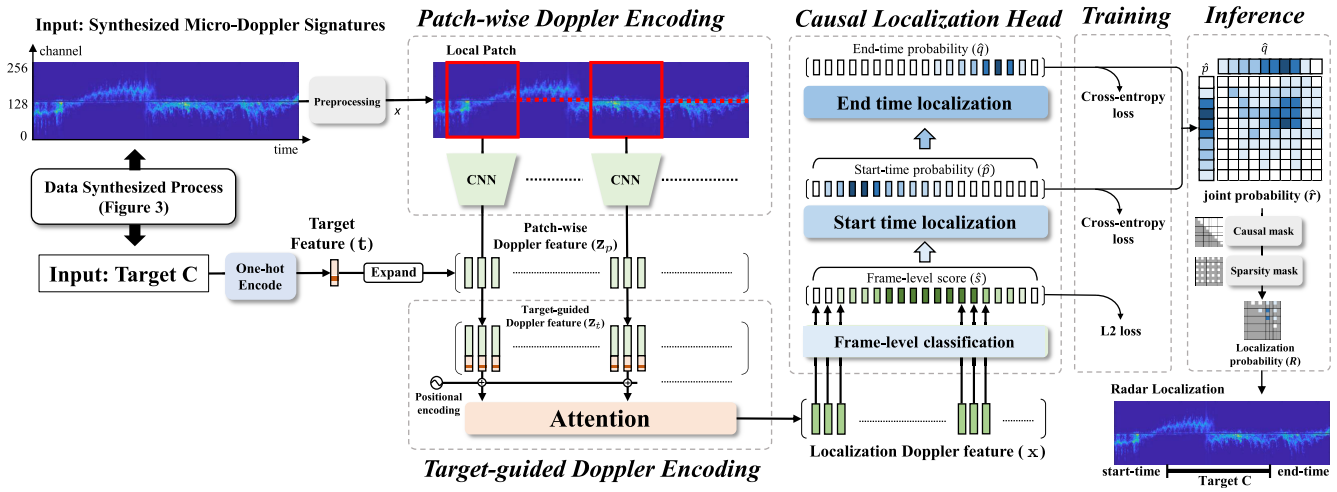


FIGURE 6. Illustrations of causal localization network (CLNet). CLNet is composed of three modules (a) Patch-wise doppler encoding which embeds MD signatures into d-dimensional features that consider human gait patterns contained in Doppler patches, (b) Target-guided Doppler Encoding which encodes the commonalities about temporal patterns pertinent to the target via temporal self-attention, (c) Causal Localization Head which predicts the temporal boundaries under our designed causal prediction approach.

preprocessing, our final MD signatures provide the signals with 205 channels.

2) INPUT REPRESENTATIONS

To give formal definitions of input representations (*i.e.*, MD signatures and target), we denote the $\mathbf{z} \in \mathbb{R}^{L\times 205}$ as MD signatures, where L is the frame length of MD signatures. The target is converted into one-hot encoding as $\mathbf{t} \in \mathbb{R}^{c}$, where c is the number of target classes.

B. PATCH-WISE DOPPLER ENCODING

The distinct movements of the arms and legs in human gait provide unique characteristics that can identify each individual. To this end, our designed Patch-wise Doppler Encoding (PDE) aims to extract features about unique characteristics of human gaits, such that PDE first prepares a sliding window in the MD signatures that can cover the human gait patterns, and extract the features using convolution neural network on the window. In detail, the width and height of the window[4] are 45 and 205, which slides along the time axis with a stride of 1. The mathematical formulation is represented as given below:

$$\mathbf{z}_p = \bigcup_{i=1}^{L} f_w(\mathbf{z}[i : i + w]), \tag{7}$$

where the $\mathbf{z}_p \in \mathbb{R}^{L\times d}$ is the patch-wise Doppler feature and $f_w : \mathbb{R}^{w\times 205} \rightarrow \mathbb{R}^{d}$ is multi-layer perception (MLP) composed of several convolution neural networks and max poolings, which maps patch-wise MD signatures into d-dimensional features. Here, the $w = 45$ is the width of the patch. In the following, the $\mathbf{z}_p$ is used for the input of the following Target-guided Doppler Encoding.

[4]See also ablation studies of this.

C. TARGET-GUIDED DOPPLER ENCODING

Target-guided Doppler Encoding (TDE) attends the features related to the input target $\mathbf{t}$ in the patch-wise Doppler features $\mathbf{z}_p$. To selectively attend information regarding the input target in the MD signature, we first provide the one-hot encoding of the target $\mathbf{t} \in \mathbb{R}^{c}$ to the Doppler features $\mathbf{z}_p$ by concatenating them as $\mathbf{z}_t = [\mathbf{z}_p||\mathbf{t}_p]W$, where $\mathbf{t}_p \in \mathbb{R}^{L\times c}$ is the expanded one-hot encoding of the target along the frame axis to concatenate (*i.e.*, $[\cdot||\cdot]$) all the features of $\mathbf{z}_p$ and $W \in \mathbb{R}^{(d+c)\times d}$ is d-dimensional embedder. Finally, $\mathbf{z}_t \in \mathbb{R}^{L\times d}$ is target-guided Doppler features. To highlight the common patterns in the features $\mathbf{z}_t$,[5] we introduce the attention method in Transformer [21], which is effective in similarity-based sequential information processing. To prepare the inputs of Transformer attention, we first update the $\mathbf{z}_t$ by applying layer normalization (LN) [22] and positional encoding (PE) [21], and then apply self-attention as given below:

$$\begin{aligned} \mathbf{z}_t &\leftarrow \mathrm{LN}(\mathbf{z}_t + \mathrm{PE}(\mathbf{z}_t)), \\ \mathbf{x} &= \mathrm{attention}(\mathbf{z}_t, \mathbf{z}_t, \mathbf{z}_t), \end{aligned} \tag{8}$$

where $\mathbf{x} \in \mathbb{R}^{L\times d}$ is the final localization Doppler features to localize the start-time and end-time pertinent to the input target $\mathbf{t}$. The $\mathbf{x}$ is used as input in the following Causal Localization Head.

D. CAUSAL LOCALIZATION HEAD

To train a model for temporal localization (*i.e.*, start-end times) in MD signatures, we propose a causal localization head (CLH), which performs three predictions: (1) frame-level prediction, (2) start-time prediction, and (3) end-time prediction. To perform the predictions, we design a causal process in Figure 7 to sequentially predict the start-end time

[5]Since human walking is repeated in a certain pattern, we also aim to obtain this patterned information from the Doppler features.

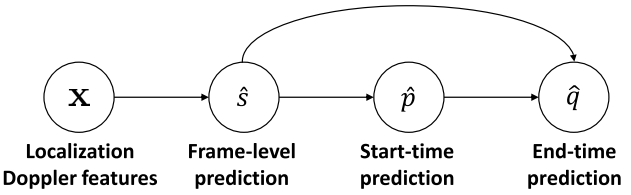


FIGURE 7. Illustration of the causal prediction process in Causal Localization Head. (x: Localization Doppler features, $\hat{s}$: Frame-level prediction, $\hat{p}$: Start-time prediction, $\hat{q}$: End-time prediction).

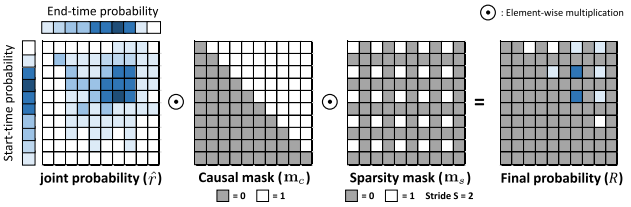


FIGURE 8. Illustration of causal masking and sparsity masking on joint probability for target localization in MD signatures.

of the input target. We elaborate on these predictions with their training objectives in the following paragraphs.

1) FRAME-LEVEL PREDICTION

The frame-level prediction provides discrete score distributions $\hat{s} = (\hat{s}[1], \cdots, \hat{s}[L])$ over L frames from input localization Doppler features $\mathbf{x} \in \mathbb{R}^{L\times d}$, where a single frame output $s[i]$ is the score between 0 and 1, denoting 1 is the signal of the input target and 0 is not. The $\hat{s} \in \mathbb{R}^{L\times 1}$ is obtained from $\mathbf{x}$ as given below:

$$\hat{s} = \sigma(\mathbf{x}W_1). \tag{9}$$

where $W_1 \in \mathbb{R}^{d\times 1}$ is the score embedding and $\sigma(\cdot)$ is the sigmoid function to scale the output between 0 and 1. To optimize the score $\hat{s}$ to be closed to 1 in the region of ground-truth temporal boundary, we prepare the score label $s = [0, \ldots, 1, 1, 1, \ldots, 0] \in \mathbb{R}^{L\times 1}$, where the s has a scalar value 1 in the region between ground-truth start-time and end-time and the other regions have 0 values. The frame-level loss $\mathcal{L}_f$ is defined using L2 loss with the label as $\mathcal{L}_f = \sum_{i=1}^{L} ||s - \hat{s}||_2^2$.

2) START-END TIME PREDICTION

The start-time prediction provides discrete probability distributions $\hat{p} = (\hat{p}[1], \cdots, \hat{p}[L]) \in \mathbb{R}^{L\times 1}$ over L frames denoting the probabilities of start-time of the target in the MD signatures. Based on the frame-level score distributions $\hat{s}$, we estimate the start-time of the target. Thus, 1D convolution filter $\text{Conv1D}_{\text{st}}$ is applied on top of $\hat{s}$ as given below:

$$\hat{p} = \text{softmax}(\text{Conv1D}_{\text{st}}(\hat{s})), \tag{10}$$

The softmax function is used to produce the probability distributions along the frame axis. The end-time prediction also provides discrete probability distributions as $\hat{q} = (\hat{q}[1], \cdots, \hat{q}[L]) \in \mathbb{R}^{L\times 1}$. The end-time is also predicted based on the two frame-level predictions and start-time predictions. However, the $\hat{q}$ is based on the concatenated two distributions about start-time $\hat{p}$ and the frame-level $\hat{s}$ as below:

$$\hat{q} = \text{softmax}(\text{Conv1D}_{\text{ed}}([\hat{s}||\hat{p}]W_2)), \tag{11}$$

where $\text{Conv1D}_{\text{ed}}$ is 1D convolution filter for end-time probability and $W_2 \in \mathbb{R}^{2\times 1}$ is learnable matrix to give frame-level weighted summation between $\hat{s}$ and $\hat{p}$. The training objectives of the $\hat{p}$ and $\hat{q}$ are locational loss as $\mathcal{L}_{loc} = -\log\hat{p}[i_{st}] - \log\hat{q}[i_{ed}]$ using the cross-entropy loss, where i_{st} and i_{ed} are ground-truth indices of frame axis in terms of start-time and end-time. The total loss is the summation of frame-level and locational losses as $\mathcal{L} = \mathcal{L}_f + \mathcal{L}_{loc}$. In the inference, the two probabilities $\hat{p}$ and $\hat{q}$ are utilized by building joint probabilities, which are explained in the following section.

E. INFERENCE

In an inference, following the work [19], CLNet considers the start-time and end-time distributions (*i.e.*, $\hat{p}, \hat{q}$) via building localization joint probability distributions $\hat{r}$ in Figure 6. The $\hat{r}$ is obtained by applying matrix multiplication between $\hat{p} \in \mathbb{R}^{L\times 1}$ and $\hat{q} \in \mathbb{R}^{L\times 1}$ as $\hat{r} = \hat{p}\hat{q}^T \in \mathbb{R}^{L\times L}$, where $\hat{r}_{i,j} = \hat{p}_i\hat{q}_j$ denotes the joint probability that the inference about temporal boundary information would start at index i and end at index j along the frame axis. Based on the distribution $\hat{r}$, as shown in Figure 8, we apply the causal mask $\mathbf{m}^c \in \mathbb{R}^{L\times L}$ to ensure the cases that the end-time precedes the start-time should have zero probability. Thus the values of lower triangular regions in $\mathbf{m}^c$ are zero and the values of upper triangular regions are one to keep the original probabilities. To provide the sparsity among the candidate probabilities in $\hat{r}$, we also apply the sparsity mask $\mathbf{m}^s \in \mathbb{R}^{L\times L}$, where the probabilities to have sparsity with distance S(*e.g.*, $S = 2$). The mathematical formulations about $\mathbf{m}^c$ and $\mathbf{m}^s$ are as given below:

$$\mathbf{m}^c_{i,j} = \begin{cases} 1 & i < j, \\ 0 & i \geq j \end{cases},$$
$$\mathbf{m}^s_{i,j} = \begin{cases} 1 & \textit{if } i \textit{ mod } S = 0 \text{ and } j \textit{ mod } S = 0, \\ 0 & \textit{else} \end{cases}, \tag{12}$$

where *mod* is the modulo operation. The index starts from zero and the final probability $R = \hat{r} \odot \mathbf{m}^c \odot \mathbf{m}^s$ is defined by applying the two masks, where $\odot$ is element-wise multiplication. The final start-end times for RHL task are predicted on the highest probabilities of R.

V. EXPERIMENTS

We validate the baseline CLNet on our proposed IDRad-TBA dataset. The details are explained in the following.

A. DATASET

IDRad-TBA dataset is based on the MD signatures of human walking in IDRad dataset [4], where the temporal annotation is provided to perform the radar human localization task.

TABLE 1. Performance comparisons for radar human localization on IDRad-TBA (test). ⋆: reconstruction-based results with the public codes.

Method	IDRad-TBA												
	IoU=0.7				IoU=0.5				IoU=0.3				mIoU
	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50	R@1	R@5	R@10	R@50	R@1
deep CNN⋆ [4]	2.31	4.83	8.31	13.42	4.62	9.62	16.42	21.32	12.14	16.82	22.41	29.62	10.64
PCA SVM⋆ [5]	1.96	3.61	7.91	13.11	3.73	7.43	10.43	18.62	9.49	13.52	17.83	22.47	8.43
DSDA⋆ [5]	3.98	6.45	11.23	16.34	8.53	13.34	19.21	23.43	14.32	21.57	26.62	33.21	11.52
CLNet (Ours)	**13.20**	**24.79**	**31.37**	**45.09**	**23.21**	**39.18**	**45.09**	**61.76**	**41.99**	**57.18**	**63.71**	**79.94**	**21.35**

TABLE 2. Ablation study on model variants of CLNet on IDRad-TBA (validation). (PDE: Patch-wise Doppler Encoding, TDE: Target-guided Doppler Embedding, CLH: Causal Localization Head) ✓⋆: end-time prediction is used as prior knowledge and start-time is predicted.

PDE	TDE	CLH	IoU=0.7		
			R@1	R@10	R@50
✓			9.74	16.70	37.19
	✓		8.36	14.32	35.35
		✓	10.02	17.43	40.12
✓	✓		10.21	18.32	41.34
✓		✓	12.32	21.81	43.91
	✓	✓	11.15	19.32	41.32
✓	✓	✓⋆	**13.41**	22.93	45.52
✓	✓	✓	13.36	**23.18**	**45.63**

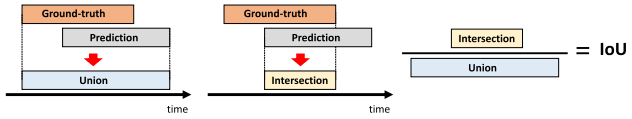


FIGURE 9. Illustration of calculating the intersection of union for evaluating radar human localization task.

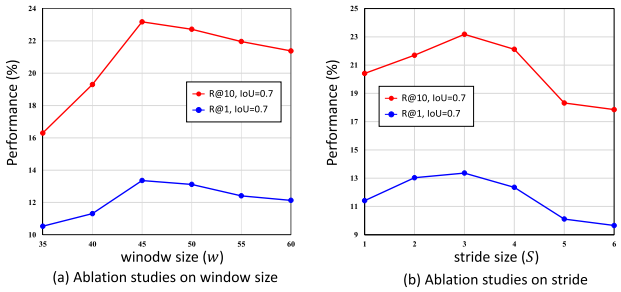


FIGURE 10. Experiments on hyperparameters about (a) patch window size (w) in PDE and (b) stride (S) for sparsity mask in CLH.

The MD signatures are synthesized to build untrimmed MD signatures containing multiple humans walking in a sequence, where all the temporal boundaries of each human are annotated as ground-truth information. Therefore IDRad-TBA provides 5K synthesized MD signatures about multiple human walking, where it is split into 80% train, 10% val, and 10% test. Each synthesized MD signatures are about 35 seconds and contain multiple annotations about start-end times corresponding to all the human in the signatures.

B. EXPERIMENTAL DETAILS

1) EVALUATION METRIC

Following the popular metrics [23], [24] of video localization, we perform recall metrics about the intersection of union between the predicted temporal boundary and the ground-truth. Shortly, it is referred to as "R@n,IoU=μ", where it denotes the percentage of targets having at least one prediction whose Intersection over Union (IoU) with ground truth is larger than μ in top-n localized temporal boundaries. The IoU calculation follows IoU = (intersection of predicted temporal region and ground truth temporal region)/(union of temporal region and ground truth temporal region), which is also illustrated in Figure 9. The number of IoU closer to 1 denotes that the model prediction is highly aligned with ground-truth. In our experiment, we validate the CLNet on the setting of $n = \{1, 5, 10, 50\}$ and $\mu = \{0.3, 0.5, 0.7\}$, which shows diverse performances of the system with several conditions. Moreover, we also measure "mIoU" which is the average IoU over all samples.

2) TRAINING DETAILS

The MD signatures are recorded based on 15 frames per second. All the values in MD signatures are normalized by the mean and the standard deviation. The 256 channels of MD signatures are truncated by the preprocessing in Section IV, resulting in 205 channels. The number of targets in IDRad-TBA is 5, such that the dimension of one-hot encoding in target features $\mathbf{t} \in \mathbb{R}^c$ is as $c = 5$. The model is trained for 150 epochs with a batch size of 48. Learnable parameters are optimized by Adam [25] with a learning rate of 0.001, while applying linear decay of learning rate. The stride for sparsity pooling is $S = 3$ along both vertical and horizontal axes.

C. EXPERIMENTAL STUDIES

1) QUANTITATIVE RESULTS

Table 1 summarizes the localization performances on the IDRad-TBA test split. As our work is the first work of radar human localization on the proposed IDRad-TBA dataset, we validate the models with many various metrics. Furthermore, there have been several previous works [4], [5] that perform radar human identification task, such that we also modify them to perform the localization task. To be specific, deep CNN [4] and DSDA [5] are the classifier to identify the target human based on the input of MD signatures about the gait of human walking. The input sizes of their models are (45×205) and (150×205), where the numbers 45 and 150 are the temporal lengths of MD signatures and the

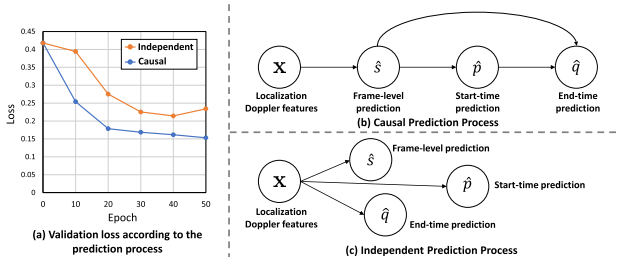


FIGURE 11. (a) Efficiency analysis on the saturation of validation loss corresponding to prediction approach, (b) Illustration of causal prediction process, (c) Illustration of the independent prediction process.

number 205 is the input channel dimension in each model. While keeping their input sizes, to perform the RHL task, these models predict synthesized MD signatures by sliding window approach. Under stride 1 of the sliding window approach, each model is available to predict frame-level target classification, where their predictions can be utilized for localization. Based on the frame-level predictions, we localize the temporal boundaries[6] of the target along the frame axis and measure their localization results with our evaluation metrics. Our baseline CLNet shows the highest performances among the candidate models for RHL task. Especially the performances of IoU=0.3 are highly improved in localizing the target. The performances of IoU=0.7 are still challenging to perform proper localization, which tells us that CLNet needs to be improved by fine-grained radar understanding.

2) ABLATION STUDY

Table 2 summarizes the ablative studies of the proposed three modules in CLNet: (1) Patch-wise Doppler Encoding (PDE), (2) Target-guided Doppler Encoding (TDE), (3) Causal Localization Head (CLH). It is confirmed that the localization performances of CLNet are incremental as each module is added. Sections III and IV of Table 2 are the ablative experiments about CLH module. When the CLH is not used in the model, CLNet simply predicts start-time and end-time independently, where the procedure of predictions is also illustrated in Figure 11 (c). Based on the comparisons of the results, we identify that the CLH contributes to the model with the most effectiveness. We consider that this is because the causal process of radar localization provides effective prior (*i.e.*, start-time) by narrowing the search space for finding end-time. It also shows similar effectiveness in taking end-time prediction as prior and predicting start-time. Furthermore, PDE is also effective for extracting information in the human gaits, which results in enhanced performances when applying the module.

Figure 10 summarizes the ablation studies on hyperparameters about CLNet: (1) window size of the patch-wise Doppler encoding and (2) stride of sparsity matrix in

[6]We clustered the predictions along the frame axis not to be sparse predictions and selected the longest temporal region to the input target.

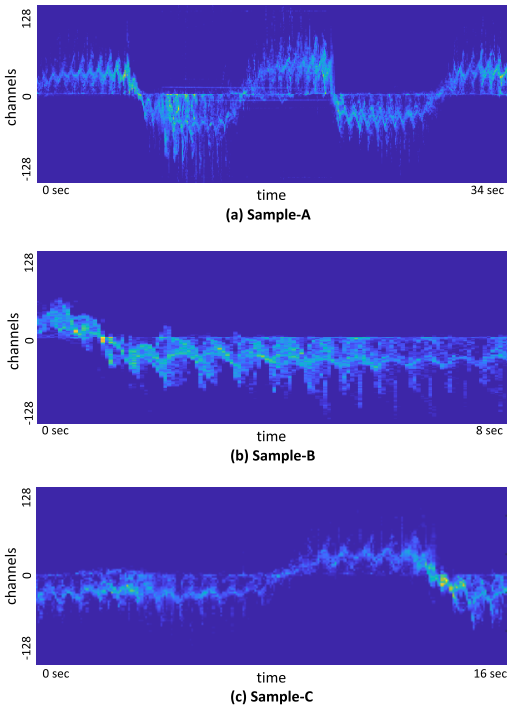


FIGURE 12. Samples of synthesized MD signatures: (a) annotations of sample-A: target 4 in (0s, 9.3s), target 2 in (8.2s, 20.9s), target 5 in (19.2s, 34.8s), (b) annotations of sample-B: target 1 in (0s, 8.3s), (c) annotations of sample-C: target 3 in (0s, 5.2s), target 2 in (4.7s, 16.8s).

causal localization head. For the experiment about patch size in PDE, large gains are confirmed when the window size is over 45, which denotes that the lower bound to contain the gait patterns should be about 45. Below the size of 45, we confirm that performance is highly decreased. This means that meaningful, discernible gait information is completely contained in more than 45 frames. After that, the performances get saturated when increasing the size over 45. For the stride of the sparsity mask in the causal localization head, we experiment with various strides, where stride 3 is the best performed. The stride over 3 shows deterioration in the performances, which denotes that there is redundancy among the predictions, which loses the opportunities to be searched to many different regions in MD signatures. Therefore, it is confirmed that the redundancy problem in the predictions is mitigated with the stride between 2 and 3. Values above 3 show a decrease in performances which means that the candidate areas containing the pertinent temporal boundaries are being removed due to sparsity masking. Figure 11 presents the efficiency analysis by the proposed causal prediction process. Figure 11 (b) illustrates the process of temporal localization in MD signatures in CLH, such that we first perform the frame-level predictions. Based on the predictions, the start-time is predicted, where the end-time is also predicted based on the two predictions (*i.e.*, frame-level predictions and start-time predictions). By providing the prior information for the localization, this process causally

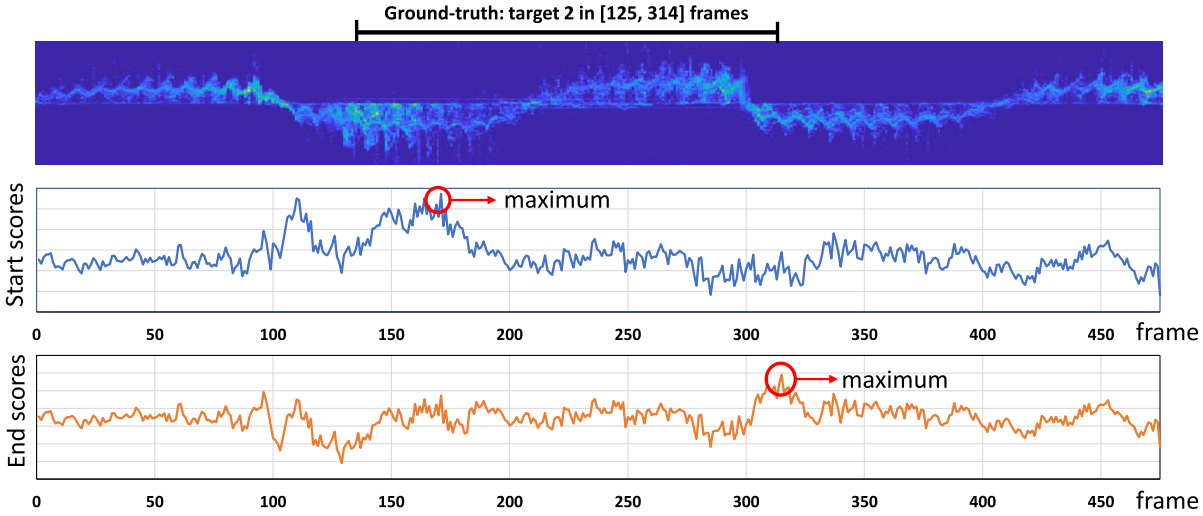


FIGURE 13. Prediction scores about input target 2 by CLNet in terms of start-time (blue curve) and end-time (red curve) along the frame axis. The red circle denotes the maximum point of the curve. The ground-truth temporal boundary is presented above the synthesized MD signatures.

narrows the search space. To identify the effectiveness of the causal prediction process, we also prepare an independent prediction process in Figure 11 (c), where all the predictions are predicted by themselves without any prior predictions. Figure 11 (a) summarizes the validation loss according to these two processes, where it is confirmed that the causal prediction process contributes to early saturation of validation loss and also further optimization of the loss compared to the independent prediction process. We consider this is because the process of progressively narrowing search space with prior information (*i.e.*, previous prediction) is effective in localizing the target MD signatures.

D. QUALITATIVE RESULTS

1) SYNTHESIZED MD SIGNATURES

Figure 12 presents examples of synthesized MD signatures in IDRad-TBA. The annotations about all the samples are also provided in the caption of Figure 12. For samples (*i.e.*, sample-A,B,C), several annotations are attached, where the sample shows natural synthesis to the extent that humans cannot tell where the stitching is applied. In addition, even when two signals in opposite positions are attached (*e.g.*, the stitched region between target 2 and target 5 in sample-A), they are also naturally connected without any seams in the MD signatures. The proposed gradual region-based Doppler stitching method properly synthesized the MD signatures for the radar human localization task.

Furthermore, to assess the alignment of our synthesized MD signatures with the raw MD signatures from the real environment, we conducted an analysis using t-SNE plots comparing the synthetic data to the raw data, as depicted in Figure 14. Although it is not available to check all the conditions necessary for real data, we approximate this using t-SNE similarity analysis, which provides us with a visual

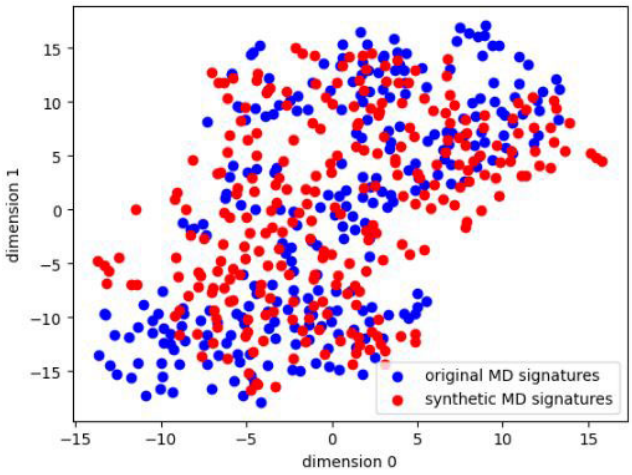


FIGURE 14. t-SNE plots between synthesized MD signatures and original MD signatures.

representation of the distribution and clustering patterns of the synthetic data in relation to the real data. These results qualitatively demonstrate substantial overlap, allowing us to infer that our synthesized data does not encompass information related to impractical scenarios. We consider this because our proposed synthesizing algorithm combines the raw MD signatures along the time axis, which preserves the characteristics of the original MD signatures.

2) PREDICTIONS OF CLNET

Figure 13 shows the predictions about start-end scores for the localization. The black line above the MD signatures shows the temporal boundary of target 2. From the input MD signature and one-hot encoding of target 2, CLNet produces the predictions of start time and end time pertinent to target 2. The max point of each prediction is located near the

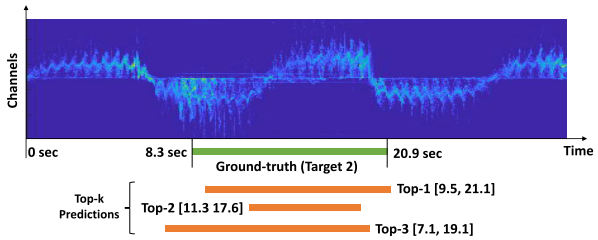


FIGURE 15. Qualitative results on radar human localization of CLNet with respect to the top-k predictions.

ground-truth start-end points. Based on the two predictions, we build joint probabilities and perform radar localization about target 2. The localization results are presented in Figure 15 (*i.e.*, The two MD signatures in Figure 13 and Figure 15 are identical signatures.), where the green box denotes the ground-truth of temporal boundaries of the target and the red boxes denote the top-k (*i.e.*, k=3) predictions of CLNet. The top-1 prediction shows a high overlap with the ground-truth, moreover, the other predictions also have overlaps. It is also notable that the CLNet predicts different lengths of MD signatures, which avoids the redundancy problem among the candidate boundaries. We consider that the sparsity mask contributes to mitigating the redundancy problem.

VI. LIMITATION

This paper proposes radar human localization in micro-Doppler signatures of human walking. The limitations of this work are summarized as follows: (1) The proposed IDRad-TBA dataset is based on the synthesizing process to build samples for the dataset, (2) The radar human localization in this work assumes MD signatures about the existence of a single person at any time. To be specific, for the first limitation, the sample data in IDRad-TBA dataset is based on the IDRad [4] dataset, which is the MD signature for radar human identification. Thus, the sample of IDRad is corresponding to a single person and IDRad-TBA synthesizes the samples of IDRad to make MD signatures with multiple humans in sample data. The second limitation of our work is that since the synthesized sample is based on MD signatures of IDRad, the IDRad-TBA shares the same limitation as IDRad. Therefore IDRad-TBA has a limitation that it does not allow the co-occurrence of multiple humans at the same time. Thus our future work is to build a dataset to perform a more general format of RHL tasks by building real environmental data under more diverse conditions such as the co-occurrence of human and outdoor environments. Furthermore, we also consider extending the work of the current training framework of CLNet to be performed in weakly-supervised settings [26], [27], which mitigates the reliance on temporal annotations to train localization in MD signatures.

VII. CONCLUSION

This paper proposed radar human localization (RHL) on micro-Doppler signatures. To perform RHL, we build dataset IDRad-TBA dataset, which synthesizes the radar signals and annotates the temporal boundaries of each human in the MD signatures. Henceforth, we present a baseline system of RHL referred to as Causal Localization Network (CLNet), where it predicts temporal boundaries pertinent to input target human in the MD signatures of human walking. Experimental results contribute to validating the possibilities of RHL task in MD signatures and the efficiency of proposed causal prediction.

REFERENCES

[1] D. Kang and D. Kum, "Camera and radar sensor fusion for robust vehicle localization via vehicle part localization," *IEEE Access*, vol. 8, pp. 75223–75236, 2020, doi: 10.1109/ACCESS.2020.2985075.

[2] J. Bai, S. Li, L. Huang, and H. Chen, "Robust detection and tracking method for moving object based on radar and camera data fusion," *IEEE Sensors J.*, vol. 21, no. 9, pp. 10761–10774, May 2021, doi: 10.1109/JSEN.2021.3049449.

[3] B. Yang, R. Guo, M. Liang, S. Casas, and R. Urtasun, "RadarNet: Exploiting radar for robust perception of dynamic objects," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 496–512.

[4] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, A. Bourdoux, W. De Neve, and T. Dhaene, "Indoor person identification using a low-power FMCW radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 7, pp. 3941–3952, Jul. 2018.

[5] S. Yoon, D. Kim, J. W. Hong, J. Kim, and C. D. Yoo, "Dual-scale Doppler attention for human identification," *Sensors*, vol. 22, no. 17, p. 6363, Aug. 2022.

[6] V. C. Chen, F. Li, S.-S. Ho, and H. Wechsler, "Micro-Doppler effect in radar: Phenomenon, model, and simulation study," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 42, no. 1, pp. 2–21, Jan. 2006.

[7] T. Gong, Y. Cheng, X. Li, and D. Chen, "Micromotion detection of moving and spinning object based on rotational Doppler shift," *IEEE Microw. Wireless Compon. Lett.*, vol. 28, no. 9, pp. 843–845, Sep. 2018, doi: 10.1109/LMWC.2018.2858552.

[8] N. Seddon and T. Bearpark, "Observation of the inverse Doppler effect," *Science*, vol. 302, no. 5650, pp. 1537–1540, Nov. 2003.

[9] D. Gusland, J. M. Christiansen, B. Torvik, F. Fioranelli, S. Z. Gurbuz, and M. Ritchie, "Open radar initiative: Large scale dataset for benchmarking of micro-Doppler recognition algorithms," in *Proc. IEEE Radar Conf. (RadarConf21)*, May 2021, pp. 1–6.

[10] X. Li, Y. He, and X. Jing, "A survey of deep learning-based human activity recognition in radar," *Remote Sens.*, vol. 11, no. 9, p. 1068, May 2019.

[11] D. Lee, H. Park, T. Moon, and Y. Kim, "Continual learning of micro-Doppler signature-based human activity classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2020.3046015.

[12] J. Martinez and M. Vossiek, "Deep learning-based segmentation for the extraction of micro-Doppler signatures," in *Proc. 15th Eur. Radar Conf. (EuRAD)*, Sep. 2018, pp. 190–193, doi: 10.23919/EuRAD.2018.8546638.

[13] S. Abdulatif, Q. Wei, F. Aziz, B. Kleiner, and U. Schneider, "Micro-Doppler based human–robot classification using ensemble and deep learning approaches," in *Proc. IEEE Radar Conf. (RadarConf18)*, Apr. 2018, pp. 1043–1048.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[15] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using an artificial neural network," in *Proc. IEEE Antennas Propag. Soc. Int. Symp.*, Jul. 2008, pp. 1–4.

[16] J. Park, R. Javier, T. Moon, and Y. Kim, "Micro-Doppler based classification of human aquatic activities via transfer learning of convolutional neural networks," *Sensors*, vol. 16, no. 12, p. 1990, Nov. 2016.

[17] Y. Lin, J. Le Kernec, S. Yang, F. Fioranelli, O. Romain, and Z. Zhao, "Human activity classification with radar: Optimization and noise robustness with iterative convolutional neural networks followed with random forests," *IEEE Sensors J.*, vol. 18, no. 23, pp. 9669–9681, Dec. 2018.

[18] P. Cao, W. Xia, M. Ye, J. Zhang, and J. Zhou, "Radar-ID: Human identification based on radar micro-Doppler signatures using deep convolutional neural networks," *IET Radar, Sonar Navigat.*, vol. 12, no. 7, pp. 729–734, Jul. 2018.
[19] S. Yoon, D. Kim, J. Kim, and C. D. Yoo, "Cascaded MPN: Cascaded moment proposal network for video corpus moment retrieval," *IEEE Access*, vol. 10, pp. 64560–64568, 2022.
[20] P. Cao, W. Xia, and Y. Li, "Heart ID: Human identification based on radar micro-Doppler signatures of the heart using deep learning," *Remote Sens.*, vol. 11, no. 10, p. 1220, May 2019.
[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.
[22] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[23] J. Gao, C. Sun, Z. Yang, and R. Nevatia, "TALL: Temporal activity localization via language query," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5277–5285.
[24] S. Yoon, J. W. Hong, E. Yoon, D. Kim, J. Kim, H. S. Yoon, and C. D. Yoo, "Selective query-guided debiasing for video corpus moment retrieval," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2022, pp. 185–200.
[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[26] S. Yoon, D. Kim, J. W. Hong, J. Kim, K. Kim, and C. D. Yoo, "VLANet: Video-language alignment network for weakly-supervised video moment retrieval," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Cham, Switzerland: Springer, Aug. 2020, pp. 156–171.
[27] S. Yoon, D. Kim, J. W. Hong, J. Kim, K. Kim, and C. D. Yoo, "Weakly-supervised moment retrieval network for video corpus moment retrieval," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 534–538.
[28] N. Rojhani, M. Passafiume, M. Sadeghibakhi, G. Collodi, and A. Cidronali, "Model-based data augmentation applied to deep learning networks for classification of micro-Doppler signatures using FMCW radar," *IEEE Trans. Microw. Theory Techn.*, vol. 71, no. 5, pp. 2222–2236, May 2023.
[29] A. Erdoğan and S. Güney, "Object classification on noise-reduced and augmented micro-Doppler radar spectrograms," *Neural Comput. Appl.*, vol. 35, no. 1, pp. 429–447, Jan. 2023.
[30] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

SUNJAE YOON (Member, IEEE) received the B.S. degree from the Daegu Gyeongbuk Institute of Science & Technology, in 2019, and the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, in 2021, where he is currently pursuing the Ph.D. degree. His research interests include video search technologies, visual-language reasoning, causal reasoning, diffusion, and remote sensing.

GWANHYEONG KOO received the B.S. degree in electrical engineering from the Daegu Gyeongbuk Institute of Science & Technology, in 2023. He is currently pursuing the master's degree with the Korea Advanced Institute of Science and Technology. His current research interests include image and video processing, multimodal learning, and diffusion.

JUN YEOP SHIM received the B.Eng. degree in electronic engineering from The University of Manchester, U.K., in 2020. He is currently pursuing the M.S. degree with the Korea Advanced Institute of Science and Technology. His current research interests include image and video processing and multimodal learning.

SOOHWAN EOM received the B.S. degree in electronic engineering from the Korea Advanced Institute of Science and Technology, in 2022, where he is currently pursuing the M.S. degree. His current research interests include image and video processing and speech representation learning.

JI WOO HONG (Member, IEEE) received the B.S. degree in mechanical engineering from Michigan State University, in 2019, and the M.S. degree from the Robotics Program, Korea Advanced Institute of Science and Technology, in 2022, where he is currently pursuing the Ph.D. degree. His research interests include 3D human pose and shape estimation and visual-language reasoning.

CHANG D. YOO (Senior Member, IEEE) received the B.S. degree in engineering and applied science from the California Institute of Technology, the M.S. degree in electrical engineering from Cornell University, and the Ph.D. degree in electrical engineering from the Massachusetts Institute of Technology. From January 1997 to March 1999, he was a Senior Researcher with Korea Telecom (KT). Since 1999, he has been a Faculty Member of the Korea Advanced Institute of Science and Technology (KAIST), where he is currently a tenured Full Professor with the School of Electrical Engineering and an Adjunct Professor with the Department of Computer Science. He was the Dean of the Office of Special Projects and the Office of International Relations.