FM-Fi 2.0: Foundation Model for Cross-Modal Multi-Person Human Activity Recognition

Yuxuan Weng, Tianyue Zheng✉, *Member, IEEE*, Yanbing Yang, *Member, IEEE*, Jun Luo, *Fellow, IEEE*

Abstract—Radio-Frequency (RF)-based Human Activity Recognition (HAR) rises as a promising solution when low-light, obstructions, or privacy concerns render computer vision impractical. However, the *scarcity* of labeled RF data due to their non-interpretable nature poses a significant obstacle. Thanks to the recent breakthrough of *foundation model*s (FMs), extracting deep semantic insights from unlabeled visual data become viable, yet these vision-based FMs fall short when applied to small RF datasets. To bridge this gap, we introduce FM-Fi 2.0, an innovative cross-modal framework engineered to translate the knowledge of vision-based FMs for enhancing RF-based, multi-person HAR systems. FM-Fi 2.0 first employs the intrinsic capabilities of FM and RF modality to associate both intra- and cross-modal features of each subject, while simultaneously filtering out irrelevant features to achieve better alignment between the two modalities. FM-Fi 2.0 also employs a cross-modal *contrastive* knowledge distillation mechanism, enabling an RF encoder to inherit the interpretative power of FMs for achieving zero-shot learning. The framework is further refined through metric-based few-shot learning techniques, aiming to boost the performance for predefined HAR tasks. Comprehensive evaluations evidently indicate that FM-Fi 2.0 rivals the effectiveness of vision-based methodologies, and the evaluation results provide empirical validation of FM-Fi 2.0's generalizability across various environments.

Index Terms—Human activity recognition, foundation model, RF sensing.

1 Introduction

With rapid developments [1], [2], Human Activity Recognition (HAR) gains significant interest in smart homes [3], [4], digital healthcare [5], [6], and human-computer interaction [7], [8]. In practice, HAR tasks can be either contact-based [9], [10], [11] or contact-free [12], [13]; the latter offers the advantage of not imposing the additional discomfort of wearing devices. Among all sensing modalities for contact-free HAR, Radio-Frequency (RF) sensing [14], [15], [16], [17] stands out by demanding minimal resource for data processing and inference, rendering it ideal for edge device integration. Additionally, it preserves privacy while providing sufficient resolution by capturing only contours without identity-specific features (e.g., facial characteristics and clothing attributes), while being free of visual constraints [18], [19], [20] such as low-light or haze. Therefore, RF-HAR is deemed as a promising solution.

Whereas being effective to specific HAR tasks, RF sensing is hindered by data scarcity and difficulties in annotation. In fact, comprehensive RF datasets are scarce, and the available ones often suffer from compatibility issues due to the diversity in RF devices. This is caused by the significant challenges in annotating RF-sensing data [21]: Unlike image data, human annotators find it impossible to intuitively recognize activities from RF data (especially when there are multiple human subjects in the scene), complicating offline annotation. As a result, annotators must resort to online labeling, posing stringent demands on their skills and increasing the difficulty in verifying data quality after annotation. Therefore, creating a comprehensive RF-HAR dataset incurs prohibitive costs yet lack guaranteed data reliability, largely confining the adoption of RF-HAR.

The recent advent of Foundation Models (FMs) [22], [23], [24] presents a promising solution for addressing the scarcity of labeled data in RF-HAR. Due to their large scale and multimodal training on massive datasets, these models have acquired comprehensive knowledge. In particular, FMs [25] are trained through an unsupervised process that aligns different data modalities within a high-dimensional space, enabling them to process and understand diverse inputs. Such capabilities enable FMs to generalize across diverse domains, and support applications such as zero-shot image classification [25], [26], [27], object detection [28], [29], and image generation [24], [30]. In particular, the comprehensive knowledge and zero-shot capability of FMs could be crucial to overcome the inherent scarcity of labeled data in RF sensing, and they may also bear the potential to push RF-HAR towards *open-set* recognition [31]. Now the question becomes: *can FMs be harnessed to interpret multi-person RF-HAR data?* A valid answer to this question is essential for advancing RF-HAR towards practical adoption.

Despite the potential of FMs in various domains, applying them to interpret RF-HAR data presents several unique challenges. First, the majority of existing FMs have been primarily developed for tasks in computer vision (CV) [23] and natural language processing (NLP) [22], [25], thus limiting their direct applicability to RF-HAR. Although cross-modal knowledge distillation (KD) [32] paves the way for knowledge transfer from image to RF modality, their efficacy in adapting to the structured embeddings of FMs remains unexplored. Second, the image and RF modalities exhibit

- *Y. Weng and T. Zheng are with the Department of Computer Science and Engineering, Southern University of Science and Technology, China. E-mail: {wengyx, zhengty}@sustech.edu.cn*
- *Y. Yang is with the College of Computer Science, Sichuan University, China. E-mail: yangyanbing@scu.edu.cn*
- *J. Luo is with the College of Computing and Data Science, Nanyang Technological University, Singapore. E-mail: junluo@ntu.edu.sg*
- ✉ *Corresponding author: Tianyue Zheng.*

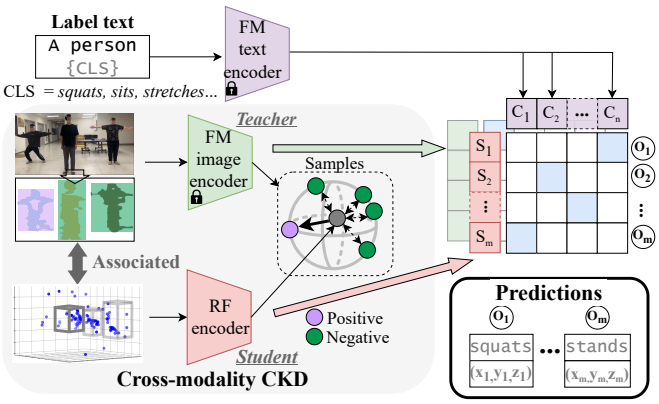


Fig. 1: Overview of FM-Fi 2.0.

inherent feature discrepancies, making the association between these modalities inherently challenging. This feature correspondence problem becomes even more complex in scenarios with multiple human subjects present in the scene. Third, while FMs produce informative embeddings, their optimal use in HAR requires further fine-tuning. However, this fine-tuning process is hindered by the scarcity (or void) of labeled data.

To tackle these challenges, we design FM-Fi 2.0, a cross-modal framework that distills the knowledge from FMs to the RF modality, as illustrated in Fig. 1. First, FM-Fi 2.0 harnesses the intrinsic capabilities of FM and RF to eliminate extraneous and background features and precisely locating individual human subjects within the scene. By leveraging the semantic and spatial relationships between these modalities, FM-Fi 2.0 effectively associates distinctive features across sensing methods for each human subject. Second, given that conventional KD does not consider the structures and interdependencies among the embeddings generated by FMs, we design a novel *contrastive knowledge distillation* (CKD) for transferring knowledge from FM to the neural model for the RF modality. As opposed to conventional KDs, our CKD stems from the mutual information between the embeddings of two modalities: since the interdependency among the embeddings' elements is captured as a form of "information", they can thus be better preserved during distillation. Finally, FM-Fi 2.0 harnesses a minimal set of annotated data to fine-tune its model via metric-based few-shot learning for further adaptation. The synergy of these mechanisms sets the stage for the RF encoder to acquire the full capabilities of the FMs, while opening the way for approaching open-set HAR given the constant improvement of FMs. In summary, our key contributions are:

- We construct the first cross-modal distillation system, FM-Fi 2.0, specifically designed to transfer knowledge from vision foundation models to RF models for multi-person HAR, and evaluate it through extensive experiments. The results demonstrate its strong performance in zero/few-shot multi-person HAR scenarios.
- We design feature association methods tailored to image and RF modalities to achieve association for individual subjects while eliminating extraneous features.
- We develop a CKD mechanism to accommodate FM's intrinsic embedding dependencies, enabling knowledge transfer from FMs to RF modality.
- We design a metric-based few-shot learning mechanism to fine-tune the RF encoder, thereby adapting and enhancing it for specific closed-set HAR tasks.

In the following, § 2 introduces the background and motivation of FM-Fi 2.0. § 3 presents the system design of FM-Fi 2.0.
S 4 introduces the datasets and system implementation, while § 5 reports the experimental setup and the evaluation results. Related and future works are discussed in § 6. Finally, § 7 concludes the paper with future directions.

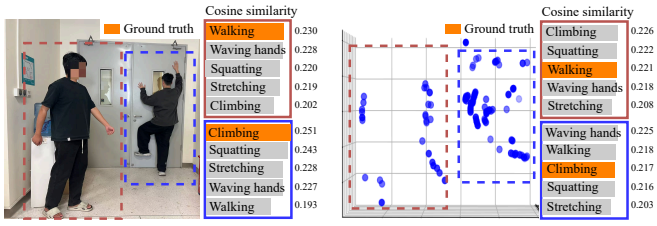
2 Background and Motivations

In this section, we introduce the background of FM for HAR and the motivations of FM-Fi 2.0's design.

2.1 FM for HAR

FMs represent a novel category of large-scale neural networks trained on datasets comprising billions of samples. The training occurs across multiple GPUs over a span of several weeks. Their rapid adoption across various domains, such as CV (e.g., DALLE [24] for image generation), NLP (e.g., GPT [22] for chatbot), and multimodal applications (e.g., CLIP [25] for image semantics understanding), have demonstrated their extensive capabilities. The enhanced image understanding in FMs is facilitated by the adoption of transformer [23], [33] architecture as encoders, which enable the derivation of complex representations. Additionally, contrastive learning [34], [35] has been exploited to align embeddings across different modalities, integrating visual data with semantic insights. Last but not least, the training methodology benefits from the use of unlabeled image-text pairs, allowing for the creation of large-scale training datasets. All these properties have enabled FMs to accurately align image and label embeddings for classification tasks regardless of sample dependency.

The interpretive power of FMs makes them ideal tools for conducting HAR. To give an example, as shown in Fig. 2a, the CLIP model successfully performs zero-shot recognition of human activities in each bounding boxes (provided by manual annotations) by computing similarities between image embeddings and textual descriptions. However, translating these FMs, which were initially trained on vision-text pairs, to RF data presents significant challenges. The fundamental issue lies in the sparsity of RF signals which obscures key physical boundaries, making it particularly challenging to differentiate between individuals. As illustrated in Fig. 2b, directly applying the CLIP model has falsely identified the point cloud captured by a mmWave



(a) Image modality. (b) RF modality.

Fig. 2: Performance of FM for HAR.

radar as wrong activities. This limitation underscores the necessity of novel methods for RF data processing to extend the applicability of FMs beyond visual data.

2.2 Challenges of Feature Alignment

Due to the infeasibility of directly applying FM to RF data, cross-modal knowledge transfer emerges as a viable solution. However, the fundamental premise of knowledge transfer, i.e., features can be effectively aligned across different modalities, faces substantial limitations when applied to image and RF modalities. First, image and RF modalities possess distinct feature sets, with each containing modality-specific characteristics that may be irrelevant to HAR. For instance, images capture extraneous environment elements such as lighting conditions and background objects, while RF data incorporates static background reflections irrelevant to HAR. Second, scenarios involving multiple subjects further complicate this misalignment, as existing knowledge transfer lacks robust mechanisms to keep instance-wise feature correspondence across subjects.

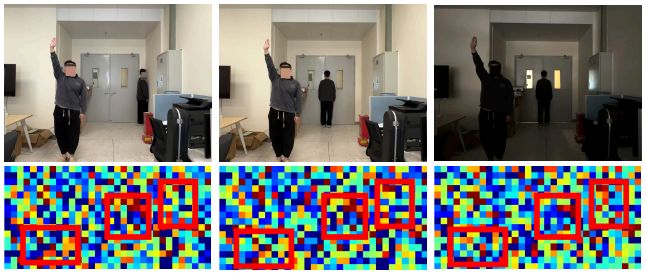
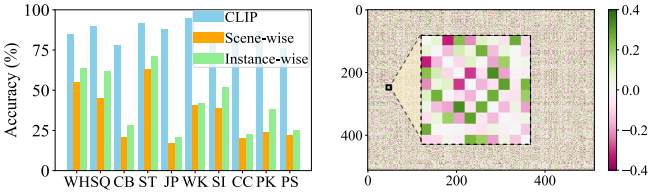


Fig. 3: Minor background variations significantly alter the output embeddings of FM.

To illustrate the challenges of cross-modal feature alignment, we analyze feature maps from the image modality under varying background subject activities and lightings, as shown in Fig. 3. The figure presents scene images in the upper row with their corresponding feature maps below. Our analysis reveals that even subtle environment changes (e.g., lighting variations or shifts in the position of background individuals) substantially impact the embedding representations. To better illustrate the representation difference, we highlight the most distinct regions in the feature maps with red bounding boxes. These perturbations in the image modality features create significant obstacles for robust cross-modal alignment. Given the sensitive nature of these representations, we hypothesize that similar instabilities manifest in the RF modality, further complicating the alignment process. As a result, there is no straightforward one-to-one correspondence between the embeddings of image and RF modalities. Consequently, this feature misalignment hinder the knowledge transfer from the image to the RF modality, necessitating the development of a method to efficiently associate features across these two modalities.

2.3 Why Conventional KD Fails for FMs?

Knowledge transfer involves transferring the knowledge from an FM to RF model by aligning their output embeddings, where we can utilize the mean squared error (MSE) loss for an element-wise comparison of embeddings between image and RF modalities. We employ a synchronized image-RF dataset in our experiment, whose classes will be detailed in § 5.1, to assess the zero-shot HAR performance, by comparing a CLIP model with an RF model trained via a standard KD [32]. In the experiment, we conduct both scene-wise and instance-wise distillation, a distinction usually not explored in single-person HAR because scene and instance levels are effectively equivalent in that context [36]. One may readily observe that a naive application of KD on FMs leads to inferior performance, as depicted in Fig. 4a: the CLIP-trained RF encoder achieves an average accuracy slightly above 36.7% in scene-wise distillation and 49.8% in instance-wise distillation. In contrast, the accuracy achieved by the baseline CLIP model exceeds 80.3%.



(a) Zero-shot HAR comparison. (b) $|\mathbf{R}^{\mathrm{IM}} - \mathbf{R}^{\mathrm{RF}}|$.

Fig. 4: Conventional KD performance.

To understand KD's ineffectiveness, we explore the interdependencies among elements of the output embeddings under the more effective instance-wise distillation setting. We compute the correlation matrices, $\mathbf{R}^{\mathrm{IM}}$ for the FM (processing the image modality) and $\mathbf{R}^{\mathrm{RF}}$ for the RF model, respectively. By subtracting $\mathbf{R}^{\mathrm{RF}}$ from $\mathbf{R}^{\mathrm{IM}}$, we obtain a difference matrix as shown in Fig. 4b. One may readily observe that the correlation difference of the two embeddings can be significant and reach up to 0.4. This finding reveals the limitation of KD: while it aligns the embeddings from the FM and RF model on an element-wise basis, it fails to account for the interdependencies among the elements of the FM's embeddings [37]. The interdependency is especially important for HAR, it is essential that latent factors representing the human subject, various body parts, and activity states should be related and active, while other irrelevant factors should also be related but suppressed. We forward reference to Fig. 8b in § 3.3 for a better correlation matrix difference that better captures the interdependencies among the elements in the embeddings.

3 System Design

We hereby present FM-Fi 2.0 with four components: i) a multimodal feature association module that aligns instance-wise subject features, ii) an RF encoder that encodes instance-wise subject information from the RF point clouds, iii) a cross-modal CKD framework for transferring semantic representations from visual feature maps to RF-based models, iv) a zero/few-shot HAR mechanism relying on learned associations between the semantics of both RF and (FM's) text modalities, and enabling FM-Fi 2.0 to quickly adapt to various closed-set HAR tasks with few labeled examples. In the following, we elaborate on each component, given the overall design depicted in Fig. 5.

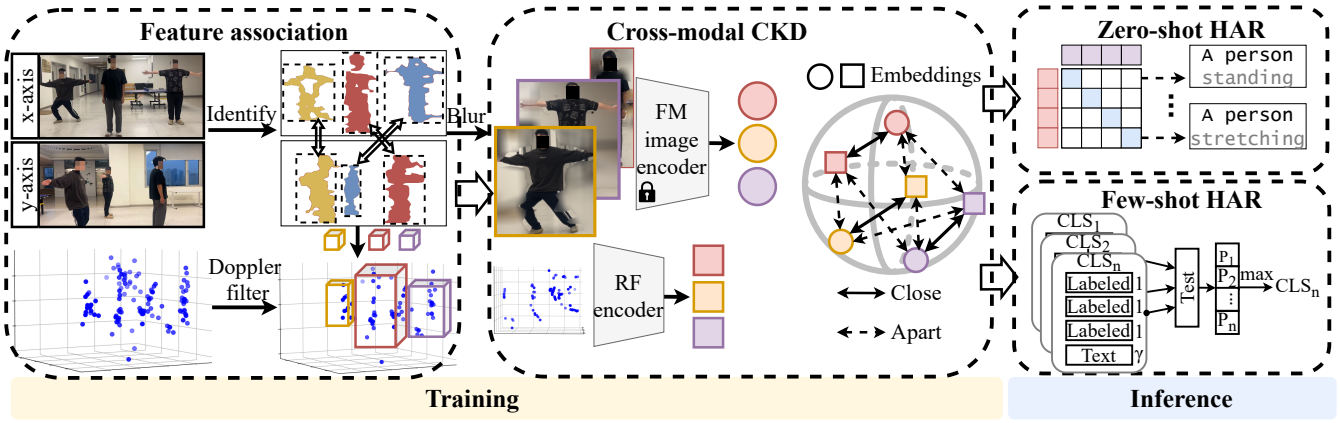


Fig. 5: Overall design of FM-Fi 2.0.

3.1 Instance-wise Feature Association

According to § 2.2, it is necessary to remove HAR-irrelevant extraneous features and keep instance-wise feature correspondence prior to knowledge transfer. To improve interpretability and reduce the consumption of computational resources, we perform instance-wise feature association and elimination of extraneous features utilizing signal properties and FM's knowledge without extra modules. Compared to single-person HAR methods [36], FM-Fi 2.0 introduces additional steps, as explained below, to explicitly associate each subject across different modalities, thus inherently suppressing irrelevant background features.

3.1.1 Image Modality

Since a single camera captures only 2-D information about human positions, it is insufficient for accurately determining a subject's location in 3-D space. To overcome this limitation, we employ two orthogonally placed cameras, with the full process illustrated in Fig. 6. In this setup, each of the cameras leverage the intrinsic capabilities of the FM to locate the human subjects. To be specific, instead of employing extra object detection or image segmentation methods for locating the subjects, we employ the image and text encoders from the teacher model in the knowledge transfer process to generate similarity maps. A similarity map M has the same dimensions as the input image, where $M(u, v)$ stems from the similarity between the corresponding pixel embedding and a text embedding $\mathbf{E}^{\mathrm{TX}}$ of *"A photo of humans doing activities with bodies, arms and legs"*:

$$M(u, v) = \mathrm{sim}(\mathcal{F}(u, v), \mathbf{E}^{\mathrm{TX}}) \tag{1}$$

where $\mathcal{F}$ consists of pixel-level embeddings of the input x^{IM}, derived from the intermediate layer data of the CLIP model, (u, v) represents the pixel coordinates, and $\mathrm{sim}(\cdot)$ computes the cosine similarity between two embeddings. Due to the parallel computation mechanism of the Transformer architectures, each input token preserves its individual representation across all layers, up to the final projection head. In ViT-based CLIP models, these tokens correspond to fixed-size image patches (e.g., 16×16 pixels). By extracting intermediate patch-level embeddings prior to the final pooling or projection stage, and subsequently applying spatial interpolation, we can construct a dense pixel-wise embedding map. This approach enables us to approximate per-pixel semantic features while keeping the CLIP encoder unchanged. This kind of pixel-level operation enables the isolation of image regions that are pertinent to human activity, allowing for the exclusion of non-essential features. As a result, the human instance maps (i.e., processed candidate region) can be expressed as $H(u, v) = \mathbb{I}\left[M(u, v) > \lambda \bar{M}\right]$, where $\mathbb{I}(\cdot)$ is the indicator function, λ is a predefined threshold parameter, and $\bar{M}$ is the mean value of the map M. Elements that exceed the threshold retain their original pixel values and proceed to the next step of fine-grained segmentation, while those below the threshold are blurred. Concurrently, different connected components in H are each assigned distinct labels and bounding boxes, as depicted in Fig. 6. Compared with other segmentation approaches, FM-Fi 2.0 eliminates the need for additional neural networks, and avoids potential issues that could arise from incompatible weighting method of input features by non-CLIP neural networks. The extensive knowledge and complex architecture of the FM contributes to its accurate outputs and reliable reasoning process. As a result, boundary maps obtained from it efficiently concentrate on the relevant features in images.

After obtaining human instance maps H, we implement a human instance tracking mechanism for both camera viewpoints while performing feature association between them. Our approach leverages the natural continuity of human activities to execute instance tracking based on frame-to-frame overlap, while simultaneously utilizing semantic similarities across different viewpoints for robust matching. Specifically, we select the i-th human instance from H^{x} and the j-th human instance from H^{y}, and compute the similarity of their corresponding original frame regions, where H^{x} and H^{y} are human instance maps from cameras providing views along the x and y directions, respectively. To mitigate errors caused by single-frame analysis, we calculate the mean similarity over a span of N frames starting from the current frame:

$$S(i, j) = \frac{1}{N} \sum_{t=1}^{N} \mathrm{sim}\big(\mathbf{E}^{\mathrm{IM}}(H_i^{\mathrm{x}}), \mathbf{E}^{\mathrm{IM}}(H_j^{\mathrm{y}})\big), \tag{2}$$

where $\mathbf{E}^{\mathrm{IM}}$ represents the vision embeddings generated by CLIP, enabling semantic comparison between instances. These similarities are ranked in descending order, with matching pairs identified starting from the highest similarity. Importantly, once an instance is matched, all subsequent pairs containing that instance are disregarded, ensuring

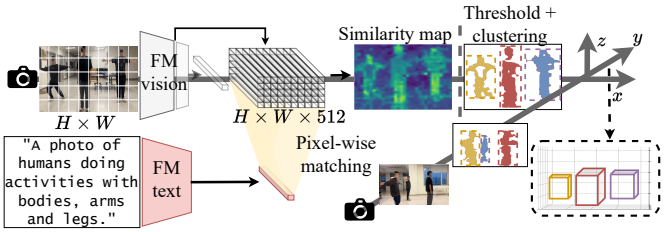


Fig. 6: Feature association of the image modality.

one-to-one correspondence. Upon completion, the bounding boxes from both viewpoints in each matched pair are integrated to form 3-D bounding boxes. The boxes serve as the locational label for the instance, which is subsequently used to train the instance-wise partitioning module in § 3.2.2.

After obtaining a pixel coordinate (u, v) in the image coordinate system, it is necessary to transform it into the real-world Cartesian coordinate system. First, the camera intrinsic matrix K is acquired:

$$K = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where f_x and f_y represent the focal lengths (in pixels) along the horizontal and vertical axes, respectively, and c_x and c_y denote the coordinates of the principal point (optical center) in the image plane. Next, the extrinsic parameters, including the rotation matrix R and the translation vector $T = [t_x, t_y, t_z]^\top$, are obtained. The depth information Z_c is computed using multi-view triangulation. Finally, the 3D Cartesian coordinates of the pixel (u, v) in the real-world reference frame are given by:

$$\begin{bmatrix} X_w \\ Y_w \\ Z_w \end{bmatrix} = R \cdot \left(Z_c \cdot K^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \right) + T. \tag{4}$$

This process enables the transformation of image pixel coordinates into the global Cartesian coordinate system.

3.1.2 RF Modality

Within the RF modality, we first eliminate static backgrounds based on the intrinsic physical properties of the wireless signals. Taking mmWave radar as an example, the sensor emits electromagnetic waves in the range of 30-300 GHz and receives the waves reflected by objects. The raw baseband data collected can be processed to derive information such as distance, angle, and velocity, which can be further transformed into point clouds. Among the information, velocity of an object is inferred through the Doppler effect, which dictates that the frequency shift of the signal is $f_d = \frac{2v}{c} f_0$, where f_d is the frequency difference between the reflected and emitted waves, f_0 is the frequency of the transmitted signal, c is the speed of light, and v is the velocity of the target object relative to the radar sensor. Signals in the point cloud with $f_d = 0$, indicative of static backgrounds, are filtered out to isolate dynamic subjects. It should be noted that, while it is theoretically possible to mistakenly filter out purely tangential activities (characterized by a Doppler velocity of zero), the likelihood of such occurrences is minimal due to the diversity of MIMO sensors and the abundance of data points associated with a single human subject in real-world scenarios.

After removing the static background, it is necessary to distinguish different individuals within the global context. This includes eliminating dynamic backgrounds and separating instance features. To achieve these, we can leverage the supervisory signals provided by the visual modality to learn whether RF signals are emanating from humans or dynamic backgrounds. Specifically, we align the mmWave point cloud with the vision modality's Cartesian coordinate system through coordinate transformation $P_{\text{camera}} = R' P_{\text{radar}} + T'$. Here, P_{camera} and P_{radar} represent points in the coordinate systems of the respective modalities, while R' and T' are the rotation and translation matrices computed based on the sensor's intrinsic and extrinsic parameters. For each point, if its transformed coordinates fall within the three-dimensional bounding boxes mentioned in § 3.1.1, it is considered to belong to the corresponding individual; otherwise, they are classified as part of the dynamic backgrounds. In addition to transformed coordinates, each point also features Doppler velocity and intensity attributes. § 3.2 further provides discusses how these features are assigned to individuals, and generate unique instance-level embeddings.

3.2 Instance-wise RF Encoder

The mmWave data collected for this study is presented as a point cloud, containing information of coordinates, Doppler frequency, and intensity, each of which is indispensable for HAR analysis. Specifically, the point cloud coordinates provide valuable insights into human posture, while intensity reveals the reflection characteristics, and Doppler frequency offers critical dynamic information regarding motion. Before being processed by the neural network, the point cloud undergoes preprocessing, during which their centroid is translated to the origin, effectively eliminating any translational biases. To extract semantic embeddings for each human, the RF encoder in FM-Fi 2.0 operates in two stages: initially, a spatial feature extraction module utilizes self-attention layers to obtain context information for each point, followed by an instance-wise embedding module that clusters these points to corresponding individuals, creating distinctive semantic representations for each person.

3.2.1 Spatial Feature Extraction

Contrary to the inherent order of image pixels, point cloud data is characterized by an absence of order. Furthermore, the coordinates of a point cloud depend on the selected coordinate system. However, neither changing the point order nor the coordinate system should affect the feature extraction outcome. To address these challenges, we revamp the design of PointNet [38] to accommodate the properties of mmWave data, as shown in Fig. 7. FM-Fi 2.0's RF encoder includes a spatial transformation network (STN) $\mathcal{T}$, attention layers, and a maxpooling module. STN aims to learn a 3×3 rotation-scaling matrix $\mathbf{W}_T$, implementing a transformation on each point as $\mathbf{x}' = \mathbf{W}_T \cdot \mathbf{x}$, where $\mathbf{x}$ and $\mathbf{x}'$ represent the original and transformed coordinates, respectively. To derive $\mathbf{W}_T$, the point cloud undergoes processing through convolutional layers and fully connected layers, outputting a 9-dimensional vector reshaped into a 3×3 matrix. Through this process, the STN captures the

relationship between the point cloud's global distribution and implicit viewpoint information, as $\mathbf{W}_T = \mathcal{T}(\mathbf{x})$. This transformation standardizes the point cloud, and improves its robustness against geometric variations.

It is important to note that, in addition to spatial coordinates (x, y, z), mmWave point clouds incorporate two additional features: Doppler frequency and intensity. The Doppler feature provides information about the moving velocity of targets, while intensity is indicative of their distance and material properties. These two features are essential for HAR and are consequently concatenated with the three-dimensional coordinates after STN processing. The resulting feature vector, now enriched with the transformed coordinates and the two additional features, is fed into a module $\mathcal{A}$, consisting of several self-attention layers to enable each point to learn contextual information on a global scale. More specifically, within each layer, we optimize three shared-weight matrices $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$ across all points. The 5-dimensional feature vector $\mathbf{p}$ of each point is transformed into corresponding query $\mathbf{Q} = \mathbf{W}_q \cdot \mathbf{p}$, key $\mathbf{K} = \mathbf{W}_k \cdot \mathbf{p}$, and value $\mathbf{V} = \mathbf{W}_v \cdot \mathbf{p}$. The weighted point vector $\mathbf{p}'$ can be calculated as $\mathbf{p}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$. We then pass the enriched feature vectors $\mathbf{p}'$ through a multilayer perceptron (MLP) ϕ to generate representations for each point.

3.2.2 *Instance-wise Partitioning*

The aforementioned representations are initially used for individual recognition. Specifically, within the individual bounding boxes provided by the visual modality, the representation of each point generates proxy coordinates through a proxy network composed of a MLP. These coordinates are learned to approximate the center of the individual bounding box. If the original point does not fall within any bounding box, it is considered part of the dynamic background, and its proxy coordinates are set to $(0, 0, 0)$, corresponding to the position of the sensor. Therefore, the loss function $\mathcal{L}_{\text{P}}$ of the instance-wise partitioning module can be written as:

$$\mathcal{L}_{\text{P}} = \mathbb{E}_i \left[b_i \|v_i - c_i\|^2 + (1 - b_i)\|v_i - (0, 0, 0)\|^2 \right], \quad (5)$$

where, b_i is a binary variable that equals 1 if point i belongs to a human, and 0 if it belongs to the background. Subsequently, the proxy coordinates are clustered using the DBSCAN algorithm. For each proxy point v_i, its ϵ-neighborhood is computed as:

$$N_\epsilon(v_i) = \{v_j \mid \|v_i - v_j\| \leq \epsilon\}. \quad (6)$$

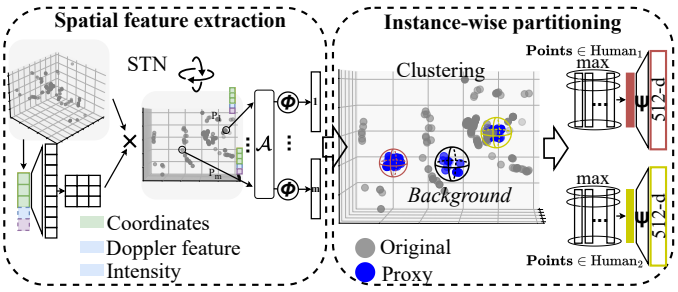


Fig. 7: Instance-wise RF encoder for cross-modal distillation.

A proxy point v_i is considered a core point if $|N_\epsilon(v_i)| \geq$ MinPts. All core points and their reachable neighbors form a cluster C_k, and the set of clusters is denoted as $\{C_1, C_2, \ldots, C_K\}$. As the output of the instance-wise partitioning module, the representations corresponding to each cluster, except for the one closest to the origin, are processed through a maxpooling mechanism. This mechanism selects the maximal value across all points for each element of the embedding, a process that remains invariant to the order of point inputs and equally emphasizes every point in the space. It should be noted that this step processes the point cloud as a whole, rather than focusing on individual points. Subsequent to another MLP, denoted as ψ, the output of the RF encoder is mapped to a 512-dimensional vector. In summary, the point cloud processing of each individual can be expressed as follows:

$$\mathbf{E}_k^{\text{RF}} = \psi \left(\underset{i \in C_k,\, k \neq k_{\text{background}}}{\text{maxpooling}} \left(\phi \left(\mathcal{A} \left(\mathcal{T}(\mathbf{X}_{\text{RF}_i}) \cdot \mathbf{X}_{\text{RF}_i} \right) \right) \right) \right). \quad (7)$$

The 512-dimensional output of the encoder guarantees compatibility with the output from the FM image encoder.

3.3 Cross-Modal CKD

Synchronized vision and RF modalities capturing the same scene offer closely related physical information, such as spatial structure, contours, and dynamic information. As a result, the gap between their semantic embedding spaces can be potentially bridged using knowledge distillation [39]. The first step in conducting KD from FM to RF models involves constructing a data bridge to link the image and RF modalities. Given the scarcity of annotated data, highlighted in Section 2.3, this bridge only employs unlabeled synchronized data gathered from a pair of camera and radar sensor. Specifically, it comprises two data types: i) unstructured data from everyday spontaneous activities, and ii) rehabilitation activity data. The former provides a large amount of data that captures real-world complexities, aiding in model generalization; while the latter includes a wide range of body movements encompassing rare movement cases, thereby offering extensive body variation and motion diversity. This comprehensive data bridge selection ensures the subsequent KD transcends mere recognition of specific movements and body parts under few environments.

Specifically, we collect datasets consisting of paired image and RF data, represented as $(\mathbf{X}_i^{\text{IM}}, \mathbf{X}_i^{\text{RF}})$, where $i = 1, \cdots, N$. These datasets are gathered from the same scenes to bridge the modalities. For each modality, data is processed by the corresponding encoder, producing embeddings $\mathbf{E}^{\text{IM}}$ and $\mathbf{E}^{\text{RF}}$. Note that in CKD, we utilize embeddings $\mathbf{E}^{\text{IM}}$ generated by the camera colocated with the radar as positive and negative samples, while cameras from another perspective are solely used for instance localization. While the representations of different modalities share some common information, they do have some differences that cannot be aligned. This means relying solely on rigid metrics like the Euclidean distance in traditional KD is insufficient, as discussed in § 2.3. Instead, we employ the mutual information between modalities as the starting point for deriving contrastive knowledge distillation (CKD) method, whose

training pipeline is illustrated in Fig. 8a. Mutual information MI can be expressed using KL divergence as:

$$MI(\mathbf{E}^{\mathrm{IM}};\mathbf{E}^{\mathrm{RF}}) = D_{\mathrm{KL}}(p(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}})\|p(\mathbf{E}^{\mathrm{IM}})p(\mathbf{E}^{\mathrm{RF}})), \quad (8)$$

where $p(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}})$ represents the true joint probability distribution of the image (IM) and RF embeddings, while $p(\mathbf{E}^{\mathrm{IM}})p(\mathbf{E}^{\mathrm{RF}})$ represents the product of their marginal distributions, i.e., the joint distribution that would arise if these embeddings were statistically independent. The KL divergence term, $D_{\mathrm{KL}}(\cdot\|\cdot)$ thus measures how much the actual joint behavior of the embeddings $p(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}})$ deviates from this baseline assumption of independence. Consequently, maximizing the mutual information $MI(\mathbf{E}^{\mathrm{IM}};\mathbf{E}^{\mathrm{RF}})$ is equivalent to maximizing the information-theoretic "distance" between the observed joint distribution of the embeddings and the distribution corresponding to their independence. A larger MI value inherently signifies that the embeddings $\mathbf{E}^{\mathrm{IM}}$ (teacher) and $\mathbf{E}^{\mathrm{RF}}$ (student) are more strongly dependent on each other. This increased dependency implies that they share more information, which is the foundation of achieving "structural consistency" across modalities. Our CKD method, by maximizing a lower bound of this MI actively encourages the student model (producing $\mathbf{E}^{\mathrm{RF}}$) to learn representations that capture information highly congruent with, and predictive of, the teacher model's representations $\mathbf{E}^{\mathrm{IM}}$. This direct optimization for maximizing the shared information between teacher and student modalities serves as the core mechanism underpinning effective knowledge transfer and improved cross-modal feature alignment, which is crucial for storing information of the embeddings. Specifically, to distill the interdependency information critical for HAR, CKD maximizes the lower bound of the mutual information MI between the image and RF embeddings $\mathbf{E}^{\mathrm{IM}}$ and $\mathbf{E}^{\mathrm{RF}}$. The mutual information: $MI(\mathbf{E}^{\mathrm{IM}};\mathbf{E}^{\mathrm{RF}}) = \mathbb{E}_{p(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}})}\left[\log\frac{p(\mathbf{E}^{\mathrm{RF}}|\mathbf{E}^{\mathrm{IM}})}{p(\mathbf{E}^{\mathrm{RF}})}\right]$. Assuming $\mathbf{E}^{\mathrm{RF}}$ follows a uniform distribution (i.e., $p(\mathbf{E}^{\mathrm{RF}}) = \frac{1}{N}$), we have:

$$MI(\mathbf{E}^{\mathrm{IM}};\mathbf{E}^{\mathrm{RF}}) = \mathbb{E}_{p(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}})}\left[\log p(\mathbf{E}^{\mathrm{RF}}|\mathbf{E}^{\mathrm{IM}})\right] + \log N.$$

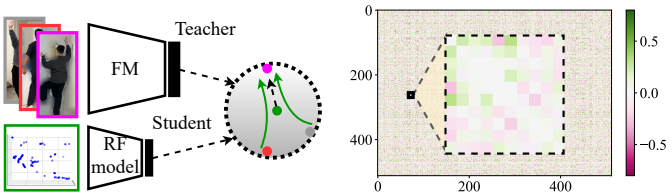
The conditional probability $p(\mathbf{E}^{\mathrm{RF}}|\mathbf{E}^{\mathrm{IM}})$ is estimated as:

$$p(\mathbf{E}^{\mathrm{RF}}|\mathbf{E}^{\mathrm{IM}}) \geq \frac{\exp(\mathrm{sim}(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}}))}{\sum_{\mathbf{E}^{\mathrm{RF}'}\in\mathcal{P}}\exp(\mathrm{sim}(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}'}))}.$$

where $\mathrm{sim}(\cdot)$ measures the similarity between $\mathbf{E}^{\mathrm{IM}}$ and $\mathbf{E}^{\mathrm{RF}}$, and $\mathcal{P}$ is the set of all possible samples $\mathbf{E}^{\mathrm{RF}'}$. Therefore we have $MI(\mathbf{E}^{\mathrm{IM}};\mathbf{E}^{\mathrm{RF}}) \geq \log N - \mathcal{L}_{\mathrm{CKD}}$, where

$$\mathcal{L}_{\mathrm{CKD}} = -\mathbb{E}_{p(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}})}\left[\log\frac{\exp(\mathrm{sim}(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}}))}{\sum_{\mathbf{E}^{\mathrm{RF}'}\in\mathcal{P}}\exp(\mathrm{sim}(\mathbf{E}^{\mathrm{IM}},\mathbf{E}^{\mathrm{RF}'}))}\right],$$

where $\mathrm{sim}(\cdot)$ is defined as $\langle\cdot,\cdot\rangle/\tau$, with $\langle\cdot,\cdot\rangle$ being the cosine similarity, and τ being the temperature scaling parameter. In CKD, for a given RF sample, positive examples are derived from the same individual within the same frame in the vision modality, while all other samples from the vision modality (including those from different individuals in the same frame and the same individual across different frames) are treated as negative examples. It should be noted that, while the mathematical structure of CKD loss may resemble conventional InfoNCE-style [40] losses, its underlying computation process is considerably different. First, the positive samples in CKD are drawn from the teacher modality's embeddings, which eliminates the need for data augmentation. Second, in InfoNCE-style methods, both encoders (e.g., vision and text) are randomly initialized and jointly optimized to align with each other in the representation space. In contrast, CKD adopts an asymmetric optimization strategy: the vision foundation model is fixed to serve as an anchor, and only the RF encoder is optimized to align with it, which significantly reducing computational overhead. Lastly, CKD leverages cosine similarity for measuring similarities of the embeddings, thereby eliminating the reliance on a critic model, as required by another cross-modal distillation baseline CRD [41].



(a) Basic idea. (b) $|\mathbf{R}^{\mathrm{IM}} - \mathbf{R}^{\mathrm{RF}}|$.

Fig. 8: Cross-modal CKD.

As shown in Fig. 8a, CKD reduces the distance between embeddings of positive RF-image pairs, while increasing the separation between negative pairs within the embedding space. This contrastive method enhances the distillation process by more effectively capturing the interdependencies among the embedding elements. Additionally, Fig. 8b shows a significant reduction of 0.2 on average, in the difference between the correlation matrices of the FM and RF, denoted as $|\mathbf{R}^{\mathrm{IM}} - \mathbf{R}^{\mathrm{RF}}|$, when utilizing CKD. This contrasts with the outcomes observed with traditional KD, as depicted in Fig. 4b. This observation underscores CKD's superiority in aligning the structural characteristics of the embeddings across diversified modalities.

3.4 Zero- and Few-Shot HAR

Given that FMs are not trained by simply mapping samples to fixed categories, but rather by understanding the relationship between image content and arbitrary textual descriptions, they are adept at handling certain zero-shot tasks, capable of accurately identifying categories not present in the training set. For instance, CLIP leverages image and descriptive text matching to categorize 1,000 classes in ImageNet within a zero-shot manner. RF models trained under its supervision exhibit similar classification capabilities. Specifically, for any HAR class described in natural language, we can embed it into an appropriate prompt, such as "A person $\{CLS\}$", where CLS denotes activities like "walking" or "squatting". Subsequently, the text description of this class is divided into individual words, known as tokens. Each token is then transformed into a corresponding numerical value that aligns with a vocabulary defined during the encoder's training phase. As a result, the CLIP text encoder processes these representations rather than the original natural language to generate a 512-D embedding.

Following cross-modal CKD, the RF encoder has been endowed with the capability of the vision FMs to embed

spatial information into the semantic space. Consequently, it can embed RF data into 512-dimensional vectors, congruent with the previously described text embedding structure. The cosine similarity between embedding vectors from different modalities serves as the criterion for their congruence, with the highest scoring category being selected for prediction $\hat{t}$. The prediction process can be formulated as $\hat{t} = \arg\max_{\mathbf{E}^{\mathrm{TX}}} \left(\frac{\mathbf{E}^{\mathrm{TX}} \cdot \mathbf{E}^{\mathrm{RF}}}{\|\mathbf{E}^{\mathrm{TX}}\| \|\mathbf{E}^{\mathrm{RF}}\|} \right)$, where $\mathbf{E}^{\mathrm{TX}}$ represents the text embedding of the label. To optimize computation, we stack the text embeddings of all candidate labels to create a matrix $W_{\text{zero-shot}} \in \mathbb{R}^{512 \times k}$, whereby score $= W_{\text{zero-shot}} \cdot \mathbf{E}^{\mathrm{IM}}$. Given that each text embedding is normalized, we identify the category corresponding to the highest score to make prediction. In multi-person scenarios, the output of FM-Fi 2.0 can be represented by a set $O = \{(\hat{t}_1, \hat{l}_1), \ldots, (\hat{t}_i, \hat{l}_i)\}$, where $\hat{l}_i$ is the instance location, and the size of O denotes the number of people in the scene.

While zero-shot learning adequately addresses most HAR tasks, for especially challenging ones characterized by less distinct language descriptions, we introduce an additional few-shot learning module. This module adopts a metric-based approach utilizing a non-parametric method to predict labels in the query set based on a weighted sum of true labels in the support set. In contrast to conventional metric-based learning, FM-Fi 2.0's embedding space is semantically rich. As such, we enhance the performance of classification by utilizing the label text embeddings generated by FMs, further exploiting the semantic information they contain. Specifically, we employ cosine similarity as our metric function following the practice of CLIP, given its superior ability to measure the similarity between semantic vectors. Thus, we determine the likelihood of an unlabeled sample belonging to class c as follows:

$$P(y_c|\mathbf{E}^{\mathrm{q}}, \mathcal{D}_c^{\mathrm{s}}) = \textstyle\sum_{\mathbf{E}_c^{\mathrm{s}} \in \mathcal{D}_c^{\mathrm{s}}} \langle \mathbf{E}^{\mathrm{q}} \cdot \mathbf{E}_c^{\mathrm{s}} \rangle + \gamma \langle \mathbf{E}^{\mathrm{q}} \cdot \mathbf{E}_c^{\mathrm{TX}} \rangle, \quad (9)$$

where $\mathcal{D}^{\mathrm{s}}$ is the support set, $\mathbf{E}^{\mathrm{s}}$ and $\mathbf{E}^{\mathrm{q}}$ denote the embeddings of a support and query sample, $\mathbf{E}_c^{\mathrm{TX}}$ represents the text embedding of class c, and γ is a hyperparameter that signifies the weight of label text. Finally, we take the maximum of the computed likelihoods to yield the prediction.

4 Dataset and Implementation

In this section, we introduce dataset collection/processing, and system implementation of FM-Fi 2.0.

4.1 Dataset

For the RF modality, we acquire data using a Texas Instruments (TI) IWR1443 Boost mmWave radar [42]. This radar operates within the 76-81 GHz frequency spectrum, offering a bandwidth of 4 GHz. It employs a frequency-modulated continuous-wave (FMCW) technique, which transmits a chirp signal that linearly increases in frequency over time. The system, upon receiving the reflected signals from the objects, constructs a point cloud. This point cloud aggregates the data collected over a time span of 200 ms, and contains information such as point coordinates (x, y, z), Doppler features d, and signal intensity I. Our dataset for CKD consists of 90,000 video samples (each 200 ms in length), totaling approximately 5 hours in duration. Given that the frequencies of most human activities lie within the 0.1-10 Hz range [43], we set the radar sampling rate to 20 Hz. After denoising with a constant false alarm rate (CFAR) filter, the resulting point cloud data become $P_i = (x_i, y_i, z_i, d_i, I_i), 1 \leq i \leq N$, where N denotes the number of points per frame.

Similarly, we position a Microsoft Kinect V2 RGB camera [44] at the same conditions as the aforementioned mmWave radar to capture human activities on the X-Z plane, while another camera is positioned orthogonally to capture activities on the Y-Z plane. These cameras are set to capture images with a resolution of 1920×1080 (1080P) and a frame rate of 30 Hz. The Kinect V2 captures raw data streams, which are then converted into JPG format to align with the input requirements of the FM. To synchronize these two modalities, which operate at different sampling rates, we initially establish specific start and end actions to assist in preliminary alignment. Subsequently, we select the lower frequency, i.e., the radar frequency, as a reference and identify the temporally closest camera frame for matching, thereby constructing our dataset.

For data acquisition, the pair of radar and camera sensors are positioned in various locations, including being mounted on different desktops, walls, and ceilings. The subjects' heights range from 152 to 186 cm, weights from 51 to 109 kg, and ages from 10 to 35 years, with an equal distribution of genders. The distance from the sensor to the target ranges from 1 to 15 m. The dataset is collected across 10 distinct environments: kitchen (KC), living room (LR), bedroom (BR), gym (GM), parking lot (PL), hallway (HW), staircase (SC), park (PK), street (ST), stadium (SD), road intersection (RI), and outdoor fitness area (OF). The kitchen, living room, bedroom, and hallway represent limited-space living environments, each furnished with scene-specific items (e.g., different furnitures, hydrants, and ladders). The gym and parking lot are spacious indoor scenes, equipped with fitness equipment and vehicles respectively, and host a modest number of individuals. As outdoor environments, park, street, and stadium are open areas featuring different plants, vehicles, large sports equipment, and pedestrians. The staircase, characterized by its narrow space and complex environment, includes stairs and railings. A road intersection is a high noise environment with dynamic backgrounds including fast-moving vehicles, bicycles, and pedestrians; and an outdoor fitness area is a strong interference setting with metallic equipment moving alongside human activities, which introduces overlapping dynamic interference. Collectively, these 10 different environments exhibit unique floor plans and background objects, underscoring the diversity of real-world scenarios.

Additionally, as elaborated in § 3.3, our dataset is divided into two main parts: everyday spontaneous activities and structured rehabilitation exercises. For the former, approximately 65,000 image-RF data pairs are collected, capturing participants performing activities in accordance with their natural behavior patterns. The latter category encompasses five exercises, each developed in accordance with professional sports rehabilitation guidelines and performed by subjects in compliance with a standardized regimen, ultimately producing approximately 30,000 sample pairs encompassing a broad range of body poses. Notably, this is a newly collected multi-person HAR dataset, which is distinct

from the dataset used in [36].

4.2 System Implementation

We conduct all experiments, including model training, inference, and saliency map generation, on 2 NVIDIA GeForce RTX 4090 GPUs equipped with 48GB of RAM in total. Regarding software, our framework is built upon Python 3.7 and PyTorch version 2.1.0, which supports CUDA 12.1. Additionally, we employ OpenAI's CLIP as our FM teacher model. The CLIP library, released by OpenAI, facilitates easy integration in Python, providing built-in data preprocessing and a selection of vision encoders. For the RF modality, we develop an mmWave point cloud encoder using PyTorch. The specific configurations are as follows:

- We choose ViT-B/32 in CLIP as our vision encoder and a custom mmWave point cloud encoder, outlined in § 3.2, featuring 1-d convolutional and linear layers with batch normalization and a 0.3 dropout rate.
- We set the similarity threshold λ in § 3.1 to 1.2 and select a Gaussian kernel for background blurring.
- Our CKD dataset consists of 90,000 pairs of image and RF data. The labeled RF dataset has 15,000 samples, and is split into validation and test sets at a 9:1 ratio.
- FM-Fi 2.0 uses continuous, non-overlapping frames for training and testing, instead of random frame sampling, to avoid overfitting due to neighboring frames.

5 Evaluation

In this section, we report a thorough evaluation on FM-Fi 2.0 in several scenarios and under various parameter settings.

5.1 Experiment Setup

5.1.1 Baselines and Environments.

To evaluate the performance of FM-Fi 2.0, we select 3 sets of baselines for comparison. First, we compare the FM-Fi 2.0's rapid adaptation capabilities in RF modality for HAR with limited samples against state-of-the-art (SOTA) meta-learning-based RF models, RF-Net [12], mmCLIP [45], and MetaSense [46]. We also compare FM-Fi 2.0 with fully supervised multi-person RF-HAR models, including PALMAR [47], RF-Action [48], and Multi-HAR [49]. Further, we compare the performance of FM-Fi 2.0 against SOTA point-cloud models, PointNet++ [50] and Point Transformer [51]. Lastly, to assess FM-Fi 2.0's performance in unseen environments, we include its teacher model CLIP [25] for comparison. Since these baselines are not designed for multi-person scenarios, we equip them with the instance-wise partitioning module described in § 3.2. Each scene in our experiments contains between 1 and 10 subjects, with an average of approximately 4 subjects per scene.

- **RF-Net** employs a dual-path architecture to discern key RF signal features for HAR and integrates a distance metric network to facilitate few-shot learning.
- **MetaSense** trains on multiple tasks calibrated to individual variances, enabling the model to quickly adapt to new conditions with minimal samples.
- **mmCLIP** aligns high-level representation space of mmWave signals and LLMs' text space to facilitate zero-shot recognition for unseen activities.
- **Multi-HAR** combines group tracking, 3D-CNN, and LSTM to enable robust per-person activity inference from clustered point cloud data.
- **PALMAR** integrates voxel-based fine-tuning, efficient clustering and tracking with an adaptive-order HMM, and adaptive deep domain adaptation.
- **RF-Action** translates the input to an intermediate skeleton-based representation, learns from both vision-based and RF-based datasets, and allows the two tasks to help each other.
- **Point Transformer** introduces a self-attention-based architecture tailored for 3D point cloud analysis that can be used for segmentation and classification tasks.
- **PointNet++** is an extension of the original PointNet architecture, introducing hierarchical feature extraction to better handle local structures in point clouds.

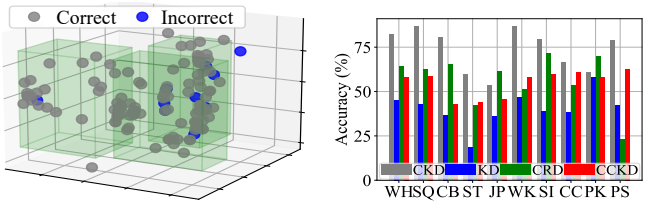
Although FM-Fi 2.0 does not limit the number of HAR classes, we test it on 10 classes for clarity: waving hands WH, squatting SQ, climbing CB, stretching ST, jumping JP, walking WK, sitting ST, cycling CC, picking PK, and pushing PS. We also prepare 10 new classes for further evaluation: running RN, standing SD, lying down LD, crawling CR, playing ball PB, dancing DN, boxing BX, lifting LF, cleaning CL, and doing Yoga YG. To gain insights into the model's predictive distribution, we also employ confusion matrices to visually demonstrate the model's performance on each class. The experiments strictly follow the IRB approved by our institution.

5.1.2 Evaluation Metric.

The two stages in the FM-Fi 2.0 pipeline operate at different levels of granularity (i.e., instance-level separation/association and activity-level recognition), therefore, a single unified metric cannot fairly evaluate the entire system. Consequently, we evaluate FM-Fi 2.0's performance in two corresponding parts. For the former, we compute precision and recall at both the point and subject level. For the latter, to enable a clearer assessment at the activity level, we first obtain a set of bounding boxes from the image modality through instance-wise feature association, with the number of boxes equal to the number of subjects. Each bounding box defines a ground truth instance point set composed of the points within it. We define Point IoU (PIoU) as $\text{PIoU} = \frac{|P_{\text{pred}} \cap P_{\text{st}}|}{|P_{\text{pred}} \cup P_{\text{st}}|}$, where P_{pred} is the set of points in the predicted instance and P_{gt} is the set of points in the ground truth instance. By definition, a higher PIoU signifies better spatial correspondence. Based on this, we set a relatively high threshold of 0.80 and evaluate the instance-wise partitioning module on the validation set (the rationale for selecting the threshold will be explained in § 5.2.1). Given this, our accuracy is computed only on predicted instances that have a PIoU greater than 0.80 with a ground truth instance. Specifically, the accuracy is defined as activity classification success rate for all detected instances, with false positives inherently counted as misclassifications.

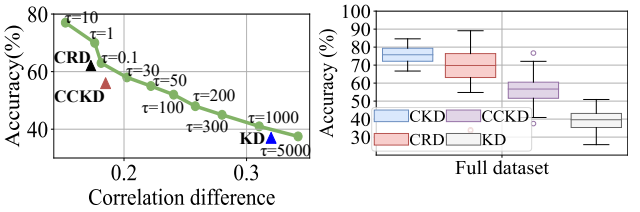
5.2 Micro-benchmark

In this section, we present micro-benchmark studies of the instance-wise partitioning module of the RF encoder, and cross-modal CKD of FM-Fi 2.0.



(a) Instance-wise partitioning. (b) Cross-modal CKD.

Fig. 9: Micro-benchmarks.



(a) Impact of temperature parameter τ. (b) Comparisons on accuracy and stability.

Fig. 10: CKD evaluation.

5.2.1 *Instance-wise Partitioning of the RF Encoder*

We evaluate the RF encoder's effectiveness in distinguishing individual subjects within global input features. The results demonstrate remarkable performance despite challenging conditions. In a test scenario featuring three individuals represented by 128 points shown in Fig. 9a, FM-Fi 2.0 achieves a point classification accuracy of 93.8%, even with ambiguous boundaries between subjects. A detailed analysis presented in Table 1 reveals consistently high performance metrics across all individuals, with precision exceeding 90.0% and recall above 85.7%. This exceptional performance can be attributed to two key features of FM-Fi 2.0: its ability to capture point-to-point relationships in a global context, and its ability to differentiate foreground from background. The latter is achieved through strategic placement of background proxy points at significant distances from human-associated points, substantially reducing clustering misclassification.

	Subject 1	Subject 2	Subject 3
Precision	94.2%	91.7%	90.0%
Recall	89.0%	97.1%	85.7%

TABLE 1: Point-level performance across individuals.

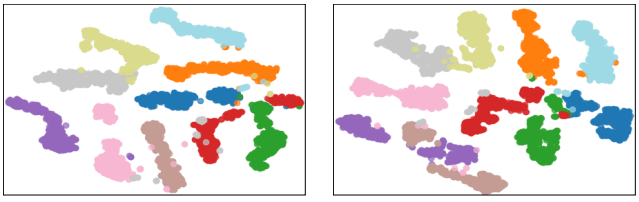
Following point-level evaluation, we transition to a more holistic instance-level assessment. For this, we utilize a PIoU threshold of 0.8, as introduced in § 5.1.2. This means a subject is considered correctly partitioned if the predicted point set P_{pred} (from the RF encoder) achieves at least an 80% overlap with the ground-truth point set P_{gt} (from the vision modality). We conduct 30 trials across diverse scenarios featuring 1 to 10 subjects, each performing different activities independently, with an average of approximately 4 subjects per scene. At this 0.8 PIoU threshold, our instance-wise partitioning module achieves an impressive average precision of 97.6% and an average recall of 98.3%. The module's ability to attain such high performance, signifying negligible false positives and missed detections, justifies the 0.8 PIoU threshold. Specifically, by requiring a substantial 80% overlap, this threshold ensures that only genuinely well-segmented instances are counted as correct, thereby providing a clear and trustworthy foundation for further assessing HAR.

5.2.2 *Cross-modal CKD*

We further compare CKD with KD, and extend the comparison to include contrastive representation distillation (CRD) [41] and correlation congruence for knowledge distillation (CCKD) [37]. We compare their performance on a 10-class zero-shot HAR task, as illustrated in Fig. 9b. We observe that CKD achieves the highest accuracy in 8 out of 10 classes, only trailing the best method by less than 8.5% in the rest 2 classes. Notably, CRD shows the highest variability in accuracies, which can be attributed to the instability inherent in its learning-based critic model used for similarity assessment. CCKD's approach, which prioritizes alignment of instance distributions between image and RF embeddings without addressing the interdependencies among elements, results in suboptimal performance. Similarly, KD's performance is compromised due to its inability to manage the interdependencies within the embeddings' elements. In summary, CKD's advantages arise from: a greater emphasis on the interdependencies of embedding elements compared to CCKD and KD, which transfers critical information to enhance performance; and using cosine similarity instead of a critic model, as in CRD, which reduces model complexity and increases robustness.

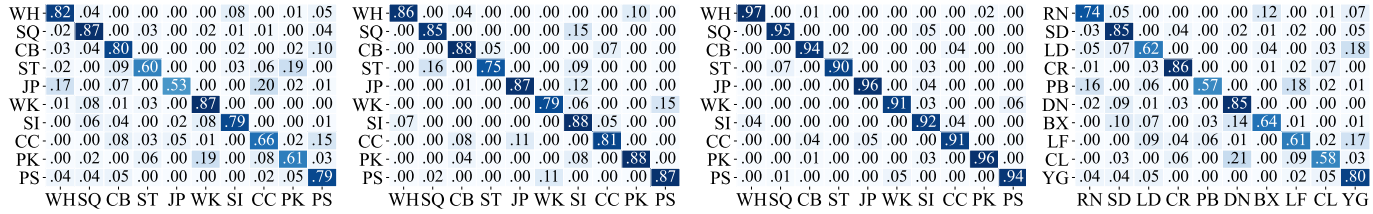
To further understand CKD's superiority, we analyze the relationship between the FM-Fi 2.0 's accuracy and the extent of interdependency information transfer. We employ the mean differences in the correlation matrices of image/RF embeddings to quantify interdependency transfer. By varying τ in the CKD loss, the correlation differences can be adjusted. We select 10 values for τ from 0.1 to 5000. Our findings shown in Fig. 10a demonstrate that the correlation difference negatively impacts FM-Fi 2.0 's accuracy ($\tau = 10$ yields the best performance). This trend further validates FM-Fi 2.0 's principle: preserving the interdependency information among the embedding elements is crucial for HAR. In contrast, the inferior results of alternative approaches (indicated by markers below the curve) can be attributed to their pronounced correlation differences, which correspond to a diminished efficacy in the transfer of interdependency knowledge.

We also conduct 50 independent runs, and perform statistical analysis of the accuracies of various distillation methods, as shown in Fig. 10b. It can be seen that CKD exhibits the highest median accuracy and narrowest interquartile range (IQR). In contrast, CRD, CCKD, and KD demonstrate lower accuracies and larger IQR. Notably, CRD



(a) CLIP. (b) FM-Fi 2.0.

Fig. 11: t-SNE plot of embeddings.



(a) Zero-shot FM-Fi 2.0. (b) 1-shot FM-Fi 2.0. (c) 3-shot FM-Fi 2.0. (d) Zero-shot (10 new classes).

Fig. 12: Confusion matrices of FM-Fi 2.0 in zero-shot and few-shot scenarios.

shows the highest variability in accuracies, which can be attributed to the instability inherent in its learning-based critic model used for similarity assessment. CCKD's approach, which prioritizes alignment of instance distributions between image and RF embeddings without addressing the interdependencies among elements, results in suboptimal performance. Similarly, KD's performance is compromised due to its inability to manage the interdependencies within the embeddings' elements.

5.3 Overall Evaluation of FM-Fi 2.0

To evaluate whether FM-Fi 2.0 has acquired CLIP's embedding capability, we first encode image frame-RF sample pairs from our test set into embedding pairs. These 512-dimensional embeddings are then reduced to 2 dimensions for visualization via t-SNE. From Fig. 11a, it is evident that the embeddings produced by the CLIP encoder are distinct and well-separated, indicating a high degree of discriminability in the embedding space and a robust capacity for image understanding. Fig. 11b shows that FM-Fi 2.0's embeddings are separable and closely aligned with the teacher model's, indicating that FM-Fi 2.0 has effectively captured the teacher model's representational power.

In Fig. 12, we show FM-Fi 2.0's performance across various zero/few-shot scenarios. It can be seen that even in the challenging zero-shot context, FM-Fi 2.0 is capable of basic HAR tasks with a notable 73.4% accuracy. FM-Fi 2.0 also achieves accuracies of 84.4% and 93.6% for 1-shot and 3-shot learning. For the 1-shot case, a significant concentration of samples along the confusion matrix diagonal, indicates that FM-Fi 2.0 maintains robust precision and recall for all categories. This level of performance enables accurate HAR task execution. With three labeled samples, the model's accuracy further improves, with the diagonal average approaching 95%, illustrating a high degree of prediction confidence. Following the few-shot learning phase, we assess FM-Fi 2.0's performance on 10 new activities mentioned in § 5.1. Fig. 12d illustrates that the accuracy on new activities aligns with the results in Fig. 12a, indicating that the few-shot learning module has a minimal impact on zero-shot performance.

We assess the impact of the number of classes on model accuracy by analyzing both zero-shot and 3-shot performance when the number of classes ranges from 5 to 20, as depicted in Fig. 13. The results reveal a decrement in accuracy as the number of classes increases, with zero-shot learning experiencing a more substantial reduction than 3-shot learning. This trend can be attributed to decreased inter-class distinction and increasing semantic overlap as the number of classes increases, undermining the performance of semantic-driven zero-shot methods. In contrast, the metric-based few-shot classification, which utilizes anchors within the embedding space to enhance decision boundaries, exhibits less performance degradation compared its zero-shot counterpart.

Furthermore, we examine the impact of teacher model performance on the effectiveness of the RF student model. As shown in Fig. 14, a stronger teacher model is associated with improved performance of the student model. This results from the teacher's ability to direct the optimization process towards a more efficient trajectory. Notably, the student model's size constraints result in decreased performance gains, indicative of an asymptotic trend. Consequently, ViT-B/32 is chosen as our teacher model backbone due to its superior accuracy of 79.3% on the zero-shot HAR task, with the corresponding student model also evaluated in the same setting, achieving 73.4% accuracy. Compared with the vision modality, the RF modality shows no performance decline, demonstrating that CKD effectively bridges the modality gap within the embedding space.

Next, we investigate the impact of practical factors such as the dataset size for CKD and model complexity on the performance of FM-Fi 2.0. As depicted in Fig. 15a, the zero-shot accuracy increases as the number of CKD data samples increases from 10,000 to 90,000, but stops increasing when the number of CKD data reaches 80,000, stabilizing at approximately 75%. This is close to the 79.3% accuracy

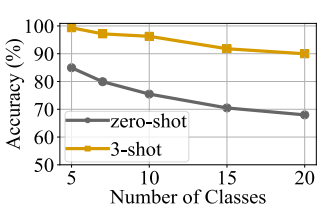


Fig. 13: Impact of the number of classes.

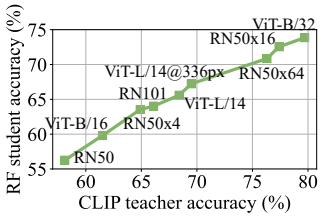
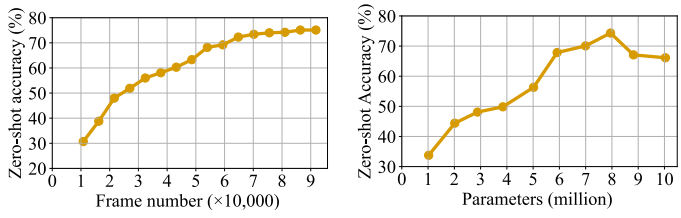
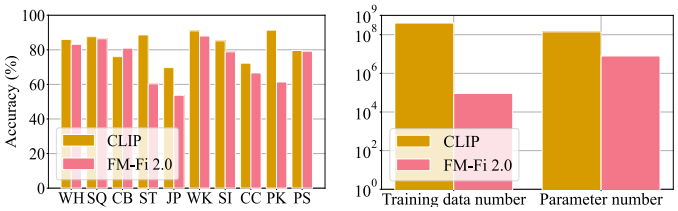


Fig. 14: Student vs. teacher accuracy.



(a) CKD dataset size. (b) Model size.

Fig. 15: Impact of practical factors.



(a) Zero-shot HAR. (b) Resource usage.

Fig. 16: Comparison with FM baseline.

of the teacher model, indicating the efficacy of FM-Fi 2.0's CKD. We then examine the impact of the number of model parameters, as shown in Fig. 15b. It can be observed that FM-Fi 2.0's zero-shot accuracy improves with as the number of parameters increases, reaching a peak of 74% when the number of parameters reaches 8 million. However, expanding the model further to 10 million parameters leads to overfitting and a notable decline in performance due to the increased model complexity.

5.4 Superiority of FM-Fi 2.0

In this section, we compare FM-Fi 2.0 with baselines. To ensure fair comparison, we equip all single-subject HAR baselines with FM-Fi 2.0's instance-wise partitioning module in § 3.2.2, which has proven to be effective in § 5.2.

5.4.1 Comparison with FM

We compare FM-Fi 2.0 with FM by assessing their zero-shot capabilities. As shown in Fig. 16a, the accuracy of FM-Fi 2.0 closely matches that of CLIP across all 10 activity classes, illustrating the overall effectiveness of FM-Fi 2.0. An interesting phenomenon is that for the class CB, the RF-based student model achieves higher accuracy than the FM-based teacher model. The improvement can be attributed to the fact that RF modality might be less susceptible to background image patterns than FM, and our feature association method enables CKD to transfer knowledge without irrelevant signals. Additionally, it should be noted that our collected dataset of 90,000 image-RF pairs is sufficient for CKD. It is also worth mentioning that FM-Fi 2.0's model, with its 8.0 million parameters, is significantly smaller than CLIP's 140 million parameters, as depicted in Fig. 16b. Although the model-to-data size ratio of FM-Fi 2.0 exceeds that of typical LLMs, it still achieves strong performance. This distinction can be attributed to two key factors: first, the knowledge distillation paradigm leverages the fact that the teacher model (i.e., CLIP) is trained on an extensive dataset, allowing it to transfer robust and useful representations to the student model. Second, our smaller dataset, which consists of both unstructured data and rehabilitation activity data, is of high quality and highly relevant to the task at hand. These observations highlight FM-Fi 2.0's ability to deliver competitive performance with considerably less data and a more compact architecture.

5.4.2 Comparison with Zero/Few-shot Baselines

We further compare FM-Fi 2.0 with three few-shot baselines MetaSense, RF-Net, and mmCLIP, where mmCLIP also supports zero-shot HAR. In the few-shot experiment, we employ 10-way-K-shot learning by sampling K instances from each of 10 classes, creating a shared training set for all models. Specifically, we pretrain mmCLIP on the synthetic dataset described in the original paper. Fig. 17 features boxplots that detail the comparative performance of them under 0, 1, 2, and 3-shot settings. In the zero-shot setting, FM-Fi 2.0 achieves 5.8% higher accuracy than mmCLIP with the instance-wise partitioning module. This is because the synthetic dataset used by mmCLIP introduces a simulation-to-reality gap, making it less effective than FM-Fi 2.0, which is directly distilled from and trained on real-world data. In the remaining three scenarios, FM-Fi 2.0 consistently outperforms the three baselines by a significant margin. Although as the number of samples increases, the median accuracy of FM-Fi 2.0 does not rise as quickly as that of the baselines, it still maintains a lead of at least 3.3% Furthermore, the interquartile range (IQR) of FM-Fi 2.0's accuracy is considerably smaller than that of the baselines, indicating less variability across multiple experiments. To better highlight the advantage of FM-Fi 2.0 over the baselines, Table 2 reports their average accuracy under different shot settings. We also explicitly indicate the performance gaps, showing that FM-Fi 2.0 achieves higher average accuracy than the baselines by 2.1% in the 0-shot setting, 0.8% to 20.8% in the 1-shot setting, 1.7% to 25.0% in the 2-shot setting, and 1.1% to 22.7% in the 3-shot setting.

5.4.3 Comparison with Supervised Baselines

We further compare FM-Fi 2.0 with three multi-person RF-HAR baselines: Multi-HAR, PALMAR, and RF-Action, and two point cloud processing baselines, PointNet++ and Point Transformer, as shown in Fig. 18 and Table 3. These models are trained on an expanded dataset (including 50,000 labeled RF samples) without CKD. For ease of comparison, we introduce an additional baseline model termed FM-Fi 2.0*, which utilizes the same RF encoder as FM-Fi 2.0 (with an

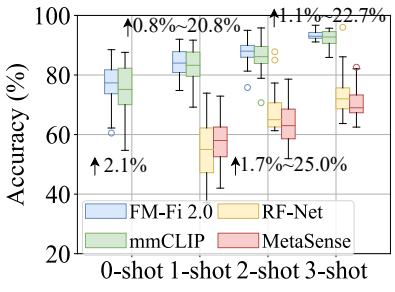


Fig. 17: Few-shot comparisons.

TABLE 2: Few-shot comparison summary.

Acc.	0-shot	1-shot	2-shot	3-shot
FM-Fi 2.0	76.9%	83.9%	88.1%	93.4%
mmCLIP	74.8% (2.1%↓)	83.1% (0.8%↓)	86.4% (1.7%↓)	92.3% (1.1%↓)
RF-Net	/	54.9% (29.0%↓)	67.5% (20.6%↓)	73.6% (19.8%↓)
MetaSense	/	57.8% (26.1%↓)	63.1% (25.0%↓)	70.7% (22.7%↓)

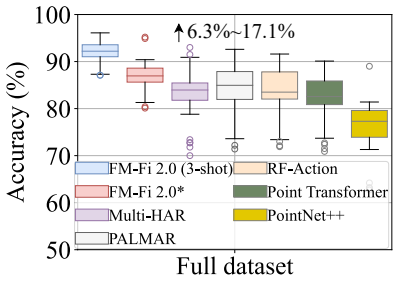
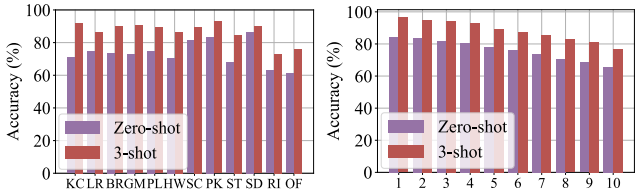


Fig. 18: Supervised comparisons.

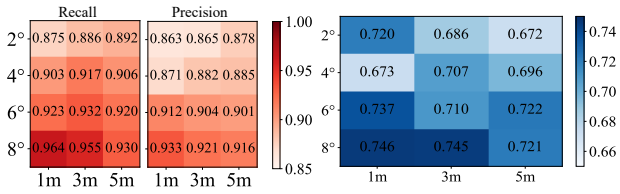
TABLE 3: Supervised comparison summary.

Avg. Acc.	Full dataset
FM-Fi 2.0 (3-shot)	93.4%
FM-Fi 2.0*	87.1% (6.3%↓)
Multi-HAR	83.7% (9.7%↓)
PALMAR	84.3% (9.1%↓)
RF-Action	84.0% (9.4%↓)
Point Transformer	82.5% (10.9%↓)
PointNet++	76.3% (17.1%↓)



(a) Different scenes. (b) Different subject numbers.

Fig. 19: Generalization across diverse settings.



(a) Instance-partitioning capability. (b) HAR accuracy.

Fig. 20: Performance under human overlapping.

ensuing multilayer perception for converting the embedding to classification result). FM-Fi 2.0* is also trained on the same 50,000-sample dataset without CKD. Using only 0.1% of the labeled data compared to the other three models, 3-shot FM-Fi 2.0 not only demonstrates superior accuracy but also greater stability in performance. Specifically, FM-Fi 2.0 achieves 6.3% to 17.1% higher average accuracy than the baselines. These results highlight the efficacy of CKD in learning robust representations while significantly decreasing the dependency on annotated RF data. Furthermore, among the three fully supervised models specifically designed for point clouds, FM-Fi 2.0* exhibits notably better performance than the other two models, PointNet++ and Point Transformer. The superior performance of FM-Fi 2.0* is due to its RF encoder which effectively integrates point cloud coordinates with Doppler features and signal intensity, thus utilizing the complete range of information available in RF data. For the three baselines specifically designed for multi-person RF-HAR, although they are trained on significantly larger labeled RF datasets, FM-Fi 2.0 still achieves notably higher accuracy and stability. This is because FM-Fi 2.0 learns rich prior knowledge from the vision FM, which is effectively transferred to zero/few-shot HAR tasks. As a result, with only limited labeled RF data, FM-Fi 2.0 outperforms fully supervised models trained solely on small RF datasets without such priors. This paradigm is not explored by previous methods and demonstrates broader applicability to diverse downstream tasks.

5.5 Generalization Capability

5.5.1 *Cross-subject and Cross-environment Evaluation.*

To evaluate the generalization capabilities of FM-Fi 2.0, we firstly conduct tests on the 10 subjects ($S1 - S10$) and 10 environments mentioned in § 4.1. Specifically, we adopt a leave-one-out strategy for 3-shot testing, where we train on data from 9 environments or subjects and test on the remaining one; for zero-shot testing, we directly conduct tests without additional training. We conduct both zero-shot and 3-shot tests in 100 settings (10 environments $\times$ 10 subjects) and the results shown in Fig. 19 are obtained by averaging across either environments or subjects.

Overall, Fig. 19a shows that performance tends to be better in outdoor scenes due to factors such as better lighting, open space, less background features, and reduced occlusion. However, street scenes yield poorer results because of the interference from rapidly moving background objects such as cars and pedestrians, which can disrupt RF signals. In contrast, for primary RF-based HAR scenarios, especially in domestic settings, FM-Fi 2.0 maintains performance levels consistent with previous tests, demonstrating exceptional capabilities. Moreover, as shown in Fig. 19b, despite a decrease in accuracy as the number of people in the scene increases, FM-Fi 2.0 still maintains robust generalization capabilities. Even with up to 10 individuals present, it achieves a zero-shot accuracy of at least 65.8% and a three-shot accuracy of at least 77.6%. While extreme noise scenarios, such as at road intersections, induce performance degradation from multipath interference, FM-Fi 2.0 sustains an average zero-shot accuracy of 63.6%. Similarly, in the high-interference environment of outdoor fitness areas, reflections from exercise equipment coupled with human motion reduce accuracy; FM-Fi 2.0 nonetheless achieves an average zero-shot accuracy of 61.3%. These outcomes underscore FM-Fi 2.0's resilience, demonstrating robust performance even under such adverse conditions.

5.5.2 *Robustness to Human Overlap Scenarios.*

We then evaluate the performance of FM-Fi 2.0 under varying degrees of human overlap in two-person scenarios, focusing on two aspects: instance-wise partitioning capability and HAR accuracy. To create different levels of spatial overlap, we vary the angle between the two subjects (2°, 4°, 6°, 8°) and their distances (1, 3, and 5 m). The average distance from the subjects to the radar is 5 m. As shown in Fig. 20a, even in highly overlapping scenarios, where the subjects are only 1 m apart with an angle of 2°, FM-Fi 2.0 maintains a recall of 87.2% and a precision of 86.3%. When the angle increases to 6°, recall remains above 92.0%, and precision stays above 90.1%. Regarding HAR accuracy, Fig. 20b shows that even under distant overlap conditions at 5m away, FM-Fi 2.0 achieves a 10-class zero-shot accuracy of over 67.2% for each subject. Across the remaining overlapping settings, the accuracy consistently remains above 67.3%. These results demonstrate the robustness of FM-Fi 2.0 in handling complex and realistic human overlap scenarios.

5.6 Hyper-parameter Searching

5.6.1 *Feature Elimination Threshold*

In this work, we determine the threshold as λ times the mean value of the entire similarity map, which establishes the lower bound score for pixels exempt from blur transformation. On one hand, a small λ preserves the original image content, but fails to efficiently eliminate background noise. On the other hand, a high λ value risks removing critical image features, depriving the model of meaningful input and thereby reducing the discriminability of the generated cross-modal supervision signal. To search for the optimal value of λ, we evaluate the zero-shot performance of FM-Fi 2.0 at different λ values from 0.4 to 1.6 One may readily observe in Fig. 21a that as λ initially increases, FM-Fi 2.0 reaches the best performance at the optimal threshold $\lambda = 1.2$. Any λ greater than that may cause the blur mask to

erode the some of the human figures, adversely impacting HAR performance. Consequently, as λ surpasses 1.2, there is a significant decline in accuracy from 79.2% to 48.6%.

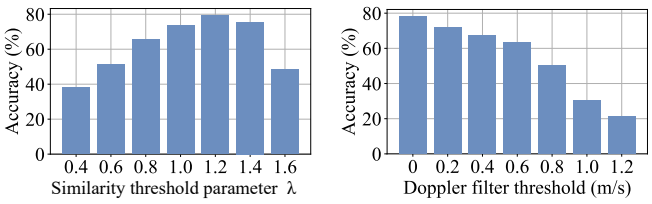
We further study the impact of velocity thresholds in the RF modality. Instead of only removing the zero-velocity component as in § 3.1.2, we set the velocity filtering thresholds from 0 to 1.2 m/s, and show the relationship between the model's accuracy and the threshold in Fig. 21b. It is observed that the model performs the best when the Doppler threshold is set to 0, which corresponds to the removal of static background. Increasing the Doppler threshold may inadvertently filter out some moving background clutter; however, it might also eliminate information pertinent to human activities, leading to a decline in model performance. When the Doppler threshold reaches 0.8 m/s, a significant portion of human activity information is lost, resulting in poor model performance. The experiment suggests that a threshold of 0, which preserves all information of moving objects while excluding static background features, optimally supports the subsequent instance-wise partitioning module in performing point selection.

5.6.2 *Weight of Label Text in Few-shot Learning*

We evaluate the impact of varying weights of label text γ on FM-Fi 2.0's performance across 1-shot, 2-shot, and 3-shot learning scenarios. Initial assessments are conducted with integer values of γ in the range 0 to 10, with results depicted in Fig. 22a. We observe that for small values, accuracy across all scenarios increased with γ, suggesting effective semantic information extraction from the RF modality by FM-Fi 2.0. As γ increases, performance across the three scenarios tends to converge due to the text embedding becoming the dominant factor. Such convergence results in performance degradation, approaching zero-shot levels as γ further increases. We aim to identify the best performance point $\gamma = 4$ to $\gamma = 6$. As Fig. 22b demonstrates, the peak performance is obtained at $\gamma = 5$, which we use as the weight of text label in few-shot learning.

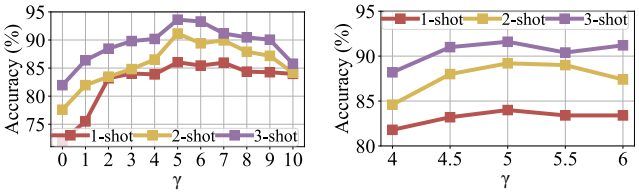
6 Related Work and Discussion

Though RF-HAR literature covers enhancing generalizability [12], [13], [46], [52], [53], [54], [55], improving the efficient utilization of scarce labeled data [56], [57], and refining model architectures [13], [58], [59], prominent RF-HAR proposals have prioritized studies on generalizability. In particular, Widar3.0 [52] introduces a domain-independent and signal-level feature, termed BVP, to enable generalizability. Another study [53] applies adversarial domain adaptation techniques [60], [61] to generalize across varying scenarios. RF-Net [12] adopts metric-based meta-learning achieve fast adaptation of its base networks in diverse environments.



(a) Image modality. (b) RF Modality.

Fig. 21: Impact of the thresholds.



(a) Preliminary search. (b) Detailed search.

Fig. 22: Impact of the weight of label text.

While prior multi-person RF-HAR methods [47], [48], [49] have advanced the field, FM-Fi 2.0 offers a distinct advantage. It uniquely harnesses the rich prior knowledge from vision foundation models, employing cross-modal CKD to transfer this knowledge to the RF modality. This pioneering approach enables robust zero-shot and few-shot HAR, a paradigm previously unexplored, and demonstrates broader applicability to diverse downstream tasks.

The emergence of FMs has brought new potentials in RF sensing in general, catering the need for more models capable of capturing rich information. In current researches, FM-Fi [36] employs cross-modal contrastive knowledge distillation (CKD) to translate the prior knowledge from vision-based FMs to enhance RF-based HAR systems. Meanwhile, mmCLIP [45] aligns mmWave signals with the textual space through cross-modality signal synthesis and activity attribute decomposition. Compared to these single-person HAR system, FM-Fi 2.0 explores the more commonly seen multi-person scenarios, and introduces several key modifications in system design compared to FM-Fi. First, to maintain instance-wise feature correspondence across modalities, it associates intra- and cross-modality features by incorporating an additional viewpoint and corresponding algorithms in the vision modality, as well as modeling spatial relationships between modalities. Second, to generate instance-wise embeddings, FM-Fi 2.0 includes a module that explicitly aggregates RF features at the instance level. Finally, the CKD negative sample pool is expanded to include different instances from the same frame. In the future, we would expect FM-Fi 2.0 to be able to support other sensing tasks including gesture detection [62], gait recognition [63], and even vibration monitoring [64], [65], by modifying the target of interest; we plan to explore FM-Fi 2.0's potential beyond HAR in future work.

7 Conclusion

Taking [66] a significant stride in advancing HAR, we have introduced FM-Fi 2.0, which harnesses the interpretative power of FMs to facilitate cross-modal RF-HAR. By employing instance-wise feature association and CKD, the innovative RF encoder in FM-Fi 2.0 effectively assimilates the semantic embedding derived from FMs. This enables precise mapping of RF data for efficient zero/few-shot HAR applications, addressing the critical challenge of data scarcity in RF-HAR. Our thorough experiment analysis across diverse and complex scenarios confirms FM-Fi 2.0's superiority over conventional baselines. This research not only demonstrates the effectiveness of our approach but also lays the groundwork for further advancements in RF-HAR, while aiming for broader RF sensing tasks in practical settings.

ACKNOWLEDGMENTS

The study is supported by Shenzhen Science and Technology Program (No. 20231120215201001) and the research start-up grant from the Southern University of Science and Technology, for which Tianyue Zheng expresses sincere gratitude. We are also grateful to the anonymous reviewers for their constructive comments. As a side note, one of the authors, Yanbing Yang, receives funding from National Natural Science Foundation of China (62272329).

REFERENCES

[1] J. Hao, A. Bouzouane, and S. Gaboury, "Recognizing Multi-Resident Activities in Non-Intrusive Sensor-Based Smart Homes by Formal Concept Analysis," *Neurocomputing*, vol. 318, pp. 75–89, 2018.

[2] L. M. Dang, K. Min, H. Wang, M. J. Piran, C. H. Lee, and H. Moon, "Sensor-Based and Vision-Based Human Activity Recognition: A Comprehensive Survey," *Pattern Recognition*, vol. 108, p. 107561, 2020.

[3] Z. Chi, Y. Yao, T. Xie, X. Liu, Z. Huang, W. Wang, and T. Zhu, "EAR: Exploiting Uncontrollable Ambient RF Signals in Heterogeneous Networks for Gesture Recognition," in *Proc. of the 16th ACM SenSys*, 2018, pp. 237–249.

[4] K. Niu, F. Zhang, J. Xiong, X. Li, E. Yi, and D. Zhang, "Boosting Fine-grained Activity Sensing by Embracing Wireless Multipath Effects," in *Proc. of the 14th ACM CoNEXT*, 2018, pp. 139–151.

[5] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep Learning Models for Real-Time Human Activity Recognition with Smartphones," *Mobile Networks and Applications*, vol. 25, pp. 743–755, 2020.

[6] A.-K. Seifert, M. G. Amin, and A. M. Zoubir, "Toward Unobtrusive In-home Gait Analysis Based on Radar Micro-Doppler Signatures," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 9, pp. 2629–2640, 2019.

[7] D. Chen, M. Wang, C. He, Q. Luo, Y. Iravantchi, A. Sample, K. G. Shin, and X. Wang, "MagX: Wearable, Untethered Hands Tracking with Passive Magnets," in *Proc. of the 27th ACM MobiCom*, 2021, pp. 269–282.

[8] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal Human Action Recognition in Assistive Human-Robot Interaction," in *Proc. of IEEE ICASSP*. IEEE, 2016, pp. 2702–2706.

[9] C. Bi, G. Xing, T. Hao, J. Huh-Yoo, W. Peng, M. Ma, and X. Chang, "FamilyLog: Monitoring Family Mealtime Activities by Mobile Devices," *IEEE Transactions on Mobile Computing*, vol. 19, no. 8, pp. 1818–1830, 2019.

[10] H. Truong, S. Zhang, U. Muncuk, P. Nguyen, N. Bui, A. Nguyen, Q. Lv, K. Chowdhury, T. Dinh, and T. Vu, "Capband: Battery-Free Successive Capacitance Sensing Wristband for Hand Gesture Recognition," in *Proc. of the 16th ACM SenSys*, 2018, pp. 54–67.

[11] A. Ferlini, D. Ma, R. Harle, and C. Mascolo, "EarGate: Gait-based User Identification with In-ear Microphones," in *Proc. of the 27th ACM MobiCom*, 2021, pp. 337–349.

[12] S. Ding, Z. Chen, T. Zheng, and J. Luo, "RF-Net: A Unified Meta-Learning Framework for RF-enabled One-Shot Human Activity Recognition," in *Proc. of the 18th ACM SenSys*, 2020, pp. 517–530.

[13] Z. Chen, C. Cai, T. Zheng, J. Luo, J. Xiong, and X. Wang, "RF-based Human Activity Recognition using Signal Adapted Convolutional Neural Network," *IEEE Transactions on Mobile Computing*, vol. 22, no. 1, pp. 487–499, 2021.

[14] S. Palipana, D. Salami, L. A. Leiva, and S. Sigg, "Pantomime: Mid-Air Gesture Recognition with Sparse Millimeter-Wave Radar Point Clouds," *Proc. of ACM UbiComp*, vol. 5, no. 1, pp. 1–27, 2021.

[15] D. Salami, R. Hasibi, S. Palipana, P. Popovski, T. Michoel, and S. Sigg, "Tesla-Rapture: A Lightweight Gesture Recognition System from mmWave Radar Sparse Point Clouds," *IEEE Transactions on Mobile Computing*, 2022.

[16] J. Hu, T. Zheng, Z. Chen, H. Wang, and J. Luo, "MUSE-Fi: Contactless MUti-person SEnsing Exploiting Near-field Wi-Fi Channel Variation," in *Proc. of the 29th ACM MobiCom*, 2023, pp. 1–15.

[17] T. Zheng, Z. Chen, S. Zhang, C. Cai, and J. Luo, "MoRe-Fi: Motion-robust and Fine-grained Respiration Monitoring via Deep-Learning UWB Radar," in *Proc. of the 19th ACM SenSys*, 2021, pp. 111–124.

[18] K. Bansal, K. Rungta, S. Zhu, and D. Bharadia, "Pointillism: Accurate 3D Bounding Box Estimation with Multi-Radars," in *Proc. of the 18th ACM SenSys*, 2020, pp. 340–353.

[19] T. Boroushaki, J. Leng, I. Clester, A. Rodriguez, and F. Adib, "Robotic Grasping of Fully-Occluded Objects using RF Perception," in *Proc. of IEEE ICRA*. IEEE, 2021, pp. 923–929.

[20] T. Zheng, Z. Chen, J. Luo, L. Ke, C. Zhao, and Y. Yang, "SiWa: See into Walls via Deep UWB Radar," in *Proc. of the 27th ACM MobiCom*, 2021, pp. 323–336.

[21] A. D. Singh, Y. Ba, A. Sarker, H. Zhang, A. Kadambi, S. Soatto, M. Srivastava, and A. Wong, "Depth Estimation From Camera Image and mmWave Radar Point Cloud," in *Proc. of IEEE/CVF CVPR*, 2023, pp. 9275–9285.

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language Models are Few-Shot Learners," *Proc. of NeurIPS*, vol. 33, pp. 1877–1901, 2020.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-Shot Text-to-Image Generation," in *Proc. of ICML*. PMLR, 2021, pp. 8821–8831.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning Transferable Visual Models from Natural Language Supervision," in *Proc. of ICML*. PMLR, 2021, pp. 8748–8763.

[26] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong *et al.*, "Robust fine-tuning of zero-shot models," in *Proc. of IEEE/CVF CVPR*, 2022, pp. 7959–7971.

[27] S. Esmaeilpour, B. Liu, E. Robertson, and L. Shu, "Zero-Shot Out-of-Distribution Detection Based on the Pre-trained Model CLIP," in *Proc. of AAAI*, vol. 36, no. 6, 2022, pp. 6568–6576.

[28] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen *et al.*, "Simple Open-Vocabulary Object Detection with Vision Transformers," in *Proc. of ECCV*. Springer, 2022, pp. 728–755.

[29] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "EVA: Exploring the Limits of Masked Visual Representation Learning at Scale," in *Proc. of IEEE/CVF CVPR*, 2023, pp. 19 358–19 369.

[30] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[31] S. Ren, L. Li, X. Ren, G. Zhao, and X. Sun, "Delving into the Openness of CLIP," in *Proc. of ACL*, 2023.

[32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," *arXiv preprint arXiv:1503.02531*, 2015.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Proc. of NeurIPS*, vol. 30, 2017.

[34] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proc. of ICML*. PMLR, 2020, pp. 1597–1607.

[35] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proc. of IEEE/CVF CVPR*, 2020, pp. 9729–9738.

[36] Y. Weng, G. Wu, T. Zheng, Y. Yang, and J. Luo, "Large Model for Small Data: Foundation Model for Cross-Modal RF Human Activity Recognition," in *Proc. of the 22nd ACM SenSys*, 2024, pp. 436–449.

[37] B. Peng, X. Jin, J. Liu, D. Li, Y. Wu, Y. Liu, S. Zhou, and Z. Zhang, "Correlation Congruence for Knowledge Distillation," in *Proc. of IEEE/CVF ICCV*, 2019, pp. 5007–5016.

[38] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *Proc. of IEEE/CVF CVPR*, 2017, pp. 652–660.

[39] L. Fan, T. Li, Y. Yuan, and D. Katabi, "In-Home Daily-Life Captioning Using Radio Signals," in *Proc. of the 16th ECCV*. Springer, 2020, pp. 105–123.

[40] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," *arXiv preprint arXiv:1807.03748*, 2018.

[41] Y. Tian, D. Krishnan, and P. Isola, "Contrastive Representation Distillation," in *Proc. of ICLR*, 2019.

[42] Texas Instruments, "IWR1443BOOST," https://www.ti.com/tool/IWR1443BOOST, 2020, accessed: 2020-09-29.

[43] H. O. Park, A. A. Dibazar, and T. W. Berger, "Cadence Analysis of Temporal Gait Patterns for Seismic Discrimination Between Human and Quadruped Footsteps," in *Proc. of IEEE ICASSP*. IEEE, 2009, pp. 1749–1752.

[44] Microsoft, "Kinect Sensor," https://developer.microsoft.com/en-us/windows/kinect/, 2020, accessed: 2020-09-29.

[45] Q. Cao, H. Xue, T. Liu, X. Wang, H. Wang, X. Zhang, and L. Su, "mmCLIP: Boosting mmWave-based Zero-shot HAR via Signal-Text Alignment," in *Proc. of the 22nd ACM SenSys*, 2024, pp. 184–197.

[46] T. Gong, Y. Kim, J. Shin, and S.-J. Lee, "MetaSense: Few-Shot Adaptation to Untrained Conditions in Deep Mobile Sensing," in *Proc. of the 17th ACM SenSys*, 2019, pp. 110–123.

[47] M. A. U. Alam, M. M. Rahman, and J. Q. Widberg, "PALMAR: Towards Adaptive Multi-inhabitant Activity Recognition in Point-Cloud Technology," in *IEEE INFOCOM 2021-IEEE conference on computer communications*. IEEE, 2021, pp. 1–10.

[48] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi, "Making the Invisible Visible: Action Recognition Through Walls and Occlusions," in *Proc. of the IEEE/CVF ICCV*, 2019, pp. 872–881.

[49] X. Zeng, Y. Shi, and A. Zhou, "Multi-HAR: Human Activity Recognition in Multi-person Scenes Based on mmWave Sensing," in *2022 IEEE 8th International Conference on Computer and Communications (ICCC)*. IEEE, 2022, pp. 1789–1793.

[50] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," *Proc. of NeurIPS*, vol. 30, 2017.

[51] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point Transformer," in *Proc. of IEEE/CVF ICCV*, 2021, pp. 16 259–16 268.

[52] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi," in *Proc. of the 17th ACM MobiSys*, 2019, pp. 313—325.

[53] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas *et al.*, "Towards Environment Independent Device Free Human Activity Recognition," in *Proc. of the 24th ACM MobiCom*, 2018, pp. 289–304.

[54] R. Gao, W. Li, Y. Xie, E. Yi, L. Wang, D. Wu, and D. Zhang, "Towards Robust Gesture Recognition by Characterizing the Sensing Quality of WiFi Signals," *Proc. of ACM UbiComp*, vol. 6, no. 1, pp. 1–26, 2022.

[55] F. Meneghello, D. Garlisi, N. Dal Fabbro, I. Tinnirello, and M. Rossi, "ShARP: Environment and Person Independent Activity Recognition with Commodity IEEE 802.11 Access Points," *IEEE Transactions on Mobile Computing*, 2022.

[56] X. Ouyang, Z. Xie, J. Zhou, J. Huang, and G. Xing, "ClusterFL: A Similarity-aware Federated Learning System for Human Activity Recognition," in *Proc. of the 19th ACM MobiSys*, 2021, pp. 54–66.

[57] X. Li, Y. He, F. Fioranelli, and X. Jing, "Semisupervised Human Activity Recognition with Radar Micro-Doppler Signatures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[58] M. Chakraborty, H. C. Kumawat, S. V. Dhavale *et al.*, "DIAT-RadHARNet: A Lightweight DCNN for Radar based Classification of Human Suspicious Activities," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[59] X. Li, S. Chen, S. Zhang, L. Hou, Y. Zhu, and Z. Xiao, "Human Activity Recognition Using IR-UWB Radar: A Lightweight Transformer Approach," *IEEE Geoscience and Remote Sensing Letters*, 2023.

[60] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in *Proc. of ICML*. PMLR, 2015, pp. 1180–1189.

[61] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky, "Domain-adversarial Training of Neural Networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.

[62] S. Zhang, T. Zheng, Z. Chen, J. Hu, A. Khamis, J. Liu, and J. Luo, "OCHID-Fi: Occlusion-Robust Hand Pose Estimation in 3D via RF-Vision," in *Proc. of IEEE/CVF ICCV*, 2023, pp. 15 112–15 121.

[63] D. Cao, R. Liu, H. Li, S. Wang, W. Jiang, and C. X. Lu, "Cross Vision-RF Gait Re-identification with Low-cost RGB-D Cameras and mmWave Radars," *Proc. of ACM UbiComp*, vol. 6, no. 3, pp. 1–25, 2022.

[64] Z. Chen, T. Zheng, and J. Luo, "Octopus: a practical and versatile wideband MIMO sensing platform," in *Proc. of the 27th ACM MobiCom*, 2021, pp. 601–614.

[65] Z. Chen, T. Zheng, C. Cai, and J. Luo, "MoVi-Fi: Motion-robust Vital Signs Waveform Recovery via Deep Interpreted RF Sensing," in *Proc. of the 27th ACM MobiCom*, 2021, pp. 392–405.

[66] J. M. Tarnawski, D. Narayanan, and A. Phanishayee, "Piper: Multidimensional Planner for DNN Parallelization," *Proc. of NeurIPS*, vol. 34, pp. 24 829–24 840, 2021.

Yuxuan Weng is a research assistant at the Southern University of Science and Technology. He received his Bachelor's degree from Sun Yat-sen University and his Master's degree from the Hong Kong University of Science and Technology. His research interests include mobile computing, RF sensing, multimodal sensing, and machine learning.

Tianyue Zheng is an Assitant Professor and Ph.D. Supervisor at the Southern University of Science and Technology (SUSTech). He received his Ph.D. degree from Nanyang Technological University, Singapore, in 2023. His research interests focus on mobile computing, RF sensing, and multimodal sensing. He has published over 30 papers in top venues. He serves program committee members for several international conferences and a reviewer for multiple journals.

Yanbing Yang received the BE and ME degrees from the University of Electronic Science and Technology of China, China, and the PhD degree in computer science and engineering from Nanyang Technological University, Singapore. He is currently an associate professor at the College of Computer Science, Sichuan University, China. His research interests include IoTs, visible light communication, and visible light sensing, as well as their applications..

Jun Luo received his BS and MS degrees in Electrical Engineering from Tsinghua University, China, and the Ph.D. degree in Computer Science from EPFL (Swiss Federal Institute of Technology in Lausanne), Lausanne, Switzerland. From 2006 to 2008, he has worked as a postdoctoral research fellow in the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada. In 2008, he joined the faculty of the School Of Computer Science and Engineering, Nanyang Technological University in Singapore, where he is currently an Associate Professor. His research interests include mobile and pervasive computing, wireless networking, machine learning and computer vision, as well as applied operations research. More information can be found at http://www.ntu.edu.sg/home/junluo.