Large Model for Small Data: Foundation Model for Cross-Modal RF Human Activity Recognition

Yuxuan Weng[1], Guoquan Wu[1], Tianyue Zheng[1,2]✉, Yanbing Yang[3], Jun Luo[4]
[1] Department of Computer Science and Engineering, Southern University of Science and Technology, China
[2] Smart City Center, Research Institute of Trustworthy Autonomous Systems, Southern University of Science and Technology, China
[3] College of Computer Science, Sichuan University, China
[4] College of Computing and Data Science, Nanyang Technological University, Singapore
Email: {wengyx, wugq2024, zhengty}@sustech.edu.cn, yangyanbing@scu.edu.cn, junluo@ntu.edu.sg

ABSTRACT

Radio-Frequency (RF)-based Human Activity Recognition (HAR) rises as a promising solution for applications unamenable to techniques requiring computer visions. 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, an innovative cross-modal framework engineered to translate the knowledge of vision-based FMs for enhancing RF-based HAR systems. FM-Fi involves a novel cross-modal *contrastive* knowledge distillation mechanism, enabling an RF encoder to inherit the interpretative power of FMs for achieving zero-shot learning. It also employs the intrinsic capabilities of FM and RF to remove extraneous features for better alignment between the two modalities. 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 rivals the effectiveness of vision-based methodologies, and the evaluation results provide empirical validation of FM-Fi's generalizability across various environments.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Human activity recognition, foundation model, RF sensing.

ACM Reference Format:
Y. Weng, G. Wu, T. Zheng, Y. Yang, and J. Luo. 2024. Large Model for Small Data: Foundation Model for Cross-Modal RF Human Activity Recognition. In *The 22nd ACM Conference on Embedded Networked Sensor Systems (SENSYS '24), November 4–7, 2024, Hangzhou, China.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3666025.3699349

✉ Corresponding author.

SENSYS '24, November 4–7, 2024, Hangzhou, China

ACM ISBN 979-8-4007-0697-4/24/11.
https://doi.org/10.1145/3666025.3699349

1 INTRODUCTION

With rapid developments [16, 27], Human Activity Recognition (HAR) gains significant interest in smart homes [15, 42], digital healthcare [55, 64], and human-computer interaction [9, 53]. In practice, HAR tasks can be either contact-based [3, 22, 62] or contact-free [11, 17]; the latter offers the advantage of not imposing the additional burden and discomfort of wearing devices. Among all sensing modalities for contact-free HAR, Radio-Frequency (RF) sensing [31, 44, 54, 75] 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 [2, 4, 74] 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 [58]: Unlike image data, human annotators find it impossible to intuitively recognize activities from RF data, 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 still lack guaranteed data reliability, largely confining the adoption of RF sensing in HAR tasks.

The recent advent of Foundation Models (FMs) [5, 18, 51] 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 [49] 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 [19, 49, 66], object detection [21, 41], and image generation [50, 51]. 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 [52]. Now the question becomes: *can FMs be harnessed to interpret RF-HAR data?* A valid answer to this question is essential for advancing RF-HAR towards practical adoption.

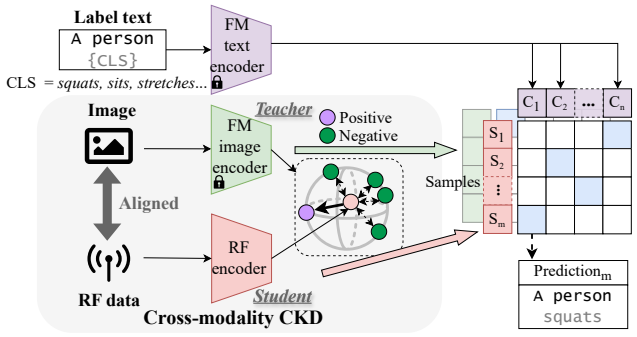


Figure 1: Overview of FM-Fi.

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) [18] and natural language processing (NLP) [5, 49], thus limiting their direct applicability to RF-HAR. Although cross-modal knowledge distillation (KD) [29] 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 inherent feature discrepancies. Specifically, the image modality include extraneous background details that obscure HAR-relevant information, whereas the RF modality often features irrevelant static backgrounds. This misalignment significantly challenges effective modality integration. 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, a cross-modal framework that distills the knowledge from FMs to the RF modality, as illustrated in Figure 1. First, 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. Second, FM-Fi employs the intrinsic capabilities of FM and RF to remove extraneous background features, thus enabling better alignment between the two modalities. In particular, the semantic space of FM is leveraged to score vision features, and the physical properties of the RF modality are explored to filter static and dynamic backgrounds. Finally, FM-Fi harnesses a minimal set of annotated data to fine-tune its model via metric-based few-shot learning, enhancing already achieved zero-shot classification to fit specific HAR tasks. The synergy of these three 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:

- To the best of our knowledge, FM-Fi is the first cross-modal distillation system specifically designed from vision FMs to RF model for zero/few-shot HAR tasks.
- We develop a CKD mechanism to accommodate FM's intrinsic embedding dependencies, enabling successful knowledge transfer from FMs to RF modality.
- We design extraneous feature elimination methods tailored to image and RF modalities, achieving a feature-aligned vision-RF dataset.
- 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.
- We construct an FM-Fi prototype and evaluate it with extensive experiments: the promising results confirm that FM-Fi enables high-performance RF-HAR for both zero-shot and few-shot HAR tasks.

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

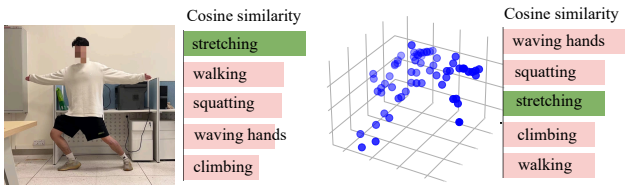
2 BACKGROUND AND MOTIVATIONS

In this section, we introduce the background of FM for HAR and the motivations of FM-Fi'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 [51] for image generation), NLP (e.g., GPT [5] for chatbot), and multimodal applications (e.g., CLIP [49] for image semantics understanding), have demonstrated their extensive capabilities. The enhanced image understanding in FMs is facilitated by the adoption of transformer [18, 63] architecture as encoders, which enable the derivation of complex representations. Additionally, contrastive learning [10, 28] 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. For instance, when analyzing an image of a person stretching, as depicted in Figure 2a, the CLIP model can accurately



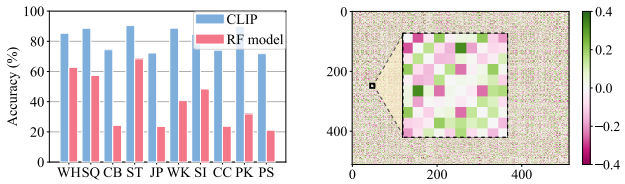
(a) On image modality. **(b) On RF modality.**

Figure 2: Performance of FM for HAR.

assess the similarity between the embedding of the image and the text, thereby achieving zero-shot HAR. However, the application of FMs, initially trained on vision-text data, to RF data introduces considerable challenges. This difficulty arises from the inherent abstractness of RF signals. As illustrated in Figure 2b, directly applying the CLIP model has falsely identified the activity *stretching* captured by a mmWave (in the form of point cloud akin to image pixels) as *waving hands*. This limitation underscores the necessity of novel methods for RF data processing to extend the applicability of FMs beyond visual data.

2.2 Why Conventional KD Fails for FMs?

One viable approach for utilizing FMs for HAR is KD; it involves transferring the knowledge from an FM to RF model by aligning their output embeddings, where we 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 [29]. One may readily observe that a naive application of KD on FMs leads to inferior performance, as depicted in Figure 3a. Specifically, for 10-class classification, the CLIP-trained RF encoder achieves an average accuracy slightly above 40%, whereas that achieved by the baseline CLIP exceeds 80%.



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

Figure 3: Conventional KD performance.

To understand KD's ineffectiveness, we explore the interdependencies among elements of the output embeddings. We compute the correlation matrices, $\mathbf{R}^{IM}$ for the FM (processing the image modality) and $\mathbf{R}^{RF}$ for the RF model, respectively. By subtracting $\mathbf{R}^{RF}$ from $\mathbf{R}^{IM}$, we obtain a difference matrix as shown in Figure 3b. 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 [46]. 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 Figure 7b in § 3.2 for a better correlation matrix difference that better captures the interdependencies among the elements in the embeddings.

2.3 Effect of Extraneous Feature

To successfully transfer knowledge from the image to RF modality, alignment between the two modalities is crucial. Both modalities, however, contain extraneous features; for instance, images may include irrelevant lighting and background objects, while RF data may be influenced by static backgrounds. As demonstrated in Figure 4, minor variations in background features, such as lighting and curtains (as illustrated in the upper row), significantly affect the embeddings (in the lower row), leading to instability. This instability is presumed to affect the RF modality as well. Furthermore, there is no straightforward one-to-one correspondence between the embeddings of image and RF modalities due to their not sharing an identical set of features. Consequently, these extraneous features hinder the knowledge transfer from image to RF modality, necessitating the development of a method to efficiently eliminate such features.

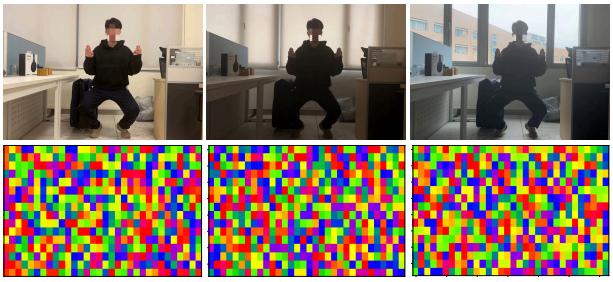


Figure 4: Minor background variations significantly alter the output embeddings of FM.

3 SYSTEM DESIGN

Based on the discussions in §2, we hereby present FM-Fi with five innovative components: i) an RF encoder that helps encoding information from the RF point clouds, ii) a cross-modal CKD framework for transferring semantic representations from visual feature maps to RF-based models, iii) a multimodal data alignment module that eliminates extraneous features, thereby improving HAR knowledge integration across modalities, iv) a zero-shot HAR mechanism relying on learned associations between the semantics of both RF and (FM's) text modalities, and v) a metric-based few-shot learning network enabling FM-Fi 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 Figure 5.

3.1 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. Drawing upon these rich features, we develop a robust RF encoder for extracting meaningful RF embeddings. 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.

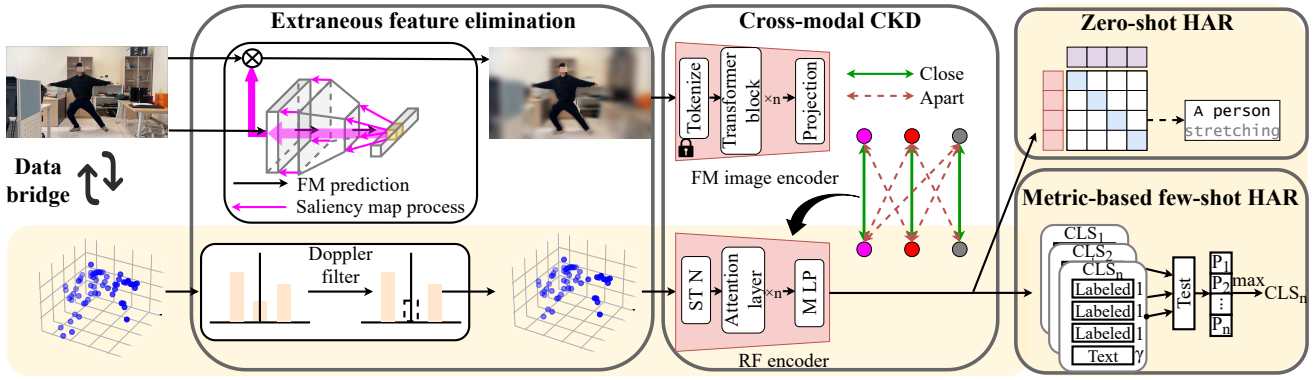


Figure 5: Overall design of FM-Fi.

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 [47] to accommodate the properties of mmWave data, as shown in Figure 6. FM-Fi'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. This module is tasked with selectively weighting individual points within the global context, thereby effectively filtering out those irrelevant to HAR. The design details of this module are further discussed in § 3.3.2.

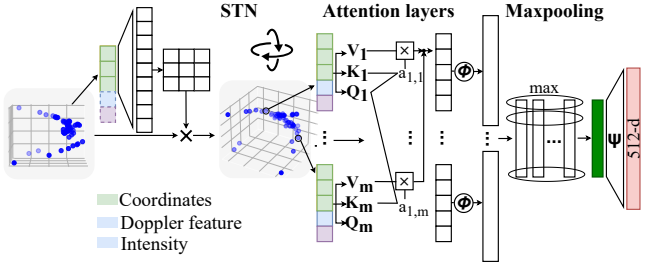


Figure 6: RF encoder for cross-modal distillation.

We then pass the enriched feature vectors through a multilayer perceptron (MLP) ϕ for dimensionality expansion, after which the updated coordinates are processed by a maxpooling module. This module 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 process of point cloud processing can be expressed as follows:

$$\mathbf{E}^{\mathrm{RF}} = \psi\left(\underset{i=1\ldots N}{\text{maxpooling}}\left(\phi\left(\mathcal{A}\left(\mathcal{T}(\mathbf{X}_{\mathrm{RF}_i}) \cdot \mathbf{X}_{\mathrm{RF}_i}\right)\right)\right)\right). \quad (1)$$

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

3.2 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 [20]. 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 2.2, 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 process 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 $(\mathbf{X}_i^{\mathrm{IM}}, \mathbf{X}_i^{\mathrm{RF}})$, 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 $\mathbf{E}^{\mathrm{IM}}$ and $\mathbf{E}^{\mathrm{RF}}$. 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 § 2.2. Instead, we employ the mutual information between modalities as the starting point for deriving contrastive knowledge distillation (CKD) method. This method is better at handling interdependencies within the embedding elements, which are 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 $\mathbf{E}^{\text{IM}}$ and $\mathbf{E}^{\text{RF}}$. The mutual information is defined as: $MI(\mathbf{E}^{\text{IM}};\mathbf{E}^{\text{RF}}) = \mathbb{E}_{p(\mathbf{E}^{\text{IM}},\mathbf{E}^{\text{RF}})}\left[\log \frac{p(\mathbf{E}^{\text{RF}}|\mathbf{E}^{\text{IM}})}{p(\mathbf{E}^{\text{RF}})}\right]$. Assuming $\mathbf{E}^{\text{RF}}$ follows a uniform distribution (i.e., $p(\mathbf{E}^{\text{RF}}) = \frac{1}{N}$), we have:

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

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

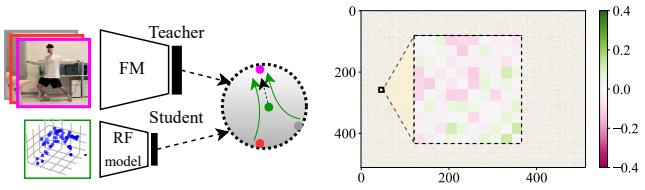
$$p(\mathbf{E}^{\text{RF}}|\mathbf{E}^{\text{IM}}) \geq \frac{\exp(\text{sim}(\mathbf{E}^{\text{IM}},\mathbf{E}^{\text{RF}}))}{\sum_{\mathbf{E}^{\text{RF}'}\in\mathcal{P}}\exp(\text{sim}(\mathbf{E}^{\text{IM}},\mathbf{E}^{\text{RF}'}))}.$$

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

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

where $\text{sim}(\cdot)$ is defined as $\langle\cdot,\cdot\rangle/\tau$, with $\langle\cdot,\cdot\rangle$ being the cosine similarity, and τ being the temperature scaling parameter. It should be noted that, while the mathematical structure of CKD loss may resemble conventional contrastive 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, the student modality interacts solely with the teacher modality for comparison, bypassing intra-modal comparisons and 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 [61].

As shown in Figure 7a, CKD reduces the distance between embeddings of positive RF-image pairs, while simultaneously 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, Figure 7b shows a significant reduction of 0.2 on average, in the difference between the correlation matrices



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

Figure 7: Cross-modal CKD.

of the FM and RF, denoted as $|\mathbf{R}^{\text{IM}} - \mathbf{R}^{\text{RF}}|$, when utilizing CKD. This contrasts with the outcomes observed with traditional KD, as depicted in Figure 3b. This observation underscores CKD's superiority in aligning the structural characteristics of the embeddings across diversified modalities.

3.3 Extraneous Feature Elimination

As described in § 2.3, our goal is to remove extraneous features to allow CKD to focus on HAR-relevant features. To improve interpretability, facilitate better integration, and reduce the consumption of computational resources, we perform feature elimination by utilizing signal properties and FM's knowledge without any extra models.

3.3.1 Image Modality. Instead of employing an extra segmentation model for annotation, we employ the image and text encoders from the teacher model in the CKD framework to generate saliency maps [57]. A saliency map has the same dimensions as the input image, where each element's magnitude quantifies the importance of the corresponding pixel in determining the model's predictive output of a human. It enables the isolation of image regions that are pertinent to human activity, allowing for the exclusion of non-essential features. Compared with other segmentation approaches, FM-Fi 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. As shown in Figure 8, an image processed through the CLIP encoder produces an embedding vector that encapsulates the spatial and contour information of the human body. We compute the similarity score $\mathcal{S}$ by comparing this vector with the text embedding of *"a photo of a human"*, which provides a structural interpretation of these attributes. Following this, we determine the gradient of $\mathcal{S}$ with respect to each input feature of the original image. The aggregate of gradients within the designated target region T signifies the relevance of that feature to the model's output. The saliency map M's individual elements can be obtained as follows:

$$M(u,v) = \textstyle\sum_{(u_t,v_t)\in T} \partial\mathcal{S}/\partial I(u_t,v_t), \tag{2}$$

where (u,v) represents the pixel coordinates and I the original image. In practice, backpropagation can be applied to the scores, generating a chain of gradients across layers equivalent to the gradient in Eqn. (2). This process infers the critical elements within each layer that the model deems essential for discrimination, culminating in the identification of salient features within the input. As a result, the processed saliency region can be expressed as $\mathcal{F}_{\text{sal}}(u,v) = \mathbb{I}\left[M'(u,v) > \lambda\right]$, where $\mathbb{I}(\cdot)$ is the indicator function and λ is the threshold, and M' denotes the normalized saliency map. Elements exceeding the threshold retain their original pixel values, whereas those below the threshold are subjected to Gaussian kernel blurring. The extensive knowledge and complex architecture of the FM contributes to its accurate outputs and reliable reasoning process. As a result, saliency maps obtained from it efficiently concentrate on the relevant features in images.

3.3.2 RF Modality. Within the RF modality, we first eliminate static backgrounds based on the intrinsic physical properties of the data through a deductive approach. Taking mmWave radar as an example, the sensor emits electromagnetic waves in the range of

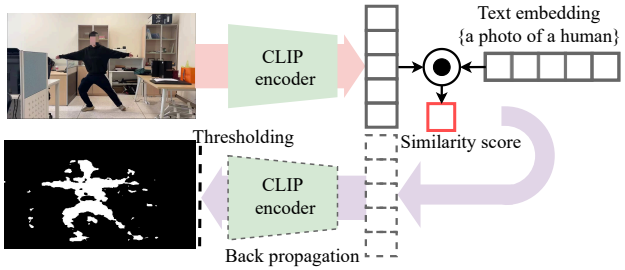


Figure 8: Generation of (threshold) HAR saliency map.

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 machine learning-friendly input features, such as point clouds. Specifically, distance is calculated based on the time interval between the emission and reception of the waves, while the angle of an object can be estimated using multiple receiving antennas. The velocity of an object is inferred through the Doppler effect, which dictates that the frequency shift of the radar waves can be formulated as $f_d = \frac{2v}{c} f_0$, 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.

However, in addition to the static background, the scene may also contain other objects that are irrelevant to HAR (e.g., a moving pet). These objects can be identified by integrating the aforementioned Doppler data, with intensity data that conveys material characteristics, in conjunction with the coordinates derived from point clouds. As detailed in § 3.1 and Figure 6, the RF encoder firstly transforms the three-dimensional coordinates through rotation and scaling, which are then concatenated with Doppler and intensity to form an enriched point cloud feature vector. To eliminate HAR-irrelevant objects, we introduce a self-attention-based module within the RF encoder which allows the RF model to autonomously discern points of interest within a global context by learning from background-free FMs during the cross-modal learning process. 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 $\mathbf{p}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$. The self-attention mechanism enables the model to learn to score based on $\mathbf{W}_q$, $\mathbf{W}_k$, and $\mathbf{W}_v$, which prioritizes points relevant to HAR by considering inter-point relationships and the influence of individual points on the global outcome. Our self-attention module explicitly expresses the focus on specific regions, thereby offering enhanced interpretability. Moreover, in contrast to standard MLPs with fixed architectures, it dynamically adjusts the weighting of points according to the input data distribution, thereby increasing the reliability of the model's decision-making.

3.4 Zero-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 action 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 numerical representations rather than the original natural language to generate a 512-dimensional text 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 $\hat{I}$. The prediction process can be formulated as $\hat{I} = \arg\max_{\mathbf{E}^{\text{TX}}} \left(\frac{\mathbf{E}^{\text{TX}} \cdot \mathbf{E}^{\text{IM}}}{\|\mathbf{E}^{\text{TX}}\| \|\mathbf{E}^{\text{IM}}\|} \right)$, where $\mathbf{E}^{\text{TX}}$ 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 $= W_{\text{zero-shot}} \cdot \mathbf{E}^{\text{IM}}$. Given that each text embedding is normalized, we identify the category corresponding to the highest score to make prediction. This matrix computation methodology prevents redundant calculations and enhances the overall efficiency.

3.5 Metric-Based Few-Shot HAR

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'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:

$$P(y_c|\mathbf{E}^{\text{q}}, \mathcal{D}_c^{\text{s}}) = \sum_{\mathbf{E}_c^{\text{s}} \in \mathcal{D}_c^{\text{s}}} \langle \mathbf{E}^{\text{q}} \cdot \mathbf{E}_c^{\text{s}} \rangle + \gamma \langle \mathbf{E}^{\text{q}} \cdot \mathbf{E}_c^{\text{TX}} \rangle, \tag{3}$$

where $\mathcal{D}^s$ is the support set, $\mathbf{E}^s$ and $\mathbf{E}^q$ denote the embeddings of a support and query sample, $\mathbf{E}_c^{TX}$ 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 the dataset collection and processing, as well as the system implementation of FM-Fi.

4.1 Dataset

For the RF modality, we acquire data using a Texas Instruments (TI) IWR1443 Boost mmWave radar [60]. This radar operates within the 76-81GHz frequency spectrum, offering a bandwidth of 4GHz. 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 200ms, 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 0.1-10Hz range [45], we set the radar sampling rate to 20Hz. 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 [40] at the same conditions as the aforementioned mmWave radar. This camera is 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 186cm, weights from 51 to 109kg, 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 meters. 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), and stadium (SD). 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. Collectively, these 10 different environments exhibit unique floor plans and background objects, underscoring the diversity of real-world scenarios.

Additionally, as elaborated in § 3.2, 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.

4.2 System Implementation

We conduct all experiments, including model training, inference, and saliency map generation, on an NVIDIA TESLA V100 GPU equipped with 16GB of RAM. 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:

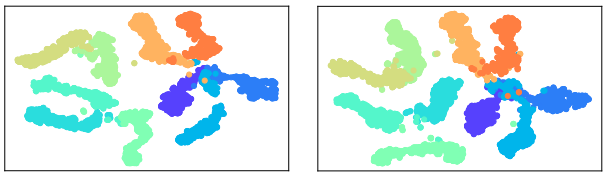
- We choose ViT-B/32 in CLIP as our vision encoder and a custom mmWave point cloud encoder, outlined in § 3.1, featuring 1-d convolutional and linear layers with batch normalization and a 0.3 dropout rate.
- For feature selection, we set a saliency threshold of 0.6, applying Gaussian blur exclusively to regions falling below this threshold using a kernel size of 30.
- We employ an Adam optimizer with a learning rate of 0.001 for both cross-modal distillation and few-shot learning, with the latter exploring 1 to 3 shots.
- 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 employs continuous, non-overlapping frames for training and testing, instead of random sampling of frames to avoid overfitting caused by neighboring frames.

5 EVALUATION

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

5.1 Experiment Setup

To evaluate the performance of FM-Fi, we select 3 sets of baselines for comparison. First, we compare the FM-Fi'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 [17] and MetaSense [26]. Further, we compare the performance of FM-Fi against SOTA point-cloud models, PointNet++ [48] and Point Transformer [72]. Lastly, to assess FM-Fi's performance in unseen environments, we include its teacher model CLIP [49] for comparison.



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

Figure 9: t-SNE plot of embeddings.

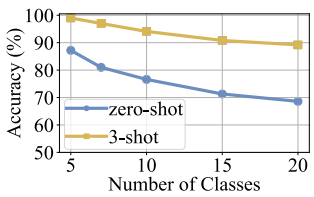


Figure 11: Impact of the number of classes.

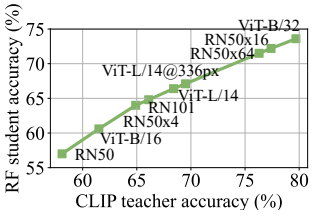


Figure 12: Student vs. teacher accuracy.

- **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.
- **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 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.

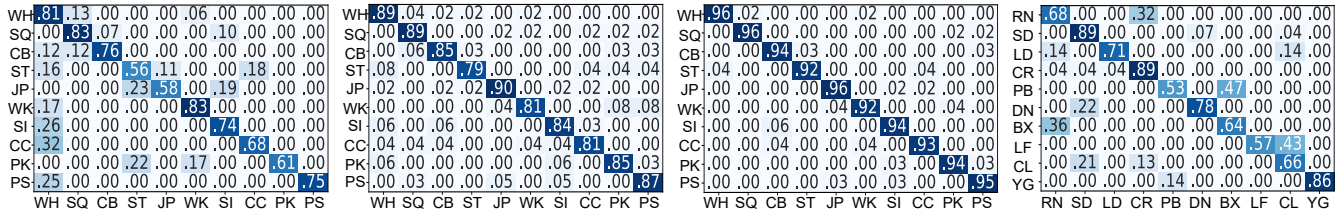
5.2 Overall Evaluation of FM-Fi

To evaluate whether FM-Fi 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 Figure 9a, 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. Figure 9b shows that FM-Fi's embeddings are separable and closely aligned with the teacher model's, indicating that FM-Fi has effectively captured the teacher model's representational power.

In Figure 10, we show FM-Fi's performance across various zero/few-shot scenarios. It can be seen that even in the challenging zero-shot context, FM-Fi is capable of basic HAR tasks with a notable 72.5% accuracy. FM-Fi also achieves accuracies of 86.0% and 94.4% 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 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's performance on 10 new activities mentioned in § 5.1. Figure 10d illustrates that the accuracy on new activities aligns with the results in Figure 10a, 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 Figure 11. 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 Figure 12, a stronger teacher model is associated with improved performance of the student model. This results from the teacher's



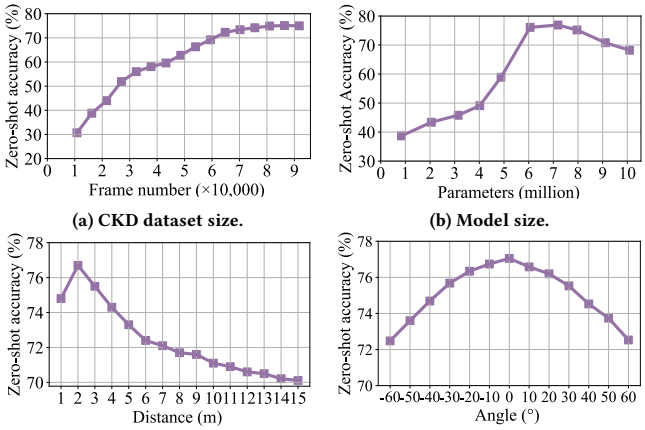
(a) Zero-shot FM-Fi. (b) 1-shot FM-Fi. (c) 3-shot FM-Fi. (d) Zero-shot (10 new classes).

Figure 10: Confusion matrices of FM-Fi in zero-shot and few-shot scenarios.

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 79.7% on the zero-shot HAR task, with the corresponding student model also evaluated in the same setting, achieving 73.6% accuracy. Compared with the vision modality, the RF modality shows no significant decline in performance, 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. As depicted in Figure 13a, the zero-shot accuracy increases as the number of CKD data samples increases from 10,000 to 90,000, but stops to increase when the number of CKD data reaches 80,000, stabilizing at approximately 75%. This is close to the 79.7% accuracy of the teacher model, indicating the efficacy of FM-Fi's CKD. We then examine the impact of the number of model parameters, as shown in Figure 13b. It can be observed that FM-Fi's zero-shot accuracy improves with as the number of parameters increases, reaching a peak of 77% when the number of parameters reaches 7 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.

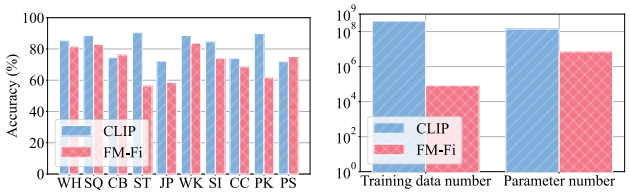
Finally, we conduct experiments to evaluate the performance of FM-Fi when the distance and angle from the subject to the sensor vary. Specifically, the subject performed activities at distances ranging from 1 to 15m and at angles from -60° to 60°, which corresponds to the radar's FoV. For each fixed distance, we calculate the average accuracy across all angles; likewise, for each fixed angle, we average the accuracy over all distances. Figure 13c demonstrates that the model's accuracy initially increases as the distance grows, then decreases beyond 2 m, peaking at 76.7%. This optimal performance at 2 m is due to the radar's ability to fully capture the subject's body at this distance. At shorter distances, the beam of the radar cannot encompass the entire body; and at larger distances, the signal-to-noise ratio diminishes, lowering the quality of the input data. Despite the degradation, the accuracy remains consistently above 70.2%, illustrating the strong generalization capability



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

(c) Subject-sensor distance. (d) Subject-sensor angle.

Figure 13: Impact of practical factors.



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

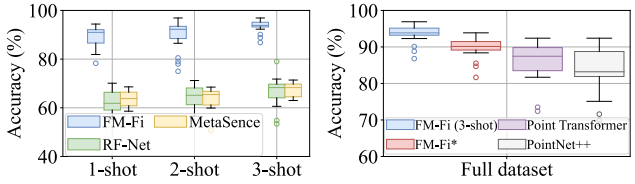
Figure 14: Comparison with FM baseline.

of FM-Fi under varying distance conditions. Similarly, Figure 13d shows that the model's accuracy decreases as the subject's absolute angle relative to the sensor increases. However, the accuracy remains above 72.9% across all angles and peaks at 76.7% when the subject directly faces the sensor (angle of 0°). The decline in performance becomes more pronounced near the edges of the FoV due to a sharper drop in radiated power and signal quality. Nonetheless, the performance degradation is modest, with the decrease not exceeding 3.8%. Collectively, these results demonstrate that FM-Fi maintains robust performance, effectively handling variations in both distance and angle, making FM-Fi delivers consistent and high accuracy, ensuring sufficient and reliable coverage for HAR.

5.3 Superiority of FM-Fi

5.3.1 Comparison with FM. We compare FM-Fi with FM by assessing their zero-shot capabilities. As shown in Figure 14a, the accuracy of FM-Fi closely matches that of CLIP across all 10 activity classes, illustrating the overall effectiveness of FM-Fi. An interesting phenomenon is that for the two classes of *CB* and *PS*, 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 the extraneous feature elimination 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's model, with its 6.9 million parameters, is significantly smaller than CLIP's 140 million parameters, as depicted in Figure 14b. Although the model-to-data size ratio of FM-Fi 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's ability to deliver competitive performance with considerably less data and a more compact architecture.

5.3.2 Comparison with Few-shot Baselines. We further compare FM-Fi with two few-shot baselines MetaSense and RF-Net. In the 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. Figure 15a features boxplots that detail the comparative performance of them under 1, 2, and 3-shot settings. In all three scenarios, FM-Fi consistently outperforms the two baselines by a significant margin. Although as the number of samples increases, the median accuracy of FM-Fi does not rise as quickly as that of



(a) Few-shot learning. (b) Supervised learning.

Figure 15: Comparison with different baselines.

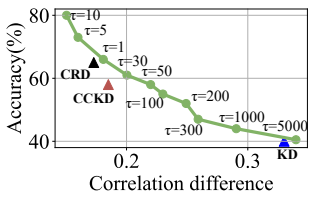


Figure 17: Impact of interdependency transfer.

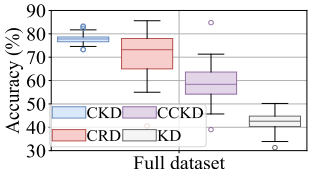


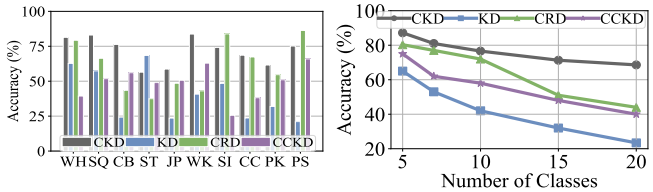
Figure 18: Comparisons on accuracy and stability.

the baselines, it still maintains a lead of at least 23.7%. Furthermore, the interquartile range (IQR) of FM-Fi's accuracy is considerably smaller than that of the baselines, indicating less variability across multiple experiments.

5.3.3 Comparison with Supervised Baselines on a Larger Dataset. We further compare FM-Fi with PointNet++ and Point Transformer. 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*, which utilizes the same RF encoder as FM-Fi (with an ensuing multilayer perception for converting the embedding to classification result). FM-Fi* 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 not only demonstrates superior accuracy but also greater stability in performance. 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, FM-Fi* exhibits notably better performance than the other two models that are specifically designed for point clouds. The superior performance of FM-Fi* 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.

5.4 CKD Evaluation

We further compare CKD with KD, as well as contrastive representation distillation (CRD) [61], and correlation congruence for knowledge distillation (CCKD) [46]. First, we compare their performance on a 10-class zero-shot HAR task, as illustrated in Figure 16a. We observe that CKD achieves the highest accuracy in 7 out of 10 classes, only trailing the best method by less than 9.8% in the rest 3 classes. Figure 16b further examines the impact of the number of activity classes. It can be observed that when the number of classes is 5, CKD leads other methods by a small margin less than 18.5%.As the number of classes increases to 20, CKD exhibits the smallest drop in accuracy, while the accuracies of all other methods drop by more than 37.3%.



(a) Class-wise accuracy. (b) Number of classes.

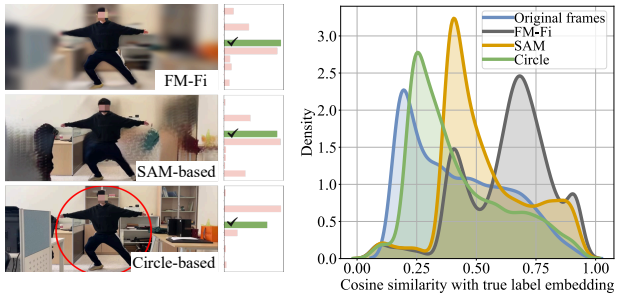
Figure 16: Generalization comparison.

To understand CKD's superiority, we analyze the relationship between the FM-Fi'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. Our findings shown in Figure 17 demonstrate that the correlation difference negatively impacts FM-Fi's accuracy (τ = 10 yields the best performance). This trend validates FM-Fi '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 near the curve) can be attributed to their pronounced correlation differences, which correspond to a diminished efficacy in the transfer of interdependency knowledge.

Finally, we conduct 50 training rounds (number of classes set to 10), and perform statistical analysis of the accuracies of various distillation methods, and show the results in Figure 18. It can be seen that CKD exhibits the highest median accuracy and narrowest IQR. In contrast, CRD, CCKD, and KD demonstrate lower accuracies and larger IQR. 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.

5.5 Feature Elimination Evaluation

5.5.1 Image Modality. In addition to a saliency map-based feature elimination method, Segment-Anything (SAM) [34] and circle prompting [56] can also be used to enhance human focus in images. SAM is the SOTA segmentation algorithm and is shown to be superior for the HAR task; whereas circle prompting emphasizes crucial information within an image through circular markings made by a human annotator. Figure 19a depicts the images processed with our saliency map-based method, SAM-based method, and circle-based method, respectively, along with the predictions made by CLIP. The ground truth classes have been highlighted



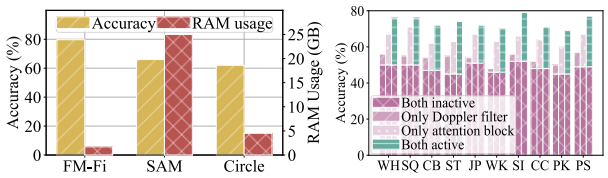
(a) Example segmentation methods and activations. (b) Probability distributions of cosine similarity.

Figure 19: Performance comparison of different image feature elimination methods.

by green for reference. It is evident that due to the extraneous background features, CLIP fails to correctly classify the outcomes from the other two prompting methods. In contrast, our approach effectively mitigates extraneous features, making CLIP solely focus on the human subject, thus enabling more refined classification. The probability distributions of the similarity between true and predicted embedding in Figure 19b further prove the advantage of FM-Fi's extraneous feature elimination module.

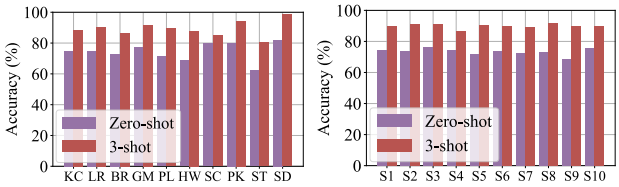
Among the three methods, FM-Fi is the only one capable of automatic background removal. This is attributed to FM-Fi's ability to autonomously identify people within images through semantic input. In contrast, the SAM-based method requires manual selection post-segmentation to remove background elements, while the circle-based approach relies entirely on manual annotation of significant objects. We have also compared the resource demands of these methods, noting that manual costs are challenging to assess directly. Therefore, we translate the manual annotation task in the circle-based method into an automated segmentation and highlight-filling task. Figure 20a shows that our approach not only leads in performance but also in reduces resource utilization, thus underscoring the superiority of the saliency map-based feature elimination module of FM-Fi.

5.5.2 RF Modality. We also evaluate the feature elimination module of the RF modality, which consists of two components: a Doppler-based object filter and an attention block. To validate the efficacy of the proposed module, we conduct ablation studies by removing the Doppler filter and attention block. As such, the experiment involves the following 4 configurations: both parts inactive, Doppler filter inactive, attention block inactive, and both parts active. We select 10-class classification for evaluation. Figure 20b presents the



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

Figure 20: Comparison of feature elimination methods.



(a) Different scenes. (b) Different subjects.

Figure 21: Generalization across diverse settings.

findings, where we plot the effect of incorporating various feature elimination components on the accuracy of zero-shot classification.

Not surprisingly, when both modules are inactive, the accuracy is the worst and only reaches less than 52.3%.The result also shows that the Doppler filter, upon excluding objects with zero velocity, has a positive but limited effect on the model. The attention block, with its automatic selection of important points, offers a more effective improvement to overall performance. The combined effect of both parts surpasses that of each individual component. This synergy can be attributed to the Doppler filter's introduction of physical priors that enhance the subsequent decision-making process of the attention block, thereby underscoring the effectiveness and necessity of our module.

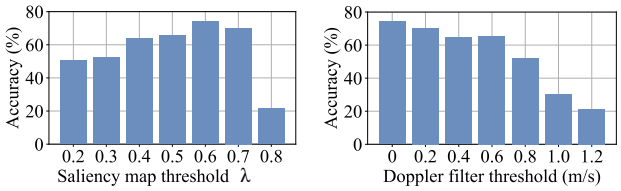
5.6 Generalization Capability

To evaluate the generalization capabilities of FM-Fi, we 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 × 10 subjects) and the results shown in Figure 21 are obtained by averaging across either environments or subjects.

Overall, Figure 21a 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 maintains performance levels consistent with previous tests, demonstrating exceptional capabilities. Moreover, physiological parameters such as age and height of participants do not affect the performance of FM-Fi, as evidenced by Figure 21b, which illustrates that our model maintains at least 68.5% zero-shot accuracy and 89.8% 3-shot accuracy across all subjects, demonstrating robust generalization capabilities.

5.7 Hyper-parameter Searching

5.7.1 Feature Elimination Threshold. According to § 3.3.1, the threshold λ is a scalar within the [0,1] range, determining the lower bound normalized 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



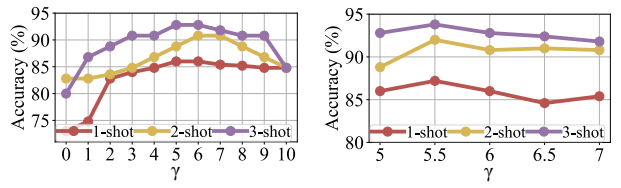
(a) Image modality. (b) RF Modality.

Figure 22: Feature elimination thresholds.

generated cross-modal supervision signal. To search for the optimal value of λ, we evaluate the zero-shot performance of FM-Fi at different λ values from 0.2 to 0.8. One may readily observe in Figure 22a that as λ initially increases, FM-Fi reaches the best performance at the optimal threshold $\lambda = 0.6$. Any λ greater than 0.6 may cause the saliency mask to erode the human figure, adversely impacting HAR performance. Consequently, as λ surpasses 0.6, there is a significant decline in accuracy from 76.2% to 21.6%.

We further study the impact of velocity thresholds in extraneous feature elimination for the RF modality. Instead of only removing the zero-velocity component as in § 3.3.2, we set the velocity filtering thresholds from 0 to 1.2m/s, and show the relationship between the model's accuracy and the threshold in Figure 22b. 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 0.8m/s, a significant portion of human activity information is lost, resulting in poor model performance. Based on these experiment observations, a threshold of 0 is selected to preserve all information of moving objects while excluding static background features, leaving the subsequent attention module to make further selections.

5.7.2 Weight of Label Text in Few-shot Learning. We evaluate the impact of varying weights of label text γ on FM-Fi'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 Figure 23a. 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. 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 $\gamma = 5$ to $\gamma = 7$. As Figure 23b demonstrates, the



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

Figure 23: Impact of the weight of label text.

peak performance is obtained at $\gamma = 5.5$. Consequently, we adopt a γ value of 5.5 as the weight of text label in few-shot learning.

6 RELATED WORK AND DISCUSSION

Though RF-HAR literature covers enhancing generalizability [11, 17, 25, 26, 33, 39, 77], improving the efficient utilization of scarce labeled data [36, 43], and refining model architectures [8, 11, 35], prominent RF-HAR proposals have prioritized studies on generalizability. In particular, Widar3.0 [77] introduces a domain-independent and signal-level feature, termed BVP, to enable generalizability. Another study [33] applies adversarial domain adaptation techniques [23, 24] to generalize across varying scenarios. RF-Net [17] adopt metric-based meta-learning achieve fast adaptation of its base networks in diverse environments.

The emergence of FMs has brought new potentials in RF sensing in general, catering the need for more models capable of capturing rich information. Therefore, FM-Fi sets itself apart from prior RF-HAR solutions by not limiting itself to HAR, because it inherits the broad recognition capability of FM. In fact, we would expect FM-Fi to be able to support other sensing tasks [12, 30, 32, 70, 71, 73, 76] including gesture detection [37, 38, 69], gait recognition [7, 59, 65], and even vibration monitoring [1, 13, 14, 67, 68], by modifying the target of interest; we plan to explore FM-Fi's potential beyond HAR in future work. Currently, it is still an open question if one can claim open-set capability for FM-enabled HAR [52]. Also, whether FM-Fi may completely inherit the knowledge of FM (obtained from massive datasets encompassing a broad spectrum of activities) needs further studies. Furthermore, the question of how to compress the RF model by quantization when transferring knowledge from the FM [6] is also of practical significance. As FM-Fi pioneers in the knowledge transferring from FM to RF-HAR, we leave these uncertainties to future exploration.

7 CONCLUSION

Taking a significant stride in advancing HAR, we have introduced FM-Fi, which harnesses the interpretative power of FMs to facilitate cross-modal RF-HAR. By employing CKD and extraneous feature elimination, the innovative RF encoder in FM-Fi 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'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] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. 2015. Smart Homes that Monitor Breathing and Heart Rate. In *Proc. of the 33rd ACM CHI*. 837–846.

[2] Kshitiz Bansal, Keshav Rungta, Siyuan Zhu, and Dinesh Bharadia. 2020. Pointillism: Accurate 3D Bounding Box Estimation with Multi-Radars. In *Proc. of the 18th ACM SenSys*. 340–353.

[3] Chongguang Bi, Guoliang Xing, Tian Hao, Jina Huh-Yoo, Wei Peng, Mengyan Ma, and Xiangmao Chang. 2019. FamilyLog: Monitoring Family Mealtime Activities by Mobile Devices. *IEEE Transactions on Mobile Computing* 19, 8 (2019), 1818–1830.

[4] Tara Boroushaki, Junshan Leng, Ian Clester, Alberto Rodriguez, and Fadel Adib. 2021. Robotic Grasping of Fully-Occluded Objects using RF Perception. In *Proc. of IEEE ICRA*. IEEE, 923–929.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *Proc. of NeurIPS* 33 (2020), 1877–1901.

[6] Kaiwen Cai, Zhekai Duan, Gaowen Liu, Charles Fleming, and Chris Xiaoxuan Lu. 2024. Self-Adapting Large Visual-Language Models to Edge Devices across Visual Modalities. *arXiv preprint arXiv:2403.04908* (2024).

[7] Dongjiang Cao, Ruofeng Liu, Hao Li, Shuai Wang, Wenchao Jiang, and Chris Xiaoxuan Lu. 2022. Cross Vision-RF Gait Re-identification with Low-cost RGB-D Cameras and mmWave Radars. *Proc. of ACM UbiComp* 6, 3 (2022), 1–25.

[8] Mainak Chakraborty, Harish C Kumawat, Sunita Vikrant Dhavale, et al. 2022. DIAT-RadHARNet: A Lightweight DCNN for Radar based Classification of Human Suspicious Activities. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–10.

[9] Dongyao Chen, Mingke Wang, Chenxi He, Qing Luo, Yasha Iravantchi, Alanson Sample, Kang G Shin, and Xinbing Wang. 2021. MagX: Wearable, Untethered Hands Tracking with Passive Magnets. In *Proc. of the 27th ACM MobiCom*. 269–282.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proc. of ICML*. PMLR, 1597–1607.

[11] Zhe Chen, Chao Cai, Tianyue Zheng, Jun Luo, Jie Xiong, and Xin Wang. 2021. RF-based Human Activity Recognition using Signal Adapted Convolutional Neural Network. *IEEE Transactions on Mobile Computing* 22, 1 (2021), 487–499.

[12] Zhe Chen, Tianyue Zheng, Chao Cai, Yue Gao, Pengfei Hu, and Jun Luo. 2023. Wider is Better? Contact-free Vibration Sensing via Different COTS-RF Technologies. In *Proc. of IEEE INFOCOM*. IEEE, 1–10.

[13] Zhe Chen, Tianyue Zheng, Chao Cai, and Jun Luo. 2021. MoVi-Fi: Motion-robust Vital Signs Waveform Recovery via Deep Interpreted RF Sensing. In *Proc. of the 27th ACM MobiCom*. 392–405.

[14] Zhe Chen, Tianyue Zheng, and Jun Luo. 2021. Octopus: a practical and versatile wideband MIMO sensing platform. In *Proc. of the 27th ACM MobiCom*. 601–614.

[15] Zicheng Chi, Yao Yao, Tiantian Xie, Xin Liu, Zhichuan Huang, Wei Wang, and Ting Zhu. 2018. EAR: Exploiting Uncontrollable Ambient RF Signals in Heterogeneous Networks for Gesture Recognition. In *Proc. of the 16th ACM SenSys*. 237–249.

[16] L Minh Dang, Kyungbok Min, Hanxiang Wang, Md Jalil Piran, Cheol Hee Lee, and Hyeonjoon Moon. 2020. Sensor-Based and Vision-Based Human Activity Recognition: A Comprehensive Survey. *Pattern Recognition* 108 (2020), 107561.

[17] Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo. 2020. RF-Net: A Unified Meta-Learning Framework for RF-enabled One-Shot Human Activity Recognition. In *Proc. of the 18th ACM SenSys*. 517–530.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929* (2020).

[19] Sepideh Esmaeilpour, Bing Liu, Eric Robertson, and Lei Shu. 2022. Zero-Shot Out-of-Distribution Detection Based on the Pre-trained Model CLIP. In *Proc. of AAAI*, Vol. 36. 6568–6576.

[20] Lijie Fan, Tianhong Li, Yuan Yuan, and Dina Katabi. 2020. In-Home Daily-Life Captioning Using Radio Signals. In *Proc. of the 16th ECCV*. Springer, 105–123.

[21] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *Proc. of IEEE/CVF CVPR*. 19358–19369.

[22] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: Gait-based User Identification with In-ear Microphones. In *Proc. of the 27th ACM MobiCom*. 337–349.

[23] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proc. of ICML*. PMLR, 1180–1189.

[24] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial Training of Neural Networks. *Journal of Machine Learning Research* 17, 59 (2016), 1–35.

[25] Ruiyang Gao, Wenwei Li, Yaxiong Xie, Enze Yi, Leye Wang, Dan Wu, and Daqing Zhang. 2022. Towards Robust Gesture Recognition by Characterizing the Sensing Quality of WiFi Signals. *Proc. of ACM UbiComp* 6, 1 (2022), 1–26.

[26] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung-Ju Lee. 2019. MetaSense: Few-Shot Adaptation to Untrained Conditions in Deep Mobile Sensing. In *Proc. of the 17th ACM SenSys*. 110–123.

[27] Jianguo Hao, Abdenour Bouzouane, and Sébastien Gaboury. 2018. Recognizing Multi-Resident Activities in Non-Intrusive Sensor-Based Smart Homes by Formal Concept Analysis. *Neurocomputing* 318 (2018), 75–89.

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proc. of IEEE/CVF CVPR*. 9729–9738.

[29] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531* (2015).

[30] Jingyang Hu, Hongbo Jiang, Tianyue Zheng, Jingzhi Hu, Hongbo Wang, Hangcheng Cao, Zhe Chen, and Jun Luo. 2024. M2-Fi: Multi-person Respiration Monitoring via Handheld WiFi Devices. *Proc. of IEEE INFOCOM* (2024).

[31] Jingzhi Hu, Tianyue Zheng, Zhe Chen, Hongbo Wang, and Jun Luo. 2023. MUSE-Fi: Contactless MUti-person SEnsing Exploiting Near-field Wi-Fi Channel Variation. In *Proc. of the 29th ACM MobiCom*. 1–15.

[32] Jinyang Huang, Bin Liu, Chenglin Miao, Xiang Zhang, Jianchun Liu, Lu Su, Zhi Liu, and Yu Gu. 2024. PhyFinAtt: An Undetectable Attack Framework Against PHY Layer Fingerprint-Based WiFi Authentication. *IEEE Transactions on Mobile Computing* 23, 7 (2024), 7753–7770.

[33] Wenjun Jiang, Chenglin Miao, Fenglong Ma, Shuochao Yao, Yaqing Wang, Ye Yuan, Hongfei Xue, Chen Song, Xin Ma, Dimitrios Koutsonikolas, et al. 2018. Towards Environment Independent Device Free Human Activity Recognition. In *Proc. of the 24th ACM MobiCom*. 289–304.

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment Anything. *arXiv preprint arXiv:2304.02643* (2023).

[35] Xiaoxiong Li, Si Chen, Shuning Zhang, Linsheng Hou, Yuying Zhu, and Zelong Xiao. 2023. Human Activity Recognition Using IR-UWB Radar: A Lightweight Transformer Approach. *IEEE Geoscience and Remote Sensing Letters* (2023).

[36] Xinyu Li, Yuan He, Francesco Fioranelli, and Xiaojun Jing. 2021. Semisupervised Human Activity Recognition with Radar Micro-Doppler Signatures. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2021), 1–12.

[37] Yang Liu, Zhenjiang Li, Zhidan Liu, and Kaishun Wu. 2019. Real-time Arm Skeleton Tracking and Gesture Inference Tolerant to Missing Wearable Sensors. In *Proc. of the 17th ACM MobiSys*. 287–299.

[38] Pedro Melgarejo, Xinyu Zhang, Parameswaran Ramanathan, and David Chu. 2014. Leveraging Directional Antenna Capabilities for Fine-grained Gesture Recognition. In *Proc. of ACM UbiComp*. 541–551.

[39] Francesca Meneghello, Domenico Garlisi, Nicolò Dal Fabbro, Ilenia Tinnirello, and Michele Rossi. 2022. ShARP: Environment and Person Independent Activity Recognition with Commodity IEEE 802.11 Access Points. *IEEE Transactions on Mobile Computing* (2022).

[40] Microsoft. 2020. Kinect Sensor. https://developer.microsoft.com/en-us/windows/kinect/. Accessed: 2020-09-29.

[41] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. 2022. Simple Open-Vocabulary Object Detection with Vision Transformers. In *Proc. of ECCV*. Springer, 728–755.

[42] Kai Niu, Fusang Zhang, Jie Xiong, Xiang Li, Enze Yi, and Daqing Zhang. 2018. Boosting Fine-grained Activity Sensing by Embracing Wireless Multipath Effects. In *Proc. of the 14th ACM CoNEXT*. 139–151.

[43] Xiaomin Ouyang, Zhiyuan Xie, Jiayu Zhou, Jianwei Huang, and Guoliang Xing. 2021. ClusterFL: A Similarity-aware Federated Learning System for Human Activity Recognition. In *Proc. of the 19th ACM MobiSys*. 54–66.

[44] Sameera Palipana, Dariush Salami, Luis A Leiva, and Stephan Sigg. 2021. Pantomime: Mid-Air Gesture Recognition with Sparse Millimeter-Wave Radar Point Clouds. *Proc. of ACM UbiComp* 5, 1 (2021), 1–27.

[45] Hyung O Park, Alireza A Dibazar, and Theodore W Berger. 2009. Cadence Analysis of Temporal Gait Patterns for Seismic Discrimination Between Human and Quadruped Footsteps. In *Proc. of IEEE ICASSP*. IEEE, 1749–1752.

[46] Baoyun Peng, Xiao Jin, Jiaheng Liu, Dongsheng Li, Yichao Wu, Yu Liu, Shunfeng Zhou, and Zhaoning Zhang. 2019. Correlation Congruence for Knowledge Distillation. In *Proc. of IEEE/CVF ICCV*. 5007–5016.

[47] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. of IEEE/CVF CVPR*. 652–660.

[48] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *Proc. of NeurIPS* 30 (2017).

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proc. of ICML*. PMLR, 8748–8763.

[50] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[51] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proc. of ICML*. PMLR, 8821–8831.

[52] Shuhuai Ren, Lei Li, Xuancheng Ren, Guangxiang Zhao, and Xu Sun. 2023. Delving into the Openness of CLIP. In *Proc. of ACL*.

[53] Isidoros Rodomagoulakis, Nikolaos Kardaris, Vassilis Pitsikalis, E Mavroudi, Athanasios Katsamanis, Antigoni Tsiami, and Petros Maragos. 2016. Multimodal Human Action Recognition in Assistive Human-Robot Interaction. In *Proc. of IEEE ICASSP*. IEEE, 2702–2706.

[54] Dariush Salami, Ramin Hasibi, Sameera Palipana, Petar Popovski, Tom Michoel, and Stephan Sigg. 2022. Tesla-Rapture: A Lightweight Gesture Recognition System from mmWave Radar Sparse Point Clouds. *IEEE Transactions on Mobile Computing* (2022).

[55] Ann-Kathrin Seifert, Moeness G Amin, and Abdelhak M Zoubir. 2019. Toward Unobtrusive In-home Gait Analysis Based on Radar Micro-Doppler Signatures. *IEEE Transactions on Biomedical Engineering* 66, 9 (2019), 2629–2640.

[56] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What Does CLIP Know About a Red Circle? Visual Prompt Engineering for VLMs. In *Proc. of IEEE/CVF ICCV*. 11987–11997.

[57] K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Proc. of ICLR*.

[58] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, and Alex Wong. 2023. Depth Estimation From Camera Image and mmWave Radar Point Cloud. In *Proc. of IEEE/CVF CVPR*. 9275–9285.

[59] Minglong Sun, Amanda Watson, and Gang Zhou. 2020. Wearable Computing of Freezing of Gait in Parkinson's Disease: A Survey. *Smart Health* 18 (2020), 100143.

[60] Texas Instruments. 2020. IWR1443BOOST. https://www.ti.com/tool/IWR1443BOOST. Accessed: 2020-09-29.

[61] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive Representation Distillation. In *Proc. of ICLR*.

[62] Hoang Truong, Shuo Zhang, Ufuk Muncuk, Phuc Nguyen, Nam Bui, Anh Nguyen, Qin Lv, Kaushik Chowdhury, Thang Dinh, and Tam Vu. 2018. Capband: Battery-Free Successive Capacitance Sensing Wristband for Hand Gesture Recognition. In *Proc. of the 16th ACM SenSys*. 54–67.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Proc. of NeurIPS* 30 (2017).

[64] Shaohua Wan, Lianyong Qi, Xiaolong Xu, Chao Tong, and Zonghua Gu. 2020. Deep Learning Models for Real-Time Human Activity Recognition with Smartphones. *Mobile Networks and Applications* 25 (2020), 743–755.

[65] Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, and Hongkai Wen. 2019. EV-Gait: Event-based Robust Gait Recognition using Dynamic Vision Sensors. In *Proc. of IEEE/CVF CVPR*. 6358–6367.

[66] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. 2022. Robust fine-tuning of zero-shot models. In *Proc. of IEEE/CVF CVPR*. 7959–7971.

[67] Zongxing Xie, Hanrui Wang, Song Han, Elinor Schoenfeld, and Fan Ye. 2022. DeepVS: A Deep Learning Approach for RF-based Vital Signs Sensing. In *Proc. of the 13rd ACM BCB*. 1–5.

[68] Bo Zhang, Boyu Jiang, Rong Zheng, Xiaoping Zhang, Jun Li, and Qiang Xu. 2023. Pi-Vimo: Physiology-inspired Robust Vital Sign Monitoring using mmWave Radars. *ACM Transactions on Internet of Things* 4, 2 (2023), 1–27.

[69] Shujie Zhang, Tianyue Zheng, Zhe Chen, Jingzhi Hu, Abdelwahed Khamis, Jiajun Liu, and Jun Luo. 2023. OCHID-Fi: Occlusion-Robust Hand Pose Estimation in 3D via RF-Vision. In *Proc. of IEEE/CVF ICCV*. 15112–15121.

[70] Shujie Zhang, Tianyue Zheng, Zhe Chen, and Jun Luo. 2022. Can We Obtain Fine-grained Heartbeat Waveform via Contact-free RF-sensing?. In *Proc. of IEEE INFOCOM*. IEEE, 1759–1768.

[71] Shujie Zhang, Tianyue Zheng, Hongbo Wang, Zhe Chen, and Jun Luo. 2022. Quantifying the Physical Separability of RF-based Multi-Person Respiration Monitoring via SINR. In *Proc. of the 20th ACM SenSys*. 47–60.

[72] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. 2021. Point Transformer. In *Proc. of IEEE/CVF ICCV*. 16259–16268.

[73] Tianyue Zheng, Zhe Chen, Chao Cai, Jun Luo, and Xu Zhang. 2020. V2iFi: In-Vehicle Vital Sign Monitoring via Compact RF Sensing. *Proc. of ACM UbiComp* 4, 2 (2020), 1–27.

[74] Tianyue Zheng, Zhe Chen, Jun Luo, Lin Ke, Chaoyang Zhao, and Yaowen Yang. 2021. SiWa: See into Walls via Deep UWB Radar. In *Proc. of the 27th ACM MobiCom*. 323–336.

[75] Tianyue Zheng, Zhe Chen, Shujie Zhang, Chao Cai, and Jun Luo. 2021. MoRe-Fi: Motion-robust and Fine-grained Respiration Monitoring via Deep-Learning UWB Radar. In *Proc. of the 19th ACM SenSys*. 111–124.

[76] Tianyue Zheng, Ang Li, Zhe Chen, Hongbo Wang, and Jun Luo. 2023. AutoFed: Heterogeneity-Aware Federated Multimodal Learning for Robust Autonomous Driving. In *Proc. of the 29th ACM MobiCom*. 1–15.

[77] Yue Zheng, Yi Zhang, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2019. Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi. In *Proc. of the 17th ACM MobiSys*. 313––325.