RFusion: Dynamic Multimodal RF Fusion for Few-Shot Human Activity Recognition

Chao Feng, Jiashen Chen, Shuo Liang, Xiaopeng Peng, Baizhou Yang, Xuan Wang, Zexuan Huang, Xianjia Meng, Xiaojiang Chen

***Abstract*—Multimodal learning plays a critical role in Radio-Frequency (RF)-based Human Activity Recognition (HAR). However, most existing methods require a large amount of dense annotation across all modalities during the pre-training and fine-tuning stages, which is laborious and expensive. In this work, we introduce RFusion, a dynamic multimodal fusion model for RF-based few-shot human activity recognition. Our model makes use of unlabeled multimodal RF data in the pre-training stage, which further benefits the performance of the few-shot fine-tuning in the absence of modalities. Specifically, RFusion introduces a novel pre-training framework which involves two novel modules (a guider and an arbiter) for a dynamic contrastive learning strategy. This novel pre-training framework allows accurate detection of the shared and unique features from diverse RF modalities. Additionally, a scalable multi-head attention scheme is also introduced to fine-tune the pre-trained model in few-shot settings. Extensive experiments on publicly available multimodal datasets demonstrate the effectiveness of RFusion. The proposed RFusion achieves average HAR accuracy improvements of 25.8% and 12.2% over a number of state-of-the-art supervised unimodal learning methods and contrastive learning baselines.**

***Index Terms*—Human Activity Recognition, RF Sensing, Multimodal Learning, Few-shot Learning**

I. Introduction

Human Activity Recognition (HAR) has been widely used in many applications, such as smart homes [1]–[6], health monitoring [7]–[10], and smart factories [11]–[15]. Compared to vision-based solutions [16], [17], HAR based on radio frequency (RF) techniques such as WiFi, RFID, and mmWave is particularly attractive due to its contact-free, privacy-preserving, and light-invariant properties.

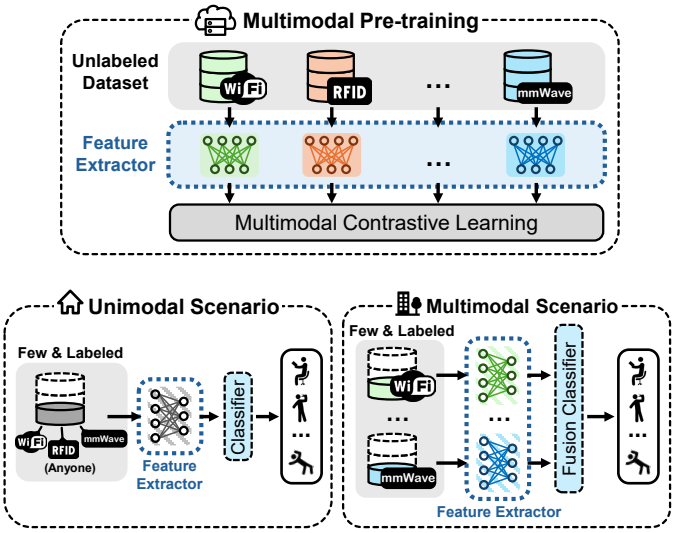


Figure 1: Application scenarios and principle of RFusion.

Early RF-based HAR systems mainly rely on a single sensing modality [18]–[20]. Recently, multimodal sensing [21]–[24] has emerged, where each sensing modality makes a unique contribution to the recognition of the same activity and offers a new perspective. Despite promising progress, directly adopting existing multimodal methods developed for images, text, and audio to RF signals is non-trivial. First, RF modalities exhibit stronger heterogeneity than typical RGB–text–audio pairs. For example, different RF modalities (e.g., WiFi, RFID, and mmWave radar) work at their respective frequencies, protocols, and hardware, which may lead to disparities in the dimensions and patterns of the measured signals. In addition, RF data faces severe label scarcity. Unlike images, where human activities can be manually annotated from videos, raw RF signals are complex-valued, phase-sensitive, and not directly interpretable. Even with auxiliary cameras or IMUs, synchronization and calibration overheads make large-scale annotation costly and labor-intensive.

What is more, many existing multimodal approaches assume that the same set of modalities is available from pre-training to deployment [25]–[27]. However, this assumption is difficult to maintain in RF modality scenarios. In reality, RF modalities are usually scarce due to deployment costs, many HAR applications only deploy one or two RF modalities, like a WiFi router in a smart home, as demonstrated in Fig. 1.

To this end, we ask the following question: *Can we*

This work was supported by the National Natural Science Foundation of China under Grant 62276211, 62302392 and W2511073, and the Project of Shaanxi Province International Science and Technology Cooperation Program under Grant 2025GH-YBXM-057 and 2024GH-YBXM-10. (Corresponding author: Xianjia Meng.)

Chao Feng is with the Shaanxi Key Laboratory of Passive Internet of Things and Neural Computing, School of Computer Science, Northwest University, Xi'an 710127, China (e-mail: chaofeng@nwu.edu.cn).

Jiashen Chen and Shuo Liang are with the Shaanxi International Joint Research Centre for the Battery-Free Internet of Things, School of Computer Science, Northwest University, Xi'an 710127, China (e-mail: 2022115102@stumail.nwu.edu.cn, 202221444@stumail.nwu.edu.cn).

Xiaopeng Peng is with Rochester Institute of Technology, Rochester NY 14623, USA (e-mail: xxp4248@rit.edu).

Baizhou Yang, Xuan Wang and Zexuan Huang are with the Xi'an Key Laboratory of Advanced Computing and System Security, School of Computer Science, Northwest University, Xi'an 710127, China (e-mail: yangbaizhou@163.com, xuanwang@nwu.edu.cn, 202421916@stumail.nwu.edu.cn).

Xianjia Meng is with the School of Computer Science, Northwest University, Xi'an 710127, China (e-mail: xianjiam@nwu.edu.cn).

Xiaojiang Chen is with the Internet of Things Research Center, School of Computer Science, Northwest University, Xi'an 710127, China (e-mail: xjchen@nwu.edu.cn).

design an RF-based multimodal framework that relies on limited labeled RF data and performs effectively in RF modality scarcity scenarios while retaining full-modality performance? If this vision is realized, we can reap several benefits. First, it would significantly reduce annotation costs. Second, this framework enables cross-modal sensing while achieving performance comparable to that of the full-modality setting. For example, a single RF modality, like WiFi, could achieve recognition results comparable to those obtained by using multiple modalities, such as WiFi, mmWave, and RFID combined.

To harvest these benefits, in this paper, we present RFusion, a multimodal RF representation learning framework that leverages unlabeled multimodal RF data for pre-training and supports few-shot adaptation under modality-missing deployments, as shown in Fig. 1. At a high level, RFusion first uses modality-adapted encoders to extract latent features from heterogeneous RF modalities. On top of these encoders, RFusion performs RF-tailored cross-modal pre-training to learn both shared and modality-unique representations, and then fine-tunes the trained model using a few labels during the deployment stage. To build this idea into a practical system, the main challenge we face is how to effectively learn the mutual information between modalities while extracting unique information from different RF modalities.

To address this challenge, in the pre-training stage, RFusion designs a novel cross-modal feature contrastive learning network. Specifically, RFusion designs a guider module to evaluate the alignment of internal feature vector intervals in real-time during contrastive learning. The module allocates more attention to poorly aligned intervals, enabling the network to extract highly consistent features across modalities. Furthermore, to ensure the guider module effectively allocates attention, RFusion employs an adversarial guiding strategy, maximizing the advantage of guided learning over unguided learning to optimize attention distribution. Next, to obtain unique information from each modality, RFusion integrates an entropy-based arbiter module that evaluates the contribution of each modality by analyzing the entropy of pseudo-classification features from different RF modalities. This module assigns relative attention scores, guiding how each modality's unique knowledge contributes to the final loss calculation and ensuring effective utilization of distinctive information to enhance overall model performance. The above schemes augment RFusion with a cross-modal capability, enabling it to effectively integrate and learn from diverse RF modalities.

In the fine-tuning stage, a multi-head attention network is designed to improve the performance of RFusion few-shot learning from various missing RF modalities scenarios. This design demonstrated improved performance in the detection of intra-modal information in the single-modal case, as well as the capturing of correlations and complementary information between modalities in the multimodal case.

The performance of our RFusion system is evaluated on two public datasets XRF55 [28] and MM-Fi [29]. The main contributions of RFusion are summarized as follows.

- We present RFusion, a multimodal RF representation learning framework that learns from unlabeled RF data and supports few-shot human activity recognition, while maintaining robust performance under diverse modality-missing configurations.
- We design an RF-tailored Guider–Arbiter pre-training scheme that couples alignment-aware guidance with unsupervised, entropy-based modality weighting, so that shared information is extracted across RF modalities while preserving their unique RF signatures.
- We validate the effectiveness of RFusion with extensive experiments. RFusion achieves average accuracy gains of 25.8% over supervised unimodal methods and 12.2% over contrastive learning baselines, demonstrating clear advantages in modality-missing scenarios.

II. Preliminary

In this section, we first present the sensing principles of common RF modalities. Then, we conduct a set of benchmarks to motivate the design of RFusion.

A. Basics of Different RF Modalities

WiFi. WiFi sensing techniques typically adopt Channel State Information (CSI) to depict the channel frequency response (CFR) of the propagation paths [30]–[34], which can be expressed as:

$$H = (H(f_1), H(f_2), ..., H(f_K)), \tag{1}$$

where $H(f_k)$ is the CFR for subcarrier k, and K is total number of subcarriers. For a MIMO WiFi system, the measured CSI for each data packet is a $N \times M \times K$ complex 3D matrix, where N and M represent the number of transmitting and receiving antennas, respectively. Note that $H(f_k) = \|H(f_k)\| e^{j \cdot \angle H(f_k)}$, where $\|H(f_k)\|$ and $\angle H(f_k)$ denote the amplitude and phase of the propagation paths, respectively. Both CSI phase and amplitude information can be used to sense human activities.

RFID. An RFID sensing system usually comprises an RFID reader and several passive RFID tags. By emitting a Continuous Wave (CW) periodic signal from the reader antenna, the passive tag is activated and modulates its data on the backscatter signals using ON-OFF keying. Finally, the reader employs the Low-Level Reader Protocol (LLRP) [35] to extract the phase readings from the received signals for sensing human activities. Generally, the received signal can be expressed as:

$$s(t) = a_S e^{j\varphi_S} + \sum_q a_q e^{j(\frac{2\pi}{\lambda} \int v_q(t)dt + \varphi_{dev})}, \tag{2}$$

where q denotes the number of reflection paths of the sensing target, $a_S e^{j\varphi_S}$ is the complex signal of the LoS path and multipath, a_q is the amplitude of reflection, $v_q(t)$ is the length change speed of the q-th path, φ_{dev} is the phase offset bias of the device. By analyzing the variations of the phase and amplitude readings, we can infer human activities.

mmWave Radar. The most commonly used mmWave radar is FMCW radar, which obtains the distance, speed and angle information of human activities by processing the received and

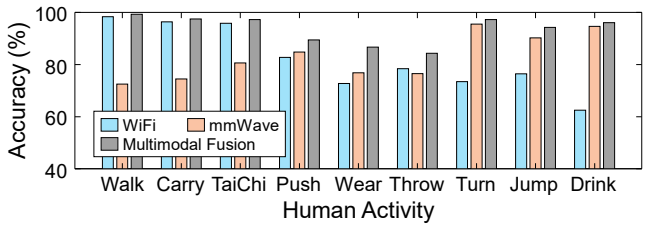


Figure 2: The improvement of multimodality over unimodality.

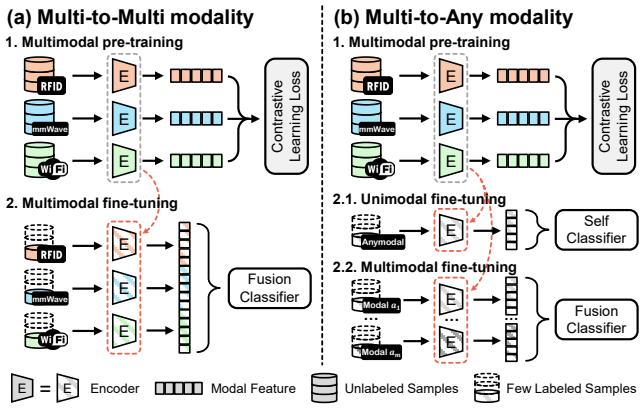


Figure 3: The existing solution is the "Multi-to-Multi" mode, but the "Multi-to-Any " [1] mode exists in the RF sensing scenarios.

Table I: Performance of different numbers of fine-tuning modalities on XRF55 dataset (Scene 3) .

Train Mode	Uni-modality	WiFi + RFID	RFID + mmWave	WiFi +mmWave	All Modality
WiFi	56.27	65.98	∼	68.10	70.27
RFID	38.80	49.69	47.75	∼	55.50
mmWave	47.56	∼	65.75	67.97	69.56

transmitted frequency modulated continuous wave (FMCW) signals. We represent the IF signal as [36]:

$$S_{IF}(t) = Ae^{-j2\pi\{f_0\tau(t)+\mu t\tau(t)-\mu^2\tau^2(t)/2\}}, \quad (3)$$

where $\tau(t) = (2R_0 + V_t)/c$, R_0 is the distance from the target to the radar, and μ is the slope of the FMCW signal. The original IF data is a four-dimensional matrix of $P \times T \times A \times Q$, where Q denotes the number of sampling frames, P is the number of chirps in each frame, T is the length of the samples, and A is the number of receiving antennas. By performing FFT on each chirp of the original FMCW data, we can obtain the distance matrix. Then, the Doppler velocity matrix can be obtained by performing FFT on the distance matrix along the range bin, and the range angle matrix can be obtained by performing FFT on the velocity matrix along the antenna dimension. A mmWave sensing system often uses range-Doppler heat maps and range-angle heat maps to distinguish different activities.

B. Motivation

In this section, we first verify the benefits of multimodal learning using a real human activity RF dataset. Secondly, we present the limitations of existing multimodal learning methods. Finally, we evaluate the impact of a small number of samples on recognition accuracy.

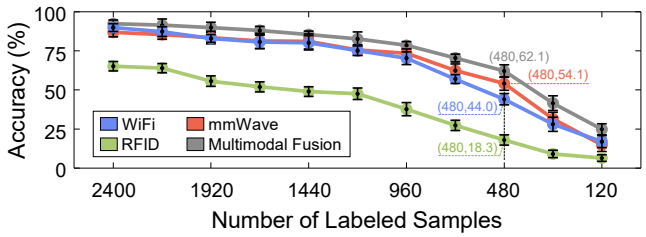


Figure 4: Performance of unimodal and multimodal models with small samples.

Benefits of Multimodality. Different RF modalities have different wavelengths, protocols, and hardware, and thus have different capabilities to represent human activities. We now conduct a benchmark to showcase this phenomenon by selecting two modalities, i.e., WiFi and mmWave, and nine activities on the XRF55 dataset [28]. Then, we build a classical 5-layer CNN for evaluation. The result is plotted in Fig. 2. We can clearly see that different modalities have different performances in different activities. In particular, the WiFi performs better for recognizing large-scale movements (e.g., walk, carry, and TaiChi), while the mmWave radar achieves better accuracy for small-scale movements (e.g., Turn, Jump, and Drink). This is because the mmWave signal has a small wavelength, and thus is more sensitive to subtle and small activities. These results imply each RF modality can provide different strengths in a specific activity sensing task, and we can improve the recognition performance by carefully fusing their information.

Impacts of Modalities. To fuse information from different modalities, existing solutions [27], [37]–[41] hold an assumption: the number of modalities in the pre-training stage is the same as in the fine-tuning stage, as shown in Fig. 3(a). However, this assumption is difficult to guarantee in RF-based sensing scenarios. This is because RF devices are harder to deploy than IMU sensors. In a real-world scenario, there are typically small numbers of RF modalities, like one or two, as shown in Fig. 3(b). This means traditional multimodality learning approaches cannot be directly used in such RF-based sensing cases, and we cannot gain the same performance as all modalities exist. To illustrate it, we conduct a benchmark experiment by varying the number of modalities in the fine-tuning stage from 1 to 3. Table. I presents the results. We can observe that as the number of missing modalities increases, the recognition accuracy decreases. Therefore, it is desired to seek a solution that ensures the sensing system can achieve good performance even when a different number of modalities are missing in the fine-tuning phase.

Impacts of Data Size. Generally, to ensure a wireless sensing system has good performance, the multimodality learning model requires a large number of data for each RF modality and needs online labeling during collection, which is extremely difficult and causes huge human effort. To showcase the impact

[1]In this work, "Multi-to-any" is restricted to any subset of the pre-trained modalities, namely single-modal, partial-modal, and full-modal configurations within this fixed modality set.

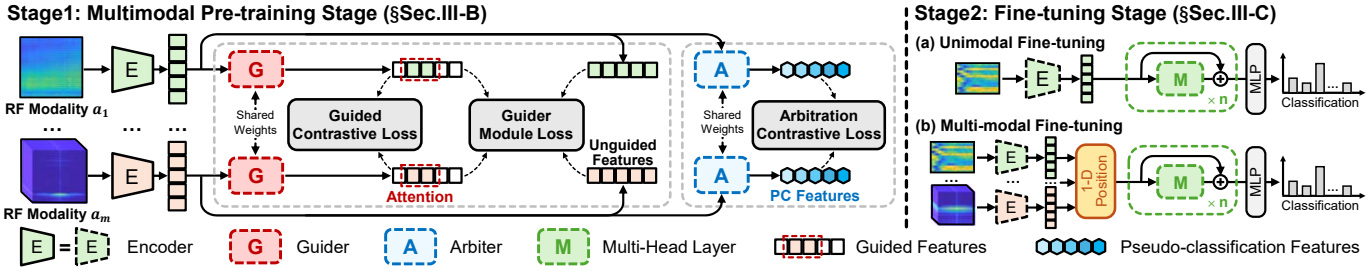
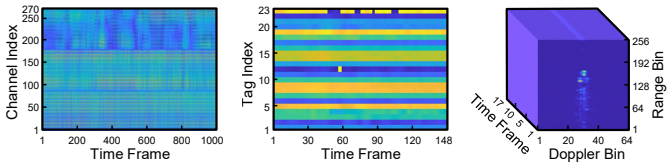


Figure 5: The framework of RFusion consists of multimodal pre-training stage and fine-tuning stage.

of different numbers of training data on the performance, we vary the number of label samples from 2400 to 120 and plot the result in Fig. 4. We can see that when the number of labeled samples is 480, the average accuracy of unimodal recognition is only 38.8%. Although the effect of multimodal fusion is improved compared to unimodal, it is only 62.1% of the recognition accuracy. This means that it is difficult to learn the characteristics of complex human activities using only limited labeled samples.



(a) Signal strength information of CSI for each WiFi channel. (b) Reflection phase information of each RFID tag. (c) Doppler-range heat map of mmWave radar.

Figure 6: Example measurements from different RF modalities, illustrating highly heterogeneous data representations.

III. RFusion System

A. Overview

Problem Formulation: The RFusion system aims to learn knowledge from a large number of unlabeled multimodal samples to enhance fine-tuning performance in missing modality scenarios (including unimodal fine-tuning and multimodal fine-tuning). The dataset of unlabeled pre-trained multimodal samples with M modalities is defined as $\mathcal{D}^M = \{\mathcal{D}^{m_1}, ..., \mathcal{D}^{m_M}\}$, where $\mathcal{D}^{m_i} = \{x_1^{m_i}, ..., x_N^{m_i}\}$ indicates that there are N samples in the m_i modality. The small set of labeled samples used for fine-tuning can be expressed as $\mathcal{D}^U, \mathcal{D}^U = \{\mathcal{D}^{m_{a_1}}, ..., \mathcal{D}^{m_{a_k}}\}$, where $m_{a_k} \subseteq M$, indicating that the modality in the fine-tuning set may be one or more of the pre-training set. Then the dataset of labeled samples for the m_{a_j} modality is $\mathcal{D}^{m_{a_j}} = \{\dot{x}_1^{m_{a_j}}, ..., \dot{x}_T^{m_{a_j}}\}$, where $T \ll N$, because the number of labeled samples in the fine-tuning set is much smaller than the number of unlabeled samples in the pre-training set.

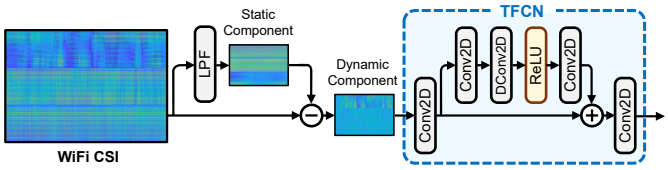
System Architecture: RFusion is a multimodal-assisted wireless sensing framework that leverages unlabeled multimodal data in the pre-training stage to improve the performance in the missing modality scenarios during the fine-tuning phase. The core idea of RFusion is to fully exploit both shared (mutual) and modality-specific (unique) information in multimodal learning, thereby maximizing the utility of multimodal data and boosting performance under limited labeled samples. The system framework is plotted in Fig. 5. In the multimodal pre-training stage, RFusion adopts a novel contrastive learning scheme to train feature encoders on large-scale unlabeled multimodal samples. A Guider module enhances the learning of mutual information and strengthens cross-modal feature alignment via dynamic attention, while an Arbiter module exploits modality-specific information by assigning different attention weights according to the classification performance of pseudo-classification features from each modality, thus achieving strong complementarity across modalities. In the fine-tuning stage, RFusion uses a small amount of labeled samples to further enhance the recognition effect of each modality. For both unimodal and multi-modal fine-tuning modes, we design a multi-head attention network to further enhance the utilization of modality information.

B. RFusion System Pre-training

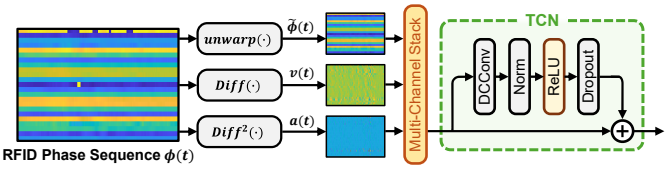
In the multimodal pre-training stage, we introduce a multimodal scheme to extract mutual information and unique information from a large number of unlabeled multimodal data. This stage consists of a feature encoder module and a contrastive learning module.

1) Modality-specific Feature Encoder: As mentioned in Sec. II, the extracted sensing measurements from different RF modalities are often diverse and non-uniform. Fig. 6 shows the raw measurements for the same human activity by using three different RF modalities. We can see that different modalities have different data dimensions. For example, Fig. 6(a) plots the WiFi raw amplitude readings sampled from 270 subcarriers, Fig. 6(b) shows the RFID phase data from 23 tags, and the radar data, plotted in Fig. 6(c), comprises a 3D Doppler-range heat map. This means we need to unify the dimensions of features from different RF modalities before performing contrastive learning. Therefore, we need to design modality-specific feature encoders tailored to the sampling structures and noise characteristics of different RF modalities to extract structured representations from heterogeneous raw RF measurements. The detailed design of each modality-specific feature encoder is shown in Fig. 7.

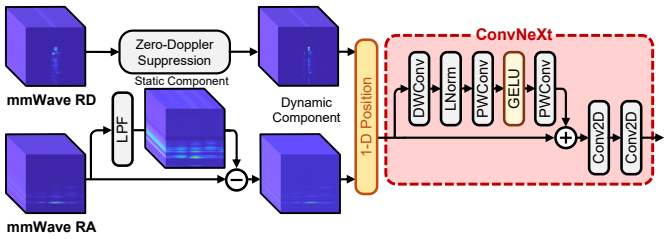
WiFi Encoder. For the WiFi modality, we adopt a time–frequency convolutional network (TFCN) to capture the



(a) WiFi Encoder.



(b) RFID Encoder.



(c) mmWave Encoder.

Figure 7: Feature encoder for each RF modality.

coupled temporal–spectral structure of CSI across subcarriers. Since CSI is often dominated by static reflections, we first remove the background component from the raw signal via background subtraction to emphasize motion-induced variations. Then, TFCN applies convolutions along both time and frequency to jointly model temporal dynamics and inter-subcarrier correlations, while dilated convolutions enlarge the temporal receptive field to efficiently capture long-range activity patterns without increasing network depth.

RFID Encoder. For the RFID modality, we adopt a temporal convolutional network (TCN) to model the one-dimensional temporal evolution of RFID phase. Since passive backscatter signals are low-dimensional and highly sensitive to multipath, we construct multi-channel dynamic inputs from phase differences and their higher-order temporal variations (e.g., velocity and acceleration) to suppress static components and emphasize motion-induced variations. The TCN is built with dilated causal convolutions to enlarge the temporal receptive field while preserving causality, enabling efficient capture of long-range, subtle activity-related phase dependencies without increasing network depth.

mmWave Encoder. For the mmWave modality, we use ConvNeXt to encode the concatenated range–Doppler (RD) and range–angle (RA) heatmaps, which capture motion dynamics and spatial distributions. To remove non-informative static reflections, we extract dynamic heatmaps by suppressing the Zero-Doppler component in RD and subtracting a low-frequency background from RA. ConvNeXt then performs hierarchical spatial feature extraction with efficient channel mixing, providing strong spatial modeling for mmWave heatmaps with a lightweight fully convolutional design.

To summarize the above encoder designs in a unified

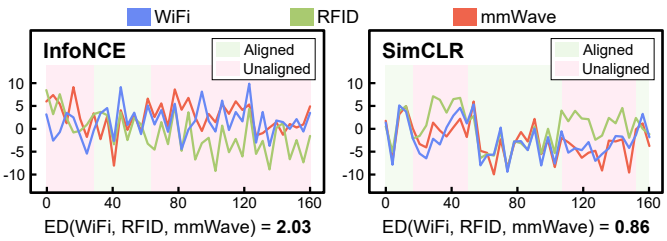


Figure 8: Illustration of feature spaces using existing methods.

formulation, suppose that each original sample is denoted as o and contains M RF modalities. Let $\tilde{o}_n^m$, $m \in \{1, 2, \ldots, M\}$, represent the preprocessed measurement of the m-th modality in the n-th sample after static background suppression. The corresponding feature representation x_n^m is obtained as

$$x_n^m = P^m\big(E^m(\tilde{o}_n^m)\big), \quad m \in \{1, 2, \ldots, M\}, \tag{4}$$

where $E^m(\cdot)$ denotes the modality-specific feature encoder instantiated by different architectures for different RF modalities, and $P^m(\cdot)$ represents the corresponding modality-specific projection network that maps encoded features into a unified embedding space.

2) Cross-modal Contrastive Learning: After obtaining the features from different RF modalities, our next goal is to learn the mutual and unique information knowledge from a large number of easily available unlabeled multimodal samples. To do this, prior methods employ advanced contrastive learning methods [42], [43]. However, we find that these methods have certain problems. As shown in Fig. 8, after contrastive learning, there are still some misaligned intervals in the feature vectors between modalities, which means that there is still a wealth of mutual information between modalities that have not been learned, making them have limited performance. Besides, these contrastive learning methods overemphasize the consistency features and ignore the unique features of different RF modalities. To quantitatively characterize such misalignment, we define the *degree of alignment* between two modality-specific features. Given the features $x_n^{m_1}$ and $x_n^{m_2}$ extracted from modalities m_1 and m_2 for the same sample n, we first apply ℓ_2 normalization:

$$\bar{x}_n^m = \frac{x_n^m}{\|x_n^m\|_2}. \tag{5}$$

We then compute the Euclidean distance between the normalized features:

$$ED_n^{m_1,m_2} = \|\bar{x}_n^{m_1} - \bar{x}_n^{m_2}\|_2. \tag{6}$$

Finally, the distance is converted into a soft alignment score through a Gaussian kernel:

$$Align_n^{m_1,m_2} = \exp\left(-\frac{(ED_n^{m_1,m_2})^2}{2\sigma^2}\right), \tag{7}$$

where a larger value indicates a higher degree of alignment between the two modality-specific feature representations. Hence, to better extract the mutual and unique features from diverse RF modalities, we respectively design a guided contrastive scheme and a pseudo-classification arbitration scheme.

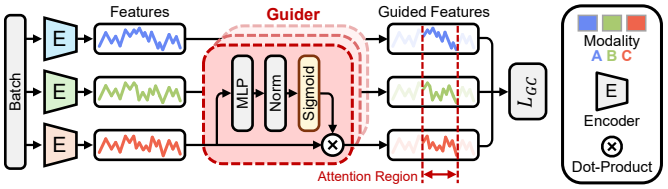


Figure 9: Design of Guider module.

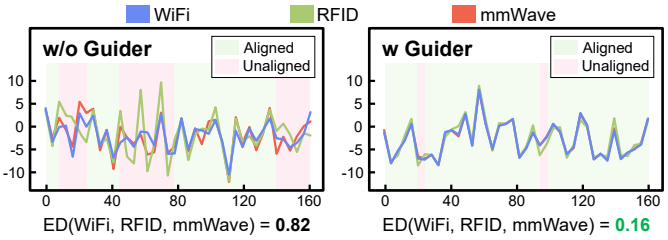


Figure 10: Guided feature alignment representation in feature space.

Guided Contrastive Learning. To extract consistency features from different RF modalities, our core idea is to design a guider module, which evaluates the alignment effect of the inter-modal feature vectors in real-time during the training process and dynamically transfers more attention from the well-aligned intervals to the poorly aligned intervals to strengthen the learning of mutual information during the pre-training process. The application framework of the guider module is shown in Fig. 9. The guider model includes two parts: the MLP network and the scoring layer. The feature vector x_i^m of each modality will be subjected to dynamic attention after passing through the guider module, and finally become the attention feature z_i^m:

$$z_i^m = Sigmoid(Norm(MLP(x_i^m))) \cdot x_i^m. \tag{8}$$

To maximize the similarity of positive pairs and minimize the similarity of negative pairs, we use the cosine distance between the features within a pair to evaluate the similarity of pairs. Then the similarity between any two samples $x_{i_1}^{m_1}$ and $x_{i_2}^{m_2}$ can be expressed as:

$$s(x_{i_1}^{m_1}, x_{i_2}^{m_2}) = \frac{(x_{i_1}^{m_1}) \cdot (x_{i_2}^{m_2})^\top}{\|x_{i_1}^{m_1}\| \cdot \|x_{i_2}^{m_2}\|} \cdot \frac{1}{t_{sim}}, \tag{9}$$

where $i_1, i_2 \in [1, N]$ represent the sample number, and $m_1, m_2 \in [1, M]$ represent the modality under the sample. t_{sim} is an adjustable temperature parameter, properly adjusting t_{sim} value can lead to improved training performance (in RFusion, t_{sim} is set to 0.07 by default).

Next, we consider the feature vectors between different modalities of the same sample as positive pairs, and the feature vectors of different samples in the batch, whether they are in the same modality or between different modalities, as negative pairs. Assuming we have a batch of data of size N, then each sample between two modalities under the RFusion method will generate 2 positive pairs and $2N-2$ negative pairs. Then, for the ith sample, the contrastive learning loss from the m_1 modality to the m_2 modality can be expressed as:

$$lg_i^{m_1,m_2} = CL(z_i^{m_1}, z_i^{m_2}) = -log \frac{exp(s(z_i^{m_1}, z_i^{m_2}))}{\sum_{j=1}^{N} [exp(s(z_i^{m_1}, z_j^{m_2})) + exp(s(z_j^{m_1}, z_i^{m_2}))]}. \tag{10}$$

Note that the weights between these guider modules are shared because they apply a unified attention score to the features of different modalities on the same sample. Then, when facing multiple modalities for joint contrastive learning (for example, M modalities), the final multimodal guided contrastive loss L_{GC} is:

$$L_{GC} = \mathcal{L}(lg, kg) = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{m_1 < m_2} [kg_{m_1} lg_i^{m_1,m_2} + kg_{m_2} lg_i^{m_2,m_1}], \tag{11}$$

where lg denotes the collection of guided contrastive loss terms, and kg is the corresponding weighting vector used for the guided contrastive learning. Since we only focus on the utilization of mutual information between modalities in the role of the guider, we set the vector of kg to 0.5 by default, that is, the same weight is applied to each modality. Finally, we evaluate the alignment effect of feature vectors under guided contrastive learning in the pre-training stage. As plotted in Fig. 10, after adding the guider module, the inter-modal feature vectors under our method are highly aligned. With the efficient use of mutual information, the clustering effect between modalities has been significantly improved after guidance (the distance between modalities has been reduced by an average of 5.4 times).

Dynamic Attention in Guiders. The previous section introduced how the guider module strengthens the training of feature vectors by dynamically allocating attention, but a question that follows is: how to make the guider clear about how to allocate attention? To address this, we observe that although the unguided contrastive loss L_{Base} is not explicitly optimized during guided contrastive learning, it consistently decreases as training proceeds. Importantly, L_{Base} typically lags behind the guided contrastive loss L_{GC}, indicating that cross-modal alignment is primarily enforced by the Guider rather than being fully absorbed by the encoder itself.

Based on this observation, we introduce an adversarial training strategy for the Guider module. Specifically, the Guider is optimized to enlarge the discrepancy between L_{GC} and L_{Base}, while the encoder parameters are optimized only through L_{GC}. This adversarial objective encourages the encoder to progressively internalize the alignment capability emphasized by the Guider. With continuous training, L_{Base} progressively approaches L_{GC}, indicating that the encoder itself has sufficiently learned the cross-modal mutual information emphasized by the Guider. As training proceeds, the gap between L_{Base} and L_{GC} gradually shrinks, indicating that the encoder has internalized the cross-modal alignment previously enforced by the Guider. As a result, the Guider can be safely removed during fine-tuning. Specifically, for

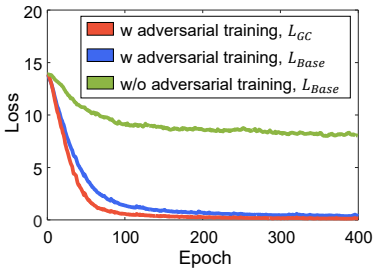


Figure 11: Impact of adversarial training on the Guider module.

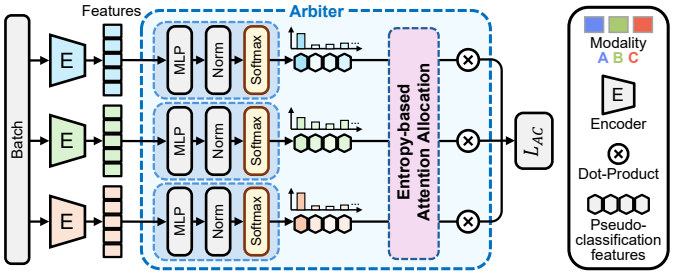


Figure 12: Design of Arbiter module.

unsupervised samples, the unguided contrastive learning loss from the m_1 modality to the m_2 modality is:

$$l_i^{m_1,m_2} = CL(x_i^{m_1}, x_i^{m_2}). \tag{12}$$

Keeping it consistent, the basic contrastive learning loss L_{Base} in a batch is:

$$L_{Base} = \mathcal{L}(l, k), \tag{13}$$

where l denotes the collection of unguided contrastive loss terms, and k is the corresponding weighting vector used for the unguided contrastive learning. Therefore, we design the loss of the guider module as follows:

$$L_{GM} = exp(\frac{L_{Base}}{L_{GC}} \cdot t_g), \tag{14}$$

where t_g is the temperature parameter. In the pre-training stage, the training of the guider module is carried out together with the guided contrastive learning. In Eqn. 14, we use $exp(\cdot)$ to increase the gradient of the guider module training. Among them, t_g is the temperature parameter to improve the training effect. We set it to 0.8 by default. In addition, the weighting vector k used in L_{Base} is consistent with the guided weighting vector k_g in L_{GC}. We set them to 0.5 to ensure that the same weight is applied between modalities.

To quantify the effect of adversarial training, we run an ablation study with and without it. As shown in Fig. 11, without adversarial training, L_{GC} decreases much faster than L_{Base}, indicating that cross-modal alignment is largely achieved by the Guider while the encoder struggles to learn it on its own. In contrast, adversarial training compels the encoder to narrow this gap, effectively distilling the Guider's alignment capability into the encoder.

Pseudo-classification Arbitration. According to our observations in Sec. II-B, in addition to strengthening the learning of mutual information between modalities, we should also make good use of the unique information of different modalities. Unfortunately, we cannot directly evaluate the complementary performance of each modality during the unlabeled pre-training stage, and thus cannot directly learn the unique information of modalities. In this regard, our core idea is to design an arbiter module which first pseudo-classifies the feature vector obtained from each modality to obtain pseudo-classification features and then calculates the entropy of pseudo-classification features of different modalities to judge the classification performance of each modality for a certain sample. Finally, a relative attention score is given to the unique information of each modality based on this performance, and the score will eventually be included in the calculation of the contrastive arbitration loss. Fig. 12 shows the working process of the arbiter module.

Specifically, the arbiter module consists of two parts: pseudo-classification feature extraction and entropy-based attention allocation. For the previous part, we use the "MLP-Norm-Softmax" network to obtain the probability distribution of all possible classification results of the input feature x_i^m, which we call the pseudo-classification feature $\hat{x}_i^m, \hat{x}_i^m \in [0, 1]$. In the attention allocation part, we calculate the entropy of each pseudo-classification feature. Suppose that the total number of classification results is C, then $\hat{x}_i^m = [p_1, p_2, ..., p_C]$, the entropy calculation formula for the pseudo-classification feature $\hat{x}_i^m$ of the mth modality of the ith sample is:

$$H(\hat{x}_i^m) = -\sum_{c=1}^{C} p_c log(p_c). \tag{15}$$

The entropy of each modality represents the degree of confusion of the pseudo-classification results of the modality. The larger the entropy, the more ambiguous the classification result, and vice versa. The smaller the entropy, the clearer the classification result. Therefore, for any sample, we do not impose too many constraints on those modalities with smaller entropy. On the contrary, for those modalities with larger entropy, we hope that they will learn from the modalities with smaller entropy. Then for a certain modality m_1, its entropy-based attention is:

$$ka_{m_1} = \frac{exp(H(\hat{x}^{m_1}))}{\sum_{m=1}^{M} exp(H(\hat{x}^m))}. \tag{16}$$

Next, we construct the contrastive learning loss between the two modalities of pseudo classification features based on L_{Base}:

$$la_i^{m_1,m_2} = CL(\hat{x}_i^{m_1}, \hat{x}_i^{m_2}). \tag{17}$$

Then, the arbitration contrastive loss L_{AC} is:

$$L_{AC} = \mathcal{L}(la, ka) \cdot t_a, \tag{18}$$

where ka denotes the entropy-based attention vector, t_a is a temperature coefficient.

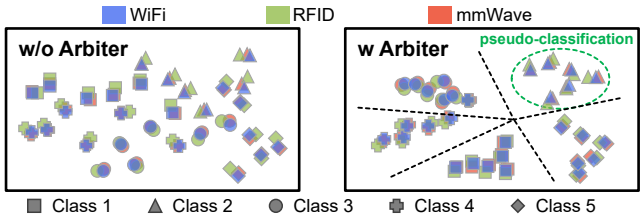
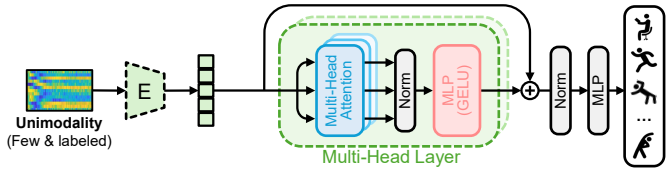
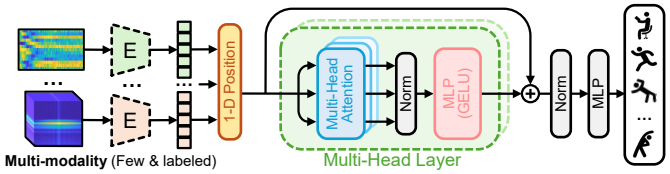


Figure 13: Inter-modality pseudo-classification effect under the influence of the arbiter.



(a) Unimodal fine-tuning mode.



(b) Multimodal fine-tuning mode.

Figure 14: Multi-head attention based fine-tuning model.

In the actual pre-training stage, the arbiter module is trained together with contrastive learning. However, we observe that at the early stage of contrastive learning, the pseudo-classification feature extractor is not sufficiently trained, so the arbiter module cannot effectively perform arbitration. Therefore, in Eqn. 18, we design t_a with a very small initial value that gradually increases with the number of training rounds, to avoid the phenomenon of incorrect arbitration in the early stage of arbiter module training. Finally, we simulate and evaluate the effectiveness of the arbiter method by randomly selecting six classes from the test set and using tSNE to visualize the feature space. As shown in Fig. 13, with the arbiter module, our method produces clear pseudo clustering in all modalities.

3) Pre-Train Loss: The total loss of RFusion in the pre-training stage consists of the three parts mentioned above: guided contrastive loss, guider module loss, and arbitration contrastive loss. That is:

$$L_{total} = L_{GC} + L_{GM} + L_{AC} \tag{19}$$

We consider that the three parts that make up the pre-training stage are equally important, so the final comprehensive loss only needs to simply add them together. The final experiment also proves the effectiveness of our approach.

C. Fine-tuning to Different Missing-modality Conditions

In the fine-tuning stage, the objective is to leverage a small amount of labeled data to enhance recognition performance. However, in practical deployment, the number of available RF modalities often differs from that in the pre-training stage, resulting in missing-modality scenarios. To address this challenge, we design two complementary fine-tuning strategies—unimodal fine-tuning and multimodal fine-tuning—both built upon a multi-head attention–based classification network for effectively exploiting the available unimodal or multimodal information.

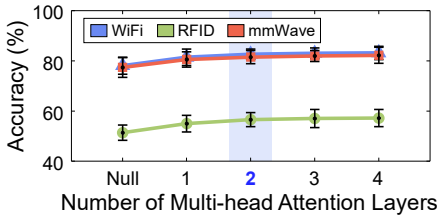


Figure 15: The impact of the number of multi-head attention layers on classification performance.

Specifically, when only a single RF modality (e.g., WiFi, RFID, or mmWave) is available, only the corresponding modality-specific encoder branch is activated, and the extracted feature is directly fed into the classifier, corresponding to unimodal fine-tuning. When multiple modalities are available, their features are concatenated and jointly processed by the classifier, while the encoder branch of any missing modality is simply omitted. This design enables RFusion to flexibly adapt to different modality availability without architectural modification.

We next introduce the design of the unimodal and multimodal fine-tuning mechanisms.

1) Unimodal Fine-tuning Case: In unimodal fine-tuning, we cannot combine information from multiple modalities, for example, we cannot apply attention weights to different modalities like Cosmo [27]. Thus, our idea is to apply multiple different self-attentions to unimodal features to enhance the use of information within the modality. As shown in Fig. 14(a), we introduce a multi-layer multi-head attention classification network behind the unimodal feature encoder, and these networks are connected internally through residuals.

2) Multi-modal Fine-tuning Case: When there are multiple modalities in the scene, our approach is shown in Fig. 14(b). We first concatenate the features of multiple modalities to use multi-head attention to learn the complementarity between modalities. Secondly, we also add a learnable position encoding layer to enable the model to better learn the relevant information between modalities.

Multi-head Attention Network. The multi-head attention is written as:

$$x' = Softmax\left(\frac{Q(x)K^T(x)}{\sqrt{d_k}}\right)V(x) + x \tag{20}$$

where the query layer, key layer, and the value layer are given respectively by Q, K, and V, and input x is the input feature vector. Through this network, we can learn different types of dependencies in the input sequence. Moreover, multi-head attention enhances the model's capability to capture complex

patterns by modeling the input through multiple independent attention weights across different subspaces.

Finally, we simulate the relationship between the number of multi-head attention layers and the fine-tuning effect. As shown in Fig. 15, the increase in the number of multi-head attention layers does improve the fine-tuning effect, but the accuracy improvement gradually becomes less obvious after more than two layers. Therefore, weighing the two factors of fine-tuning effect and model size, we finally chose a multi-head attention layer with two layers of residual connections in the fine-tuning stage.

IV. Experiments

A. Datasets

We use two public multimodal datasets for human activity recognition, XRF55 and MM-Fi, to evaluate the performance of RFusion. Each dataset contains four multimodal scenarios, covering common RF modalities such as WiFi, RFID, and mmWave. We show the information summary of each scenario in the dataset in Table. II.

XRF55 [28] dataset contains multimodal data of 55 human activity categories collected from 39 individuals. The dataset provides three RF modalities: WiFi, RFID, and mmWave. Among them, the WiFi modal samples come from the CSI of a total of 9 WiFi links under three WiFi receivers, and each link sends an OFDM signal of 30 subcarriers. The samples of the RFID modality come from the backscattered phase information of 23 RFID tags collected by the RFID reader. The samples of the mmWave modality are the distance-Doppler heat map and distance-angle heat map obtained by processing the FMCW signal, and the two are spliced along the dimensional range. The XRF55 dataset contains four scenes in total. For each scene, we divide 70% of the data into pre-training datasets and the remaining data into testsets.

MM-Fi [29] dataset consists of multimodal data of 27 human activities collected from 40 individuals, including two RF modalities: WiFi and mmWave. WiFi modal samples come from the CSI of three WiFi channels after smoothing and filtering, and each WiFi channel sends a signal composed of 114 subcarriers. The mmWave modal samples are point cloud data of position-Doppler velocity and signal strength, in which the number of sampling points is increased by frame aggregation. Notably, the provided mmWave point clouds have already undergone static component suppression during dataset preprocessing, and thus no additional operation is required. The dataset contains four scenes. For each scene, we use approximately 70% of the data for pre-training, with the remaining data reserved for testing.

Table II: Dataset summary.

Dataset	Scene	Modality	#Activities	#Train Samples	#Test Samples
XRF55	Scene1	WiFi, RFID, mmWave	55	23100	9900
	Scene2			2310	990
	Scene3			2310	990
	Scene4			2310	990
MM-Fi	Scene1	WiFi, mmWave	27	3219	1284
	Scene2			2767	1168
	Scene3			2926	1201
	Scene4			2736	1147

B. Baseline Methods

To evaluate the performance of RFusion, we selected six advanced unsupervised multimodal contrastive learning methods as baselines. To ensure a fair comparison under the RF few-shot and missing-modality setting, all baseline methods follow the same training protocol as RFusion: they are first pre-trained in a multimodal manner using all available RF modalities, and then fine-tuned under a few-shot setting where only the modalities available in the target scenario are activated. In the fine-tuning stage of all baselines, we used the Multimodal-GRU classifier model consistent with Cosmo. The information of all baselines is shown below.

Cosmo [27] is an advanced multimodal human activity recognition method based on contrastive learning. It uses a quality-oriented attention fine-tuning mechanism and a feature fusion contrastive learning method, which can effectively learn the consistency and complementarity between modalities. In the experiment, except for canceling the attention guidance mechanism during single-modal fine-tuning, the other parts retain the design of Cosmo unchanged.

CMC [37] is a contrastive learning method designed for computer multi-view vision tasks, which directly compares features from different modalities of unlabeled data to train feature encoders. In the experiment, we keep the design of CMC unchanged.

CC [44] is an advanced multimodal contrastive learning method that introduces pseudo clustering in the pre-training stage to obtain better clustering effects. In the experiment, we retain the complete design of CC in the pre-training stage.

SimCLR [43] is an excellent contrastive learning method for computer vision. It achieves good recognition results by means of data enhancement and the introduction of nonlinear transformation. In the experiment, we keep the contrastive learning loss calculation method of SimCLR unchanged.

InfoNCE [42] is a basic unsupervised contrastive learning method based on information theory. It trains model parameters by dividing positive and negative samples and comparing their similarities. We also compared the performance of the InfoNCE contrastive learning method in the experiment.

Transformer is a general attention-based model whose representation learning capability is also well suited for multimodal fusion. It employs cross-attention to explicitly model inter-modal interactions and capture complementary information across modalities. In the experiment, modality-specific features are first extracted and then fused through cross-attention layers to obtain a joint representation for downstream recognition.

C. Experiment Setup

RFusion and other baselines are deployed on a Python-based Pytorch platform running on a processor with an AMD Ryzen 7950x (5.7GHz), 64GB memory, and an Nvidia RTX4080 GPU. In both the pre-training and fine-tuning stages, the learning rate is set to 0.001, the default batch size in the pre-training stage is 32, and the default batch size in the fine-tuning stage is set to 16. In Sec. V-B and Sec. V-C, we select Scene 3 of the XRF55 dataset as the default experimental scene for experiments. In all experiments, we independently conduct 5

Table III: Performance on XRF55 dataset.

Dataset	XRF55 (label rate: 24%)													
Scene	Scene1							Scene2						
Available modality	WiFi	RFID	mmWave	WiFi + RFID	RFID + mmWave	WiFi + mmWave	All	WiFi	RFID	mmWave	WiFi + RFID	RFID + mmWave	WiFi + mmWave	All
Transformer	63.86	41.22	75.10	66.95	78.43	82.47	83.57	67.91	30.01	66.06	72.41	70.42	81.11	84.18
CC	64.34	42.19	76.19	66.98	79.64	82.35	84.00	72.84	31.41	71.73	73.66	72.74	82.22	83.64
CMC	64.14	39.46	76.05	67.75	79.48	82.80	83.62	69.29	31.21	66.76	73.87	69.08	80.84	82.19
Cosmo	59.23	41.77	73.57	61.46	76.79	75.65	84.01	72.04	31.58	66.36	59.22	69.97	82.06	85.09
InfoNCE	64.21	33.87	72.25	67.93	78.05	82.66	83.49	68.38	28.88	67.97	72.74	73.38	82.21	83.05
SimCLR	63.89	42.27	75.66	65.06	78.36	82.44	83.92	70.84	30.91	70.02	70.59	71.92	81.35	83.79
Unimodal	63.2	15.52	79.15	-	-	-	-	54.06	14.15	56.04	-	-	-	-
RFusion	**74.18 +9.84**	**50.84 +8.57**	**83.69 +7.50**	**75.68 +7.75**	**86.05 +6.41**	**89.79 +7.13**	**92.31 +8.30**	**83.44 +10.60**	**51.51 +19.93**	**81.57 +9.84**	**87.37 +13.50**	**85.72 +12.34**	**91.48 +9.26**	**93.79 +8.70**

Dataset	XRF55 (label rate: 24%)													
Scene	Scene3							Scene4						
Available modality	WiFi	RFID	mmWave	WiFi + RFID	RFID + mmWave	WiFi + mmWave	All	WiFi	RFID	mmWave	WiFi + RFID	RFID + mmWave	WiFi + mmWave	All
Transformer	73.41	46.59	72.08	80.14	80.33	83.34	86.06	75.06	40.72	72.53	77.95	73.45	82.31	84.21
CC	73.77	51.35	73.25	81.78	79.09	82.47	85.64	77.35	38.52	72.74	75.81	71.10	84.11	84.57
CMC	74.48	50.51	68.85	81.55	73.77	83.09	84.56	75.92	40.88	70.18	74.48	70.69	81.45	83.98
Cosmo	72.64	47.25	71.61	78.03	72.95	84.11	87.56	77.35	40.90	73.15	77.86	73.51	82.78	84.64
InfoNCE	73.56	45.49	71.41	80.12	77.56	84.02	84.37	74.89	41.13	73.15	76.94	74.81	83.09	83.68
SimCLR	74.34	48.69	72.84	81.84	78.84	84.13	85.20	78.76	42.11	73.84	78.17	70.49	84.37	84.56
Unimodal	55.83	33.02	60.52	-	-	-	-	48.43	16.52	60.93	-	-	-	-
RFusion	**85.69 +11.31**	**63.98 +12.63**	**84.01 +10.76**	**90.83 +8.99**	**89.01 +8.68**	**94.24 +10.11**	**96.88 +9.32**	**88.07 +9.31**	**52.22 +10.11**	**84.29 +10.45**	**88.51 +10.34**	**84.95 +10.14**	**92.88 +8.51**	**93.76 +9.12**

Table IV: Performance on MM-Fi dataset.

Dataset	MM-Fi (label rate: 24%)											
Scene	Scene1			Scene2			Scene3			Scene4		
Available modality	WiFi	mmWave	All	WiFi	mmWave	All	WiFi	mmWave	All	WiFi	mmWave	All
Transformer	53.12	75.93	82.15	53.89	70.74	76.21	51.33	74.57	80.33	56.73	76.53	82.91
CC	53.35	77.96	78.29	56.67	70.46	73.26	56.41	74.50	75.39	56.07	77.11	80.01
CMC	53.98	75.78	78.32	56.25	70.20	74.56	57.33	73.83	76.27	48.94	75.88	79.22
Cosmo	54.06	78.14	82.21	55.39	71.88	76.48	51.00	75.08	81.12	59.29	78.16	83.21
InfoNCE	59.84	74.37	75.68	49.05	70.63	71.15	45.50	71.75	72.68	53.16	74.47	76.68
SimCLR	51.32	74.84	76.32	50.94	69.01	72.49	56.33	70.25	76.24	52.46	74.91	78.97
Unimodal	57.73	70.84	-	39.21	57.87	-	39.66	57.41	-	44.36	63.29	-
RFusion	**76.33 +16.49**	**86.57 +8.43**	**92.81 +10.60**	**67.48 +10.81**	**79.82 +7.94**	**85.12 +8.64**	**65.07 +7.74**	**84.23 +9.15**	**89.33 +8.21**	**71.37 +12.08**	**88.38 +10.22**	**93.12 +9.91**

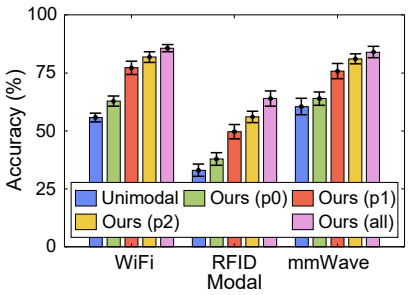
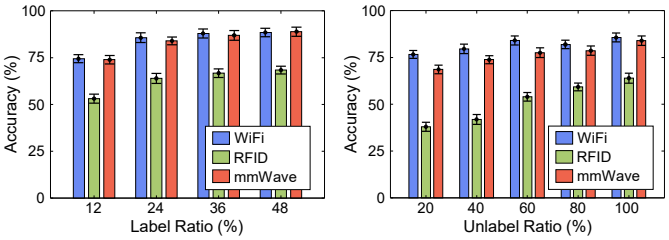


Figure 16: Ablation experiment of RFusion.



(a) Impact of labeled sample ratio. (b) Impact of unlabeled sample ratio.

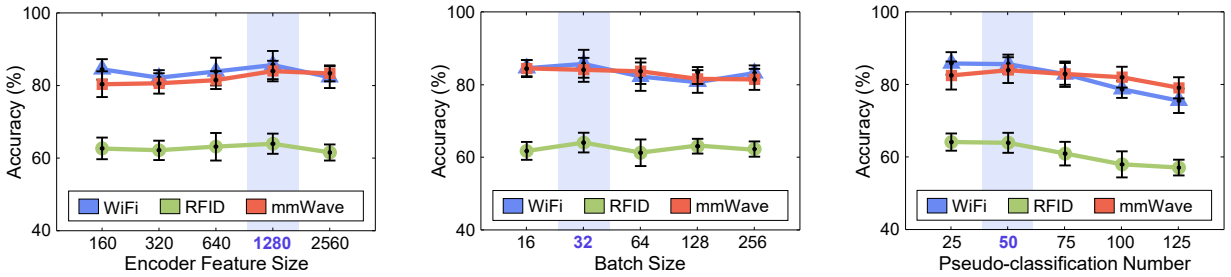
Figure 17: Impact of sample labelling ratio on pre-training and fine-tuning.

tests and record the TOP1 accuracy of each measurement result to evaluate the performance of RFusion and other baselines.

V. Evaluation

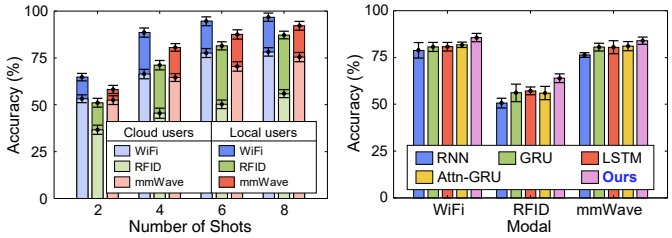
A. Overall Performance

Due to the scarcity of labeled data in practical human activity recognition, we fine-tune all methods using only 24% labeled samples. Table III (XRF55) and Table IV (MM-Fi) report results under different modality-availability settings, including unimodal, partial-modality, and full-modality (All) configurations, which explicitly evaluates performance when one or more modalities are missing at deployment time. Overall, baseline methods show noticeable performance fluctuations across different available-modality settings, and their accuracy drops more evidently in unimodal or partial-modality cases. In contrast, RFusion consistently achieves the best performance across all modality configurations on both datasets. Across datasets and modality configurations, RFusion achieves an average Top-1 accuracy of 82.41%, outperforming InfoNCE, Cosmo, CMC, SimCLR, CC, and Transformer by 12.95%, 12.18%, 12.37%, 12.07%, 11.28%, and 11.75%, respectively, and exceeding the supervised unimodal baseline by 25.79%. These results confirm that RFusion is particularly effective for missing-modality scenarios.



(a) Impact of feature size parameters. (b) Impact of batch size parameters. (c) Impact of pseudo classification number.

Figure 18: Impact of different training parameters settings.



(a) Impact of labeled sample source. (b) Performance of multi-head attention layers.

Figure 19: Impact of fine-tuning data and model design on RFusion performance.

B. Ablation Studies

The contribution of RFusion mainly consists of three parts: the guider module, the arbiter module, and the multi-head attention network. In order to evaluate the performance of each part of RFusion, we conduct an ablation experiment in the default scenario. We additionally set up a supervised unimodal method, Ours (p0) equipped with a temporal multi-head attention module with the same parameter size as the Guider module, Ours (p1) with only the guider module, and Ours (p2) with both the guider module and the arbiter module for comparison. As shown in Fig. 16, each part of RFusion has a certain contribution. In the pre-training stage, Ours (p1) improves over Ours (p0) by an average of 12.66%, and Ours (p2) further improves over Ours (p1) by 5.41%. After adding the multi-head attention network, the fine-tuning effect of RFusion improved by an average of 4.3%. These results demonstrate the effectiveness of our proposed schemes.

C. Diverse Factors on RFusion

In this subsection, we investigate the factors affecting the system performance.

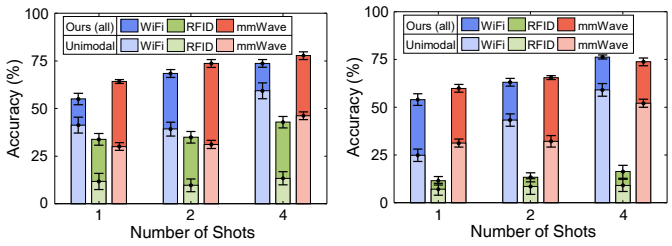
1) Impact of Labeled Sample Ratio: We use different numbers of labeled samples to evaluate the impact of the scale of labeled samples on RFusion. We use the data used in the pre-training stage as the source of labeled samples for fine-tuning. Specifically, we take 12% to 48% of the data, label it, and apply it to the fine-tuning stage. We conducted experiments in the default scenario. As shown in Fig. 17(a), with the increase in the number of labeled samples, the performance of all modalities under the RFusion method shows a significant increase. When RFusion has 24% of the labeled data, the average accuracy of the modality reached 72.17%, which is enough to prove that our method can effectively reduce the need for a large amount of labeled data.

2) Impact of Unlabeled Sample Ratio: Since RFusion needs to use a large amount of unsupervised data for pre-training, the amount of unsupervised data in the pre-training stage is crucial for the RFusion system. We use 20% to 100% of the training data in the default scenario for pre-training, and then use 24% of the pre-training data for fine-tuning testing in the fine-tuning stage. As shown in Fig. 17(b), the performance of RFusion improves as the size of pre-training data increases, which shows that RFusion can extract more useful information from a large amount of pre-training data.

3) Impact of Different Training Parameters Settings: To evaluate the robustness of the RFusion system to training parameters, we showcase the performance of RFusion in the default scenario with different contrast feature sizes, different numbers of pseudo-classifications, and different pre-training batch size parameters. As shown in Fig. 18(a) and Fig. 18(b), the RFusion system is robust to both contrast feature size and training batch parameter size during the pre-training stage. However, in Fig. 18(c), it is more sensitive to the pseudo-classification number parameter. This is because when the number of pseudo-classifications does not match the actual number of classes in the data set, it will mislead RFusion's arbitration performance in contrastive learning. Notably, the pseudo-classification number only controls the grouping capacity of the arbitration module and does not rely on any ground-truth labels, so the pre-training objective remains unsupervised. In practice, this parameter can be set conservatively or tuned during the few-shot fine-tuning stage without affecting the unsupervised pre-training formulation.

4) Impact of Labeled Sample Source: During fine-tuning, RFusion relies on a small amount of labeled data. In practice, these labels may come from different sources, e.g., "cloud users" (subjects seen during training) or "local users" (subjects encountered at deployment). To study the impact of label source, we compare its performance under these two sources with varying amounts of fine-tuning data. As shown in Fig. 19(a), RFusion yields better performance using local data for fine-tuning. For example, with only 4 labeled samples per class, RFusion reaches an average recognition accuracy of 80.12%.

5) Impact of Fine-tuning Model: To assess the impact of the fine-tuning classifier, we replace RFusion's multi-



(a) Performance of Cross-person. (b) Performance of Cross-scene.

Figure 20: Performance of cross different domains.

head attention classifier with several commonly used alternatives, including RNN, GRU, LSTM, and Attn-GRU (as in Cosmo [27]), while keeping all other settings unchanged. As shown in Fig. 19(b), the multi-head attention classifier yields the best results and improves the average single-modality accuracy by 4.1% over the alternatives, indicating that it is more effective for unimodal deployment.

6) Performance of Cross different domains: We now evaluate the performance of RFusion when crossing different domains.

Cross-person. To evaluate cross-person performance, we pre-trained the model using all samples from the first 29 subjects in Scene 1 of the XRF55 dataset and subsequently tested it on all samples from the remaining one subject who was not seen during pre-training. During the test, we used 1/2/4 test samples of each action from the remaining 1 subject to fine-tune the pre-trained model, with the remaining samples serving as the final test set. We compared the performance of using our method and not using our method(i.e., a supervised unimodal baseline). As shown in Fig. 20(a), under the 4-shot fine-tuning stage, the average recognition accuracies of 73.67% (WiFi), 42.90% (RFID), and 77.82% (mmWave) for the 3 modals, respectively. Compared with the overall performance without cross-person (Table III), the accuracy drop is at most 7.94%, whereas without our method, the baseline suffers a much larger drop of up to 31.57%. This indicates that RFusion adapts effectively to cross-person scenarios.

Cross-scene. To further evaluate the performance of RFusion in both cross-environment and cross-person tasks, we used all samples from the 30 subjects in Scene 1 for pre-training and all samples from the 3 subjects in Scene 2 as the test set. The fine-tuning stage was similar to that described in the Cross-person evaluation, where 1/2/4 test samples of each action from each subject were used for fine-tuning. As shown in Fig. 20(b), under the 4-shot fine-tuning stage, the average recognition accuracies were 76.25% (WiFi), 16.33% (RFID), and 73.82% (mmWave). The RFID modality performs the worst, consistent with the recent work [28]. This is mainly due to the intrinsic limitations of passive backscatter sensing: RFID measurements are low-dimensional with limited spatial–temporal resolution, and are highly sensitive to tag orientation, antenna coupling, and multipath distortion. Even so, our approach still delivers a maximum average-accuracy improvement of 21.71% over not using our method, demonstrating that RFusion remains effective in more complex cross-scene settings.

Table V: Computational and memory overhead of RFusion under different modality settings.

Available modality	Params	Occupy memory	CPU Inference Time	GPU Inference Time
WiFi	9.87 M	173 MB	4.59 ms	1.50 ms
RFID	8.37 M	167 MB	4.07 ms	1.72 ms
mmWave	11.38 M	307 MB	6.30 ms	1.64 ms
WiFi+RFID	11.85 M	183 MB	6.61 ms	2.08 ms
RFID+mmWave	13.56 M	317 MB	8.11 ms	2.19 ms
WiFi+mmWave	15.62 M	324 MB	9.46 ms	1.94 ms
All	18.27 M	332 MB	12.41 ms	2.50 ms

D. Efficiency and Deployment Analysis

In this experiment, we evaluate the deployment efficiency of RFusion under different modality configurations by reporting its parameter count, memory footprint, and per-sample inference latency on CPU and GPU. The results are shown in Table V. Overall, RFusion achieves millisecond-level inference across all settings, with 4.07 to 12.41 ms on CPU and 1.50 to 2.50 ms on GPU. The model complexity grows approximately linearly with the number of available modalities, while the quadratic cost introduced by attention-based fusion remains negligible in practice due to the short latent sequence length. These results indicate that RFusion maintains a favorable balance between recognition performance and computational efficiency, supporting practical real-time deployment.

VI. Related Works

In this section, we introduce the prior works that are closely related to our RFusion.

A. RF-based Human Activity Recognition

Owing to their contact-free nature and lack of privacy concerns, various radio frequency (RF) sensing technologies have been widely adopted for human activity recognition. They include WiFi [30]–[34], RFID [45]–[47], and mmWave [48]–[51]. For example, WiLDAR [52] designs a lightweight HAR system that can extract information from Wi-Fi CSI without additional complex operations. TSCNN [53] leverages the time-space convolutional network to improve the recognition accuracy of RFID HAR and realize the generation of new data sets. Wang *et al.* [54] propose a millimeter-wave point cloud data enhancement method to improve the generalization effect of the model in different scenarios. Unlike previous works that focus solely on a single RF modality, this paper introduces a multimodal-assisted sensing paradigm. By leveraging unlabeled multimodal RF data during pre-training, we enhance fine-tuning performance under modal-scarce scenarios. This enables RFusion to achieve multimodal-level performance while requiring minimal effort during inference, such as using only a single RF modality.

B. Multimodal-based Human Activity Recognition

Compared with single-modality approaches which tend to have poor robustness, multimodality has been widely studied in perception tasks [21]–[24] due to its advantages through information fusion. DeepSense [55] introduces the idea of interactive learning between sensors to enhance multimodal

training effectiveness. AttnSense [56] proposes an attention-based multimodal fusion mechanism, demonstrating improved performance and interpretability. XRF55 [28] collaboratively combines three RF modalities to enhance sensing performance. Although these multimodal approaches have achieved promising results, acquiring labeled samples remains challenging in practice. Due to the high cost of manual labeling and potential privacy concerns, it is often infeasible to obtain large quantities of multimodal labeled data for training.

To address this issue, efforts have been made to leveraging unlabeled data for human activity recognition [57]–[62]. For example, Cosmo [27] is a two-stage contrastive fusion learning framework, the training of its encoders mostly relies on unlabeled samples and requires only a small number of multimodal labeled data during fine-tuning. MaskFi [26] is a Transformer-based multimodal training method that masks parts of the input during pre-training and utilizes unlabeled WiFi and visual data for HAR. Autoencoder model [63], which is trained on unlabeled samples, has also been explored. It was fine-tuned with limited labeled data to approximate supervised learning performance. While these works reduce the model dependencies on labeled datasets, they are not directly applicable to RF-based HAR tasks. In practice, it is relatively easy to collect large-scale unlabeled multimodal RF data. However, Ma *et al.*'s autoencoder does not fully exploit the rich multimodal information during pre-training. Moreover, due to the diverse deployment scenarios of RF technologies, it is common that only a subset of RF modalities is available during inference. Cosmo, for example, relies on relative attention across multiple modalities during fine-tuning (e.g., accelerometers and IMUs in wearable devices), and MaskFi similarly requires the simultaneous availability of multiple modalities. These approaches assume the full set of modalities is accessible during fine-tuning and do not address scenarios with missing RF modalities. In contrast, RFusion is designed to address the challenges of missing RF modalities while still maintaining the benefits of full-modal performance.

VII. Discussion

In this section, we discuss several limitations and untapped opportunities with RFusion.

Unseen Gesture Recognition. In practical scenarios, undefined or previously unseen gestures frequently occur. However, the current version of RFusion performs well only on gestures it has been trained on, and struggles to effectively recognize unseen gestures. In future work, we can explore how to adapt to new gestures based on the knowledge learned from previously performed gestures.

Multi-user Scenarios. Multi-user sensing remains a well-known challenge in RF sensing, as reflection signals from multiple targets can become mixed at the receiver, causing mutual interference. The current version of RFusion focuses primarily on single-target scenarios. For multi-user cases, we believe that recent advances in multi-dimensional signal processing [64] offer promising opportunities to enable effective multi-user sensing. We leave it as our future work.

Applying to Other Fields. Although we focus on multimodal fusion across RF technologies (WiFi, RFID, and mmWave) for HAR, the proposed framework can be extended to other sensing modalities (e.g., accelerometers and gyroscopes) and applied to broader domains such as healthcare and industrial monitoring. By transferring knowledge from multimodal data, it can help low-cost modalities approach multimodal performance, reducing deployment cost.

Generalizing to unseen modalities. The current version of RFusion does not automatically generalize to completely new sensing modalities that are unseen during pre-training. Supporting a new modality would require adding a corresponding encoder and adapting it to the shared feature space.

Integrating Generative Models. Diffusion-based generative models provide a complementary direction to RFusion: they can be used to augment unlabeled RF data and even synthesize missing modalities for modality completion. RFusion, in turn, focuses on learning robust multi-modal representations. Thus, a promising future work is to combine the two: applying RFusion on a joint corpus of real and generated RF signals, and treating generated modalities as additional views to further enhance cross-modal alignment and robustness under missing-modality and few-shot conditions.

VIII. Conclusion

In this paper, we propose RFusion, a multimodal-assisted sensing paradigm that leverages unlabeled RF multimodal data in the pre-training stage to enhance the fine-tuning performance under modal scarcity scenarios. RFusion fully learns the mutual and unique information in multimodal data in the pre-training phase through the guider module and the arbiter module, and uses multi-head attention to improve the fine-tuning effect in missing modality scenarios with a small number of labeled samples. Our evaluation shows that RFusion has a significant improvement over the other latest baselines.

References

[1] A. Sanchez-Comas, K. Synnes, and J. Hallberg, "Hardware for recognition of human activities: A review of smart home and aal related technologies," *Sensors*, vol. 20, no. 15, p. 4227, 2020.

[2] D. Bouchabou, S. M. Nguyen, C. Lohr, B. LeDuc, and I. Kanellos, "A survey of human activity recognition in smart homes based on iot sensors algorithms: Taxonomies, challenges, and opportunities with deep learning," *Sensors*, vol. 21, no. 18, p. 6037, 2021.

[3] S. Mekruksavanich and A. Jitpattanakul, "Lstm networks using smartphone data for sensor-based human activity recognition in smart homes," *Sensors*, vol. 21, no. 5, p. 1636, 2021.

[4] A. Benmansour, A. Bouchachia, and M. Feham, "Human activity recognition in pervasive single resident smart homes: State of art," in *2015 12th International Symposium on Programming and Systems (ISPS)*. IEEE, 2015, pp. 1–9.

[5] M. Z. Uddin, D.-H. Kim, J. T. Kim, and T.-S. Kim, "An indoor human activity recognition system for smart home using local binary pattern features with hidden markov models," *Indoor and Built Environment*, vol. 22, no. 1, pp. 289–298, 2013.

[6] B. Yang, L. Chen, X. Peng, J. Chen, Y. Tang, W. Wang, D. Fang, and C. Feng, "Rf-sauron: Enabling contact-free interaction on eyeglass using conformal rfid tag," *IEEE Internet of Things Journal*, 2025.

[7] L. Bibbò, R. Carotenuto, and F. Della Corte, "An overview of indoor localization system for human activity recognition (har) in healthcare," *Sensors*, vol. 22, no. 21, p. 8119, 2022.

[8] D. Bhattacharya, D. Sharma, W. Kim, M. F. Ijaz, and P. K. Singh, "Ensem-har: An ensemble deep learning model for smartphone sensor-based human activity recognition for measurement of elderly health monitoring," *Biosensors*, vol. 12, no. 6, p. 393, 2022.

[9] K. Venkatachalam, Z. Yang, P. Trojovskỳ, N. Bacanin, M. Deveci, and W. Ding, "Bimodal har-an efficient approach to human activity analysis and recognition using bimodal hybrid classifiers," *Information Sciences*, vol. 628, pp. 542–557, 2023.

[10] M. Straczkiewicz, P. James, and J.-P. Onnela, "A systematic review of smartphone-based human activity recognition methods for health research," *NPJ Digital Medicine*, vol. 4, no. 1, p. 148, 2021.

[11] N. Kannen and A. Subasi, "Smart factories of industry 4.0: determination of the effective smartphone position for human activity recognition using deep learning," in *Advanced Signal Processing for Industry 4.0, Volume 2: Security issues, management and future opportunities*. IOP Publishing Bristol, UK, 2023, pp. 3–1.

[12] C.-L. Yang, S.-C. Hsu, Y.-W. Hsu, and Y.-C. Kang, "Har-time: Human action recognition with time factor analysis on worker operating time," *International Journal of Computer Integrated Manufacturing*, vol. 36, no. 8, pp. 1219–1237, 2023.

[13] P. Alli and J. Dinesh Peter, "Elevated cnn based secured sensor image data communication for har: Iiot," in *International Virtual Conference on Industry*. Springer, 2021, pp. 211–220.

[14] C. Hofmann, C. Patschkowski, B. Haefner, and G. Lanza, "Machine learning based activity recognition to identify wasteful activities in production," *Procedia Manufacturing*, vol. 45, pp. 171–176, 2020.

[15] S. Suh, V. F. Rey, S. Bian, Y.-C. Huang, J. M. Rožanec, H. T. Ghinani, B. Zhou, and P. Lukowicz, "Worker activity recognition in manufacturing line using near-body electric field," *IEEE Internet of Things Journal*, 2023.

[16] G. Bhola and D. K. Vishwakarma, "A review of vision-based indoor har: state-of-the-art, challenges, and future prospects," *Multimedia Tools and Applications*, vol. 83, no. 1, pp. 1965–2005, 2024.

[17] Y. Jang, I. Jeong, M. Younesi Heravi, S. Sarkar, H. Shin, and Y. Ahn, "Multi-camera-based human activity recognition for human–robot collaboration in construction," *Sensors*, vol. 23, no. 15, p. 6997, 2023.

[18] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366–1401, 2022.

[19] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[20] N. Wang, Y. Xiao, X. Peng, X. Chang, X. Wang, and D. Fang, "Contextdet: Temporal action detection with adaptive context aggregation," *arXiv preprint arXiv:2410.15279*, 2024.

[21] W. Gao, L. Zhang, Q. Teng, J. He, and H. Wu, "Danhar: Dual attention network for multimodal human activity recognition using wearable sensors," *Applied Soft Computing*, vol. 111, p. 107728, 2021.

[22] M. M. Islam, S. Nooruddin, F. Karray, and G. Muhammad, "Multi-level feature fusion for multimodal human activity recognition in internet of healthcare things," *Information Fusion*, vol. 94, pp. 17–31, 2023.

[23] A. Franco, A. Magnani, and D. Maio, "A multimodal approach for human activity recognition based on skeleton and rgb data," *Pattern Recognition Letters*, vol. 131, pp. 293–299, 2020.

[24] M. M. Islam and T. Iqbal, "Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1729–1736, 2021.

[25] F. Luo, S. Khan, Y. Huang, and K. Wu, "Activity-based person identification using multimodal wearable sensor data," *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 1711–1723, 2022.

[26] J. Yang, S. Tang, Y. Xu, Y. Zhou, and L. Xie, "Maskfi: Unsupervised learning of wifi and vision representations for multimodal human activity recognition," *arXiv preprint arXiv:2402.19258*, 2024.

[27] X. Ouyang, X. Shuai, J. Zhou, I. W. Shi, Z. Xie, G. Xing, and J. Huang, "Cosmo: contrastive fusion learning with small data for multimodal human activity recognition," in *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*, 2022, pp. 324–337.

[28] F. Wang, Y. Lv, M. Zhu, H. Ding, and J. Han, "Xrf55: A radio frequency dataset for human indoor action analysis," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–34, 2024.

[29] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[30] C. Wu, B. Wang, O. C. Au, and K. R. Liu, "Wi-fi can do more: toward ubiquitous wireless sensing," *IEEE Communications Standards Magazine*, vol. 6, no. 2, pp. 42–49, 2022.

[31] Y. Li, D. Wu, J. Zhang, X. Xu, Y. Xie, T. Gu, and D. Zhang, "Diversense: Maximizing wi-fi sensing range leveraging signal diversity," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–28, 2022.

[32] Z. Jiang, T. H. Luan, X. Ren, D. Lv, H. Hao, J. Wang, K. Zhao, W. Xi, Y. Xu, and R. Li, "Eliminating the barriers: Demystifying wi-fi baseband design and introducing the picoscenes wi-fi sensing platform," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4476–4496, 2021.

[33] W. Wang, J. Chen, and T. Hong, "Occupancy prediction through machine learning and data fusion of environmental sensing and wi-fi sensing in buildings," *Automation in Construction*, vol. 94, pp. 233–243, 2018.

[34] H. Yan, Y. Zhang, Y. Wang, and K. Xu, "Wiact: A passive wifi-based human activity recognition system," *IEEE Sensors Journal*, vol. 20, no. 1, pp. 296–305, 2019.

[35] R. Standard, "Low level reader protocol (llrp) 2 version 1.1 3," *nature*, vol. 18, p. 19, 2005.

[36] A. G. Stove, "Linear fmcw radar techniques," in *IEE Proceedings F (Radar and Signal Processing)*, vol. 139, no. 5. IET, 1992, pp. 343–350.

[37] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*. Springer, 2020, pp. 776–794.

[38] R. Brinzea, B. Khaertdinov, and S. Asteriadis, "Contrastive learning with cross-modal knowledge mining for multimodal human activity recognition," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 01–08.

[39] S. Dixon, L. Yao, and R. Davidson, "Modality aware contrastive learning for multimodal human activity recognition," *Concurrency and Computation: Practice and Experience*, p. e8020, 2024.

[40] C. Guo, Y. Zhang, Y. Chen, C. Xu, and Z. Wang, "Modality consistency-guided contrastive learning for wearable-based human activity recognition," *IEEE Internet of Things Journal*, 2024.

[41] B. Khaertdinov, E. Ghaleb, and S. Asteriadis, "Contrastive self-supervised learning for sensor-based human activity recognition," in *2021 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2021, pp. 1–8.

[42] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[44] Y. Li, P. Hu, Z. Liu, D. Peng, J. T. Zhou, and X. Peng, "Contrastive clustering," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 10, 2021, pp. 8547–8555.

[45] L. Lasantha, N. C. Karmakar, and B. Ray, "Chipless rfid sensors for iot sensing and potential applications in underground mining—a review," *IEEE Sensors journal*, vol. 23, no. 9, pp. 9033–9048, 2023.

[46] N. Khalid, R. Mirzavand, H. Saghlatoon, M. M. Honari, A. K. Iyer, and P. Mousavi, "A batteryless rfid sensor architecture with distance ambiguity resolution for smart home iot applications," *IEEE Internet of Things Journal*, vol. 9, no. 4, pp. 2960–2972, 2021.

[47] C. L. Baumbauer, M. G. Anderson, J. Ting, A. Sreekumar, J. M. Rabaey, A. C. Arias, and A. Thielens, "Printed, flexible, compact uhf-rfid sensor tags enabled by hybrid electronics," *Scientific reports*, vol. 10, no. 1, p. 16543, 2020.

[48] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[49] G. Li, Z. Zhang, H. Yang, J. Pan, D. Chen, and J. Zhang, "Capturing human pose using mmwave radar," in *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2020, pp. 1–6.

[50] C. Xu, X. Zheng, Z. Ren, L. Liu, and H. Ma, "Uhead: Driver attention monitoring system using uwb radar," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, pp. 1–28, 2024.

[51] H. Liu, K. Cui, K. Hu, Y. Wang, A. Zhou, L. Liu, and H. Ma, "mtranssee: Enabling environment-independent mmwave sensing based gesture recognition via transfer learning," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 1, pp. 1–28, 2022.

[52] F. Deng, E. Jovanov, H. Song, W. Shi, Y. Zhang, and W. Xu, "Wildar: Wifi signal-based lightweight deep learning model for human activity recognition," *IEEE Internet of Things Journal*, 2023.