RF-Net: A Unified Meta-Learning Framework for RF-enabled One-Shot Human Activity Recognition

Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo*

School of Computer Science and Engineering, Nanyang Technological University, Singapore

Email: {di0002ya, chen.zhe, tianyue002, junluo}@ntu.edu.sg

ABSTRACT

Radio-Frequency (RF) based device-free Human Activity Recognition (HAR) rises as a promising solution for many applications. However, device-free (or contactless) sensing is often more sensitive to environment changes than device-based (or wearable) sensing. Also, RF datasets strictly require on-line labeling during collection, starkly different from image and text data collections where human interpretations can be leveraged to perform off-line labeling. Therefore, existing solutions to RF-HAR entail a laborious data collection process for adapting to new environments. To this end, we propose RF-Net as a meta-learning based approach to one-shot RF-HAR; it reduces the labeling efforts for environment adaptation to the minimum level. In particular, we first examine three representative RF sensing techniques and two major meta-learning approaches. The results motivate us to innovate in two designs: i) a dual-path base HAR network, where both time and frequency domains are dedicated to learning powerful RF features including spatial and attention-based temporal ones, and ii) a metric-based meta-learning framework to enhance the fast adaption capability of the base network, including an RF-specific metric module along with a residual classification module. We conduct extensive experiments based on all three RF sensing techniques in multiple real-world indoor environments; all results strongly demonstrate the efficacy of RF-Net compared with state-of-the-art baselines.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**.

KEYWORDS

Human activity recognition, RF sensing, meta-learning.

*The first two authors contribute equally to this research. Our open-source codes and datasets can be found at https://github.com/di0002ya/RFNet.git.
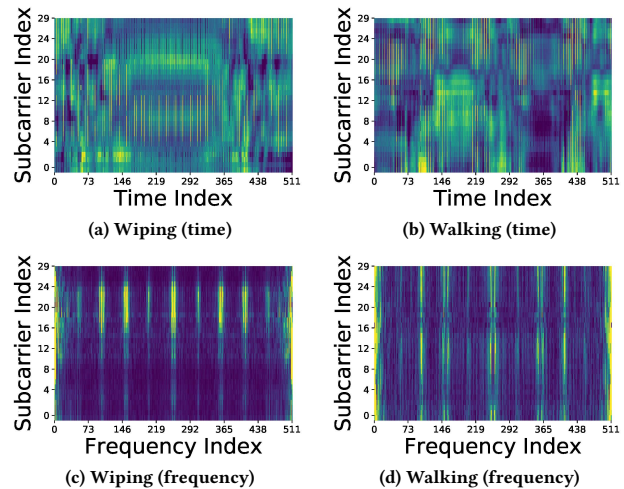
Figure 1: Wi-Fi CSI heatmaps of two different activities: wiping the board and walking around. It is virtually impossible to intuitively recognize them by human eyes.

1 INTRODUCTION

Recently, Human Activity Recognition (HAR) has been attracting attention increasingly in both academia and industry, giving its promising potential in real-world applications such as smart home [11, 20, 24], health care [13, 16, 27], and fall detection [22, 52, 60]. Generally, two categories for HAR have been explored: *device-based* (or wearable) and *device-free* (or contactless). Device-based HAR leverages the wearable devices such as smart phones or watches to recognize human activities [4, 19, 27, 53]. However, it may cause discomfort and extra burden, which leads to the alternative method of device-free HAR. This later method exploits camera-based image data [21, 36, 37], acoustic/ultrasonic signals [17, 58, 63], and Radio-Frequency (RF) signals [11, 26, 75] to achieve HAR.

Research efforts leveraging the efficacy of deep learning have been intensively carried out on camera-based approaches [35–37], but exploiting image data bears a risk of privacy infringement especially in private places [20]. Meanwhile, acoustic/ultrasonic sensing often requires bulky devices and incurs a high energy consumption, hence rarely adopted in practice even though the performance can be excellent in noise-free environments [17, 63]. Fortunately, RF sensing offers just enough resolution to perform HAR without infringing personal privacy, and it is far more energy-efficient than sound, as RF signals are not generated by mechanical vibration. Moreover, its performance is robust even under various temperature or lighting conditions. Therefore, RF-HAR is deemed as the most promising solution, where Wi-Fi is often adopted.

Shuya Ding, Zhe Chen, Tianyue Zheng, and Jun Luo

Early RF-HAR exploited RSSI (Received Signal Strength Indicator) to analyze human activities [48, 70], collected by COTS (Commercial-Off-The-Shelf) Wi-Fi cards. However, RSSI can only represent coarse information of Wi-Fi signals, rather than fine-grained multipath effects generated by human activities. Therefore, many efforts have been recently devoted to extracting comprehensive information in Wi-Fi CSI (Channel State Information) from Intel 5300 [15] Wi-Fi cards [10, 26, 59, 62]. As illustrated in Figure 1, body movements of different activities bring variations to CSI, so features generated from CSI could be leveraged for HAR. Whereas Wi-Fi sensing exploits devices originally designed for wireless communication purposes (hence inherently limited in sensing performance), dedicated RF sensing techniques are gaining momentum very recently. These techniques are typically supported by two types of radios, namely FMCW (Frequency-Modulated Continuous Wave) radio [2, 73] and impulse radio [30, 44, 74].

Although RF sensing, in general, has achieved great performance in HAR, it does have two major weaknesses. On one hand, similar to other device-free sensing techniques [36, 41, 63], RF sensing requires high volume of training data to re-train its model if the sensing environment is altered (e.g., changing the radio locations and/or the room furniture layout). This weakness is inherent to device-free sensing as, unlike device-based sensing, it senses both the subject and the background. On the other hand, as illustrated in Figure 1, human cannot intuitively recognize different activities from RF sensing data (as opposed to images). Consequently, human interpretation cannot be exploited for conducting off-line labeling, which is essential for image and text data collections. As a result, RF sensing has to endure a much more laborious process (compared with computer vision and natural language processing) to gather labeled data sufficiently, in order for RF-HAR approaches to be able to adapt to new environments with satisfactory performance.

Fortunately, recent developments on meta-learning [12, 47, 56] have offered us a chance to enhance the environment adaptivity of RF sensing. Roughly categorized into three types: model-based, optimization-based, and metric-based, meta-learning aims to *adapt to new tasks rapidly with few labeled observations*. Whereas model-based approaches [42, 47] incur a high computational complexity, the other two may potentially help RF-HAR. Optimization-based approaches [12, 25, 46] adjust the training algorithm to find good initialization weights or learning rate, but their generalization ability is questionable and the incurred overhead can still be high [46, 51]. Metric-based approaches [31, 50, 56] classify an unlabeled observation by its similarity to the labeled data. As they focus on learning an optimal similarity metric (instead of directly tuning the learning architecture), they likely incur the lowest complexity.

Leveraging the power of meta learning, we propose RF-Net for one-shot RF-HAR. In other words, RF-Net performs HAR accurately in a new environment with only one observation for each label. Specifically, we first carefully examine representative RF sensing techniques along with major meta-learning approaches. The results motivate us to innovate in two aspects: i) a dual-path base network for classifying activities, and ii) a metric-based meta-learning framework to improve the fast adaption capability of the base network, as illustrated in Figure 2. For the base HAR network, we combine spatial module along with attention-based temporal module, aiming to learn influential spatial-temporal features from both time and
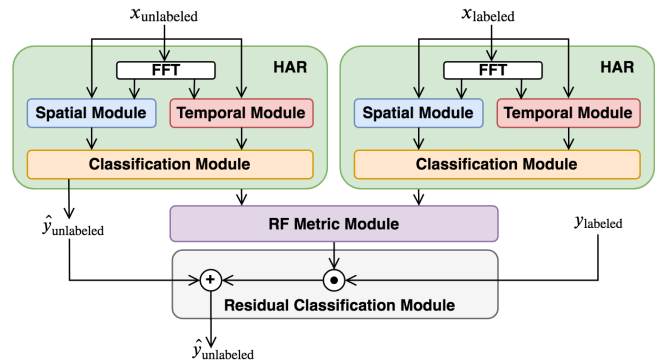

Figure 2: RF-Net overview.

frequency domains. Our meta-learning framework contains a parametric RF-specific module designed to train a powerful distance metric, instead of simply relying on traditional non-parametric metrics (e.g., Euclidean or Cosine distances). This helps to achieve a better generalization when applying this metric to conduct classifications in new environments. Essentially, we employ the base network to perform both activity recognition and feature extraction, and exploit them to (meta)-train the distance metric via a residual classification module. In summary, our major contributions are:

- We propose RF-Net, a unified meta-learning framework for RF-enabled one-shot HAR, delivering the capability of being adaptive to new environments with very few labeled data.
- We innovatively design a dual-path activity recognition (base) network, aiming to learn influential features from general RF signals for enhancing HAR accuracy.
- We equip our meta-learning framework with a novel RF-specific module designed to train a powerful distance metric, so as to achieve a better generalization.
- We conduct extensive experiments on multiple RF sensing techniques and in many indoor environments. We demonstrate the superior performance of our proposed RF-Net compared against multiple baselines.

This paper is organized as follows. Background and related works are first examined in Section 2. Then our RF-Net is presented in Section 3. Extensive experiment results are reported in Section 4, and our paper is concluded in Section 5.

2 BACKGROUND AND LITERATURE

We carefully study representative RF sensing techniques along with major meta-learning approaches in this section, aiming to better motivate our design. The rationale behind this study is that, as RF sensing induces far more complicated input data than those under wearable sensing (where only up to three time series from the three axes of an IMU sensors present [28]), deep understanding and innovative treatments are hence needed.

2.1 Modeling RF Sensing

Generally, all RF sensing approaches explore RF *channel state information* (CSI) to differentiate human activities. Therefore, we model the RF channel first, then examine three typical RF signals. According to [54], considering a pair of transmitter and receiver
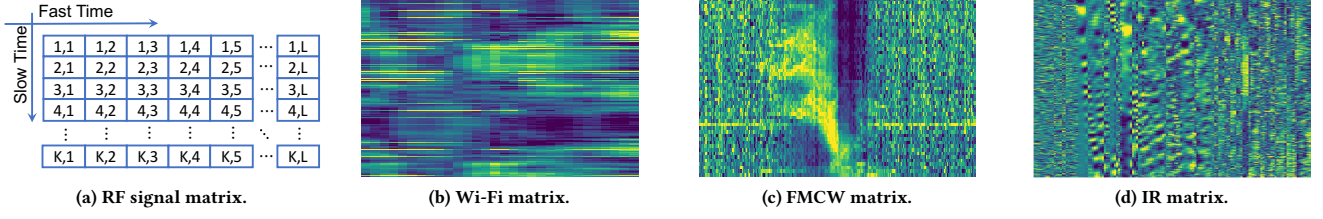
(a) RF signal matrix.

(b) Wi-Fi matrix.

(c) FMCW matrix.

(d) IR matrix.

Figure 3: The RF signal matrix and corresponding heatmaps for three RF techniques in terms of the "walking" activity.

in an indoor environment with P propagation paths, we have the following baseband RF channel model given a carrier frequency f_c:

$$h(t) = \sum_{p=1}^{P} \alpha_p e^{j2\pi f_c \tau_p} + n(t), \tag{1}$$

where α_p is the amplitude of p-th path, and $n(t)$ is Gaussian noise. Moreover, $\tau_p = \tau_p^S + \tau_p^D$ where τ_p^S and τ_p^D are the p-th time delays caused by static reflections and motion reflections, respectively. For a transmitted signal $s(t)$, the received signal becomes $y(t) = (h * s)(t)$, where $*$ denotes convolution. In the following, we stick to only one path (thus removing the subscript p), and we omit noise term for brevity.

Wi-Fi Radio. Wi-Fi communications utilize Orthogonal Frequency Division Multiplexing (OFDM) to encoding digital data on multiple subcarriers. Let L denote the number of OFDM subcarriers, s_ℓ denotes the ℓ-th subcarrier, then the received signal y_ℓ^W can be represented as follows:

$$y_\ell^{\mathrm{W}} = \alpha_\ell e^{j2\pi f_\ell \tau} s_\ell, \quad \ell \in \{1, ..., L\}, \tag{2}$$

where α_ℓ and f_ℓ are the amplitude and frequency of the ℓ-th subcarrier, respectively. Here the time variable t disappears because the bandwidth of Wi-Fi is so narrow that it is approximately deemed as time invariant. The channel state can then be estimated as $\hat{h}_\ell = y_\ell^W / s_\ell$, whose phase $\angle \hat{h}_\ell$ contains temporal feature τ. Given a total K received packets, each of them offers a CSI vector $[\hat{h}_\ell]_{\ell \in \{1, \cdots, L\}}$. Combining all these data, we obtain the ensemble input sample as a $K \times L$ signal matrix. Here we term the row index as *slow-time* and the column index as *fast-time*, as they represent sampling at different temporal scales. Figure 3a illustrates this signal model, while Figure 3b provides an example for a Wi-Fi matrix.

FMCW Radio. FMCW is a special type of radar implementation. Different from normal radios, FMCW radio transmits analog signals modulated in a continuously increasing frequency across a wide bandwidth. Consequently, it sweeps across time with a fine-grained (frequency) resolution, equivalently convertible to a fine-grained range resolution. For FMCW radio, we denote the bandwidth by B and sweeping time span by T^S. According to [39], the received FMCW signal $y^F(t)$ is:

$$y^{\mathrm{F}}(t) = \alpha \Pi(t - \tau) e^{-j2\pi(\beta \tau t + f_c \tau - 0.5\beta \tau^2)}, \tag{3}$$

where $\beta = B/T^S$ and $\Pi(t)$ is a rectangle function ranged from $-T^S/2$ to $T^S/2$. We term a signal represented by Eq (3) a *frame*; it uses a frequency $\beta \tau$ and a phase $f_c \tau - \beta \tau^2 / 2$ to represent how the signal changes over time. Therefore, we perform an L-point FFT

on each frame and get a set of L frequency components, which correspond to fast-time samples. Then we combine all K frames together to construct a $K \times L$ signal matrix (similar to that of Wi-Fi), as shown in Figure 3c.

Impulse Radio. Whereas FMCW radio uses a varying frequency to sweep time, impulse radio (IR) transmits a pulse signal with an extremely short time duration. This enables IR signals to also occupy a wide bandwidth B, again leading to a fine-grained time resolution. The received pulse signal $y^I(t)$ is:

$$y^{\mathrm{I}}(t) = \alpha e^{j2\pi f_c \tau} e^{-0.5(t-0.5T_{\mathrm{tx}}-\tau)^2 \epsilon_{\mathrm{tx}}^{-2}}, \tag{4}$$

where $T_{\mathrm{tx}} = \frac{1}{B}$ is the signal duration, and $\epsilon_{\mathrm{tx}} = \frac{1}{2\pi B_{-10\,\mathrm{dB}}(\log_{10}(e))^{1/2}}$ is the standard deviation determining the -10 dB bandwidth. Similar to FMCW, if we consider L samples of a pulse as a *frame*, combining K such frames again gives us a $K \times L$ signal matrix, where the fast-time directly corresponds to the temporal sample indices of a frame, as shown in Figure 3d.

Remark: If we deem this matrix as an image (common input to deep learning models such as CNN [32, 33]), we may stack multiple such matrices derived from different tx-rx pairs together, so that each tx-rx pair corresponds to an input channel. Although these three RF sensing techniques share a similar model, they offer very different time (thus range) resolutions: the 20 MHz bandwidth of Wi-Fi can only translate to a range resolution about 15 m, but both FMCW and impulse radios, with more than 1 GHz bandwidth, can achieve a centimeter-level resolution instead.

2.2 RF Meets Learning: Status and Challenges

Existing RF-sensing solutions mostly leverage machine learning techniques to extract features and classify activities [3, 10, 38, 55, 57, 59, 72, 73]. Although their adopted RF signals and designed algorithms vary, their input data can all be unified under the RF signal matrix introduced in Section 2.1 as subcategories. For example, the proposals in [1, 3, 10, 55, 57, 59] extract Doppler shift only along the slow time axis in Figure 3b, in order to capture the velocities of moving parts of a target and in turn to perform classifications. Other proposals [72, 73] utilize multiple tx-rx pairs to extract range information embedded along the fast time dimension in Figure 3c, so as to predict human poses. Essentially, existing solutions only retrieve partial information provided by RF signals; this motivates us to consider a full exploitation of spatial, temporal, frequency features offered by the signal matrix, so as to handle all three RF sensing techniques using a unified learning framework.
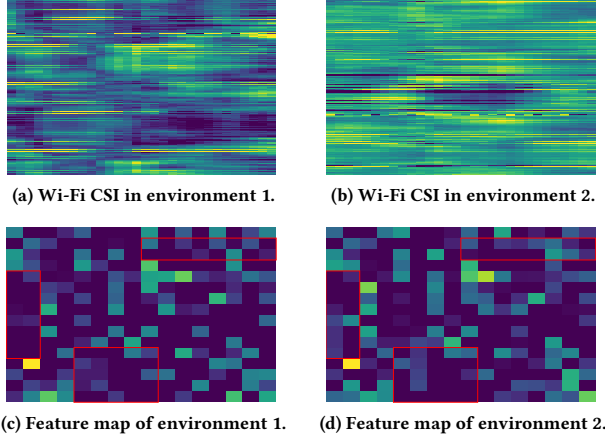
(a) Wi-Fi CSI in environment 1.

(b) Wi-Fi CSI in environment 2.

(c) Feature map of environment 1.

(d) Feature map of environment 2.

Figure 4: Environment influence: the same activity "walking" leads to distinct signal matrices and extracted features.

However, exploiting more features is a double-edged sword: it improves activity classification under noise-free circumstances, but it also makes RF sensing more sensitive to environment changes in practice. Figure 4 demonstrates the influence of this sensitivity. As explained in Section 1, re-training a learning model to suit a new environment can be extremely expensive, as RF signal matrices, totally different from images and texts, are not human understandable, enforcing a laborious on-line labelling. One solution is to employ environment-invariant features/models, hoping to retain classification accuracy without re-training in a new environment. Previous works [57, 75, 76] utilize handcrafted features for this purpose. However, these features depend on prior information such as the position and orientation of a target, likely incurring another laborious process in obtaining these information. Another solution is transfer learning [26, 69]. The basic idea is to learn "transferable" knowledge applicable across a pair of source and target environments. However, when applied to a new target environment, the whole fine-tuning process has to be performed again with a substantial amount of labeled data. In order to better tackle this challenge, we need to consider recently emerged alternatives.

2.3 Meta-Learning Basics

Human level intelligence requires that learning models can mimic human behavior to learn from known tasks (environments in our context) and adapt to new tasks quickly with only a few labeled observations. Recently, meta-learning [12, 47, 56] has emerged to achieve the aforementioned intelligence. The key idea of meta-learning is twofold: i) learning knowledge from source environments with rather abundant data, and ii) exploiting accumulated knowledge to learn similarities and differences in all target environments, requiring only a minimum level of labeled data. In meta-learning, we denote the dataset for the ℓ-th environment by $\mathcal{D}_\ell$. Each $\mathcal{D}_\ell$ is split into a *support set* $\mathcal{D}_\ell^S$ for learning and a *query set* $\mathcal{D}_\ell^Q$ for training and testing. We drop the subscript ℓ in the following discussions as the learning procedure is identical in all source datasets. Generally, a *base network* f_Φ predicts the probability $P_\Theta(y|\mathbf{x}, \mathcal{D}^S)$ of class y in source environments, given a support

set $\mathcal{D}^S$ and an input $\mathbf{x}$ in $\mathcal{D}^S$, where Θ is the *meta-parameter* that parameterizes the conditional probability.

$$\Phi^* = \arg\max_\Phi \mathbb{E}_{\mathcal{D}^S} \left[\sum_{(\mathbf{x},y) \in \mathcal{D}^S} P_\Theta(y|\mathbf{x}, \mathcal{D}^S) \right]. \tag{5}$$

Then the optimal meta-prameter Θ^* is obtained by maximizing the expectation over all query sets in the source environments as:

$$\Theta^* = \arg\max_\Theta \mathbb{E}_{\mathcal{D}^Q} \left[P(\Phi|\mathcal{D}^Q) \right]. \tag{6}$$

Upon a new target environment with dataset $\hat{\mathcal{D}}$ (containing $\hat{\mathcal{D}}^S$ as the support set with a minimum level of labeled data, and the rest unlabeled for final testing), Θ^* is transferred to this environment and the base network f_Φ is refined according to $\hat{\mathcal{D}}^S$. The two major meta-learning approaches applicable in sensing context differ in how $P_\Theta(y|\mathbf{x}, \mathcal{D}^S)$ is modelled.

2.3.1 Optimization-Based. This line of research focuses on the optimization-based training algorithms, aiming to cope with few observations or to converge within few optimization steps. Essentially, training algorithms are adjusted to find good initialization weights as Θ, so that f_Φ could be generalized to new tasks. These approaches model $P_\Theta(y|\mathbf{x}, \mathcal{D}^S)$ as $P_{g_{\Theta_g(\mathcal{D}^S)}}(y|x)$, where

$$g_{\Theta_g(\mathcal{D}^S)} = g_{\Theta_g} \left(\Theta_0, \{\nabla_{\Theta_0}\mathcal{L}(\mathbf{x}_i, y_i)\}_{(\mathbf{x}_i, y_i) \in \mathcal{D}^S} \right), \tag{7}$$

is the *meta learner* that generates initialization weights with the gradient of loss $\{\nabla_{\Theta_0}\mathcal{L}(\mathbf{x}_i, y_i)\}_{(\mathbf{x}_i, y_i) \in \mathcal{D}^S}$ and starting weights Θ_0 as input. Model-Agnostic Meta-Learning (MAML) [12] claims to be applicable to any network learnt via gradient descent, whereas later proposals (e.g., Reptile [43], Meta-SGD [34], TAML [25]) all intend to improve learning efficiency along various directions. However, as stated in [46, 51], the generalization ability of these approaches is questionable. To be specific, when they are applied to a base network f_Φ with a high-dimensional parameter space, tuning initialization weights Θ directly via $g_{\Theta_g}(D^S)$ could result in generalization difficulty given very few observations. Recently, MetaSense [14] adopts MAML to construct an adaptive wearable sensing system. Compared with RF sensing data, the dimension of wearable sensing data obtained by IMUs is much lower. Therefore, they adopt shallow neural network as f_Φ to avoid the generalization challenge described above. RF sensing (given the data model described in Section 2.1) certainly demands a powerful base network f_Φ (with a high-dimensional parameter space for Θ), rendering optimization-based meta-learning approaches largely inapplicable.

2.3.2 Metric-Based. This category of non-parametric approaches intend to classify a new observation $\mathbf{x}$ by a weighted sum of the labels in $\mathcal{D}^S$ as follows:

$$P_\Theta(y|\mathbf{x}, \mathcal{D}^S) = \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}^S} k_\Theta(f_\Phi(\mathbf{x}), f_\Phi(\mathbf{x}_i))y_i, \tag{8}$$

where k_Θ is an optimal metric function for measuring the similarity between $\mathbf{x}$ and $\mathbf{x}_i$. Essentially, they aim to learn an embedding function f_Φ that transforms inputs into a representation suitable for classification via the similarity comparison. Although several proposals have been made under this category (e.g., Siamese networks [31], Matching networks [56], Prototypical networks [50]),

they mostly differ in the choice of the (input) embedding vectors and non-parametric distance metric (e.g., Cosine similarity). Compared with optimization-based approaches, the meta-training phase of metric-based approaches is rather straightforward: it mainly focuses on learning a powerful distance metric to achieve generalization rather than directly tuning f_Φ. This property has made metric-based meta-learning approaches less constrained by the complexity of f_Φ, and it has also motivated us to adopt metric-based approaches for achieving one-shot RF-HAR in this paper.

3 RF-NET: ONE-SHOT HAR

Based on our discussions in Section 2, we hereby present RF-Net comprising two novel designs: i) a meta-learning framework that involves a parametric RF-specific module for training a powerful distance metric, and ii) a dual-path base network that fully exploits the high-dimensional features contained in the signal matrix (thus applicable to all three RF sensing datasets). We first describe the problem formulation. Then we elaborate our meta-learning framework and dual-path base HAR network.

3.1 Problem Formulation

In this paper, the ultimate goal of RF-Net q_Ω, parameterized by Ω, is to perform one-shot RF-HAR, i.e., adapting to every new environment rapidly with a single labeled observation per class. To achieve it, we need a base HAR network f_Φ, parameterized by Φ, to extract features from input observations (i.e., RF signal matrix). As introduced in Section 2.1, we deem an RF signal matrix $\mathbf{x} \in \mathbb{R}^{K \times L \times N_r}$ as an image, where K is slow time dimension, L is fast time dimension, and N_r is the number of tx-rx pairs. This base network f_Φ is then wrapped into a meta-learning framework to be generalized to new environments. In order to maintain the generalization capability of f_Φ given its high-dimensional parameter space, we adopt a metric-based meta-learning framework g_Θ parameterized by Θ. Therefore, RF-Net q_Ω includes the base network f_Φ and meta-learning framework g_Θ, hence $\Omega = \Phi \cup \Theta$.

The procedure for learning q_Ω is planned as follows. We first train RF-Net with N_e source environment training datasets $\mathcal{D} = \{\mathcal{D}_\ell\}_{\ell=1}^{N_e}$. For each epoch, we train q_Ω on environment datasets, learning from environment to environment to mimic how RF-Net would be tested when presenting in a new environment. More specifically, for each environment dataset $\mathcal{D}_\ell$, we sample support observations as support set $\mathcal{D}_\ell^S = \{\mathbf{x}_{\ell,j}^S, y_{\ell,j}^S\}_{j=1}^{N_c}$, $\mathbf{x}_{\ell,j}^S \in \mathbb{R}^{N_s \times K \times L \times N_r}$ and query observations as query set $\mathcal{D}_\ell^Q = \{\mathbf{x}_\ell^Q, y_\ell^Q\}$, $\mathbf{x}_\ell^Q \in \mathbb{R}^{1 \times K \times L \times N_r}$, where N_s denotes the number of observations (i.e., $N_s = 1$ for *one-shot* learning), N_c is the number of activity categories, and y_ℓ denotes activity label. Note that the support set $\mathcal{D}_\ell^S$ and query set $\mathcal{D}_\ell^Q$ belong to the same environment space but observations are disjoint, i.e., $\mathcal{D}_\ell^S \cap \mathcal{D}_\ell^Q = \varnothing$. We drop the subscript ℓ as the learning procedure is identical in all environments. Essentially, the objective function of RF-Net q_Ω can be formulated as follows:

$$\Omega^* = \arg\max_\Omega \mathbb{E}_\mathcal{D}\left[\sum_{\mathcal{D}^Q}\sum_{\mathcal{D}^S} g_\Theta\left(f_\Phi\left(\mathbf{x}^Q\right), f_\Phi\left(\mathbf{x}^S\right), y^S\right)\right]. \quad (9)$$

3.2 Meta Framework

We intend to design a trainable metric-based meta-learning framework specifically for RF signal matrices. As illustrated in Figure 5a, our proposed meta framework g_Θ consists of two modules:

- RF Metric Module g_{metric}: it aims to train a powerful distance metric for measuring observations similarities via a linear mapping layer. Meanwhile, it takes into consideration the multiple features generated by the dual-path base network f_Φ. Consequently, this module could provide a holistic interpretation of $\mathbf{x}$ toward better generalization.
- Residual Classification Module g_c: it intends to incorporate the capability of f_Φ further as a recognition network, in addition to its feature extractor role in g_{metric}. This module allows the base network to assist in (meta-)training the distance metric g_{metric}.

3.2.1 RF Metric Module. RF metric module first employs our dual-path base HAR network (see Section 3.3) as a (deep) feature extractor $f_\Phi : \mathbb{R}^{D_\mathbf{x}} \to \mathbb{R}^{D_\mathbf{z}}$, mapping an RF signal matrix $\mathbf{x}$ to a feature representation $\mathbf{z}$, where $D_\mathbf{x}$ and $D_\mathbf{z}$ are the dimensionalities of RF signal matrix and the corresponding embedding, respectively. Different from previous works such as [10, 26, 56], we expect that the metric space of multiple features is able to capture a more complex representation than simply relying on single feature representation. Therefore, considering the RF signal matrix described in both time and frequency domains, we leverage M features to represent the RF signal matrix $\mathbf{x}$ as follows:

$$\{H_m^{\text{feat}} = f_\Phi(\mathbf{x})\}_{m=1}^M.$$

In our problem setting, H_m^{feat} includes H^{time} and H^{freq} respectively extracted from time and frequency domains, as well as H^{fuse} combining features from both domains to derive spatial and temporal information. We refer to Section 3.3 for further elaborations on these features. Given these features, we employ cosine distance $d(a, b) = -\dfrac{a \cdot b}{\|a\|\|b\|}$ to obtain a distance set $\{\lambda_m\}_{m=1}^M$ between features generated by support observations and query observations as follows:

$$\{\lambda_m\}_{m=1}^M = \left\{d\left(H_m^{\text{feat,S}}, H_m^{\text{feat,Q}}\right)\right\}_{m=1}^M.$$

In order to further increase the RF metric space, we design RF metric module to be trainable instead of solely relying on cosine distance. We stack these M distances $[\lambda_1, \lambda_2, ..., \lambda_M]$ into a vector form $\mathbf{\Lambda}$. Given a query observation, we propose to combine each distance measure λ_m in $\mathbf{\Lambda}$ via learnable linear mapping weights $\boldsymbol{\eta}$. Therefore, we could compute the activity probabilities of query observations by weighting the labels of support observations y^S as follows:

$$p_{\boldsymbol{\beta}}(y = j|\mathbf{x}) = \text{softmax}(-\mathbf{\Lambda}_j\boldsymbol{\beta}), \boldsymbol{\beta} = \boldsymbol{\eta}y^S, \quad (10)$$

where $\boldsymbol{\eta} \in \mathbb{R}^{M \times N_c}$ and $\Lambda_j \in \mathbb{R}^{1 \times M}$. To further realize the role of $\boldsymbol{\beta}$, we analyze that for j-th class, the class-wise cross-entropy loss function is given by

$$L_j(\boldsymbol{\beta}) = \sum_{\mathbf{x} \in \mathcal{D}^Q}\left[\Lambda_j\boldsymbol{\beta} + \log\left(\sum_{n=1}^{N_c} e^{-\Lambda_n\boldsymbol{\beta}}\right)\right]. \quad (11)$$
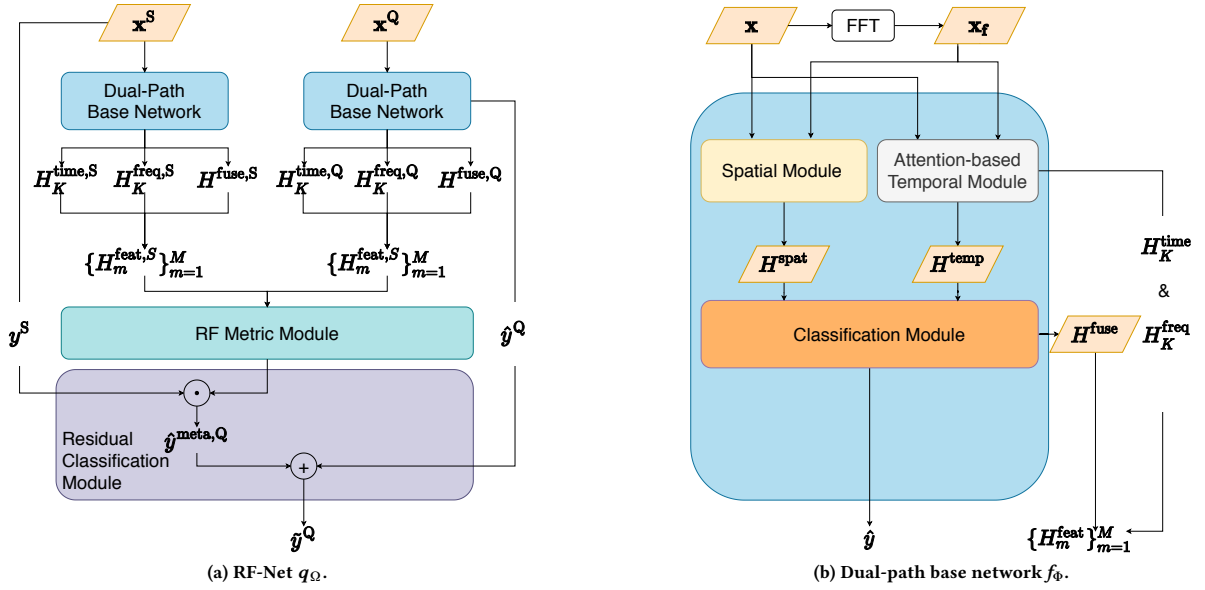
(a) RF-Net q_Ω.

(b) Dual-path base network f_Φ.

Figure 5: Overall design of RF-Net q_Ω and its dual-path base network f_Φ.

For each sample of query set $x \in \mathcal{D}^Q$, the second-order partial derivative of Eq (11) with respect to $\boldsymbol{\beta}$ is

$$\nabla^2 L_j(\boldsymbol{\beta}) = \frac{1}{\mathbf{1}^T \gamma} \text{diag}(\gamma) - \frac{1}{(\mathbf{1}^T \gamma)^2} \gamma \gamma^T, \quad (12)$$

where $\gamma = [\gamma_1, \cdots, \gamma_{N_c}]$ and $\gamma_k = e^{-\Lambda_k \boldsymbol{\beta}}$. If $\nabla^2 L_j(\boldsymbol{\beta}) \geq 0$, $L_j(\boldsymbol{\beta})$ in Eq (11) is convex. Therefore, we need to verify that $v^T \nabla^2 L_j(\boldsymbol{\beta}) v \geq 0$ for all v, but we have:

$$v^T \nabla^2 L_j(\boldsymbol{\beta}) v = \frac{(\sum_{n=1}^{N_c} \gamma_n v_n^2)(\sum_{n=1}^{N_c} \gamma_n) - (\sum_{n=1}^{N_c} v_n \gamma_n)^2}{(\sum_{n=1}^{N_c} \gamma_n)^2},$$

which is indeed non-negative due to Cauchy-Schwarz inequality, i.e., $(\sum_{n=1}^{N_c} v_n \gamma_n)^2 \leq (\sum_{n=1}^{N_c} \gamma_n v_n^2)(\sum_{n=1}^{N_c} \gamma_n)$. Now we have proven the convexity of $L_j(\boldsymbol{\beta})$, which in turn indicates that our RF metric module can quickly learn the optimal parameters $\boldsymbol{\beta}$ for combining multiple representations together.

3.2.2 Residual Classification Module. In this module, different from traditional metric-based meta-learning framework, we aim to employ the base network f_Φ for recognition too, rather than solely as a feature extractor explained in Section 3.2.1. To be specific, we first compute $\hat{y}^{\text{meta,Q}}$, i.e., classify query observations by weighting the labels of support observation y^S, as described in Eq (10). Meanwhile, we exploit f_Φ as recognition network to compute logits directly $\hat{y}^Q = f_\Phi(\mathbf{x}^Q)$. Afterwards, we incorporate $\hat{y}^Q$ into $\hat{y}^{\text{meta,Q}}$ via a residual connection, so that the final predicted logits of RF-Net $\tilde{y}^Q$ is computed as $g_c(g_{\text{metric}}(f_\Phi(\mathbf{x}^S), f_\Phi(\mathbf{x}^Q)), y^S, \hat{y}^Q) = g_{\text{metric}}(f_\Phi(\mathbf{x}^S), f_\Phi(\mathbf{x}^Q))y^S + \hat{y}^Q$. It enables the base network to reinforce the meta-training of the distance metric g_{metric}.

3.2.3 Training Strategy. We carefully devise our training strategy for learning RF-Net. As described in **Algorithm 1**, given RF-Net q_Ω including the base network f_Φ and meta-learning framework g_Θ,

we exploit training dataset $\mathcal{D}$ and testing dataset $\hat{\mathcal{D}}$ in training and testing stages, respectively. In training stage (line 2-11), we train f_Φ and g_Θ in the leader-follower asymmetric manner between inner- and meta-training to enhance the overall capability on performing one-shot RF-HAR with q_Ω. More specifically, we first inner train f_Φ with $\mathcal{D}^S$ (line 7). We minimize the inner training loss $\mathcal{L}(f_\Phi(\mathbf{x}^S), y^S)$

Algorithm 1 RF-Net training.

Require: Training dataset $\mathcal{D}$, Testing dataset $\hat{\mathcal{D}}$, Base network f_Φ, RF-Net q_Ω, Meta-learning framework g_Θ, hyperparameters α and β, as well as the maximum number of iterations iter

1: % Training
2: **while** iter $! = 0$ **do**
3:     Sample environment minibatch $\{\mathcal{D}_\ell\} \sim \mathcal{D}$
4:     **for** $\mathcal{D}_\ell \in \{\mathcal{D}_\ell\}$ **do**
5:         Sample support and query observations $\mathcal{D}_\ell^S, \mathcal{D}_\ell^Q \sim \mathcal{D}_\ell$
6:         $\mathcal{D}_\ell^S = \{\mathbf{x}_{\ell,j}^S, y_{\ell,j}^S\}_{j=1}^{N_c}, \mathcal{D}_\ell^Q = \{\mathbf{x}_\ell^Q, y_\ell^Q\}, \mathcal{D}_\ell^Q \cap \mathcal{D}_\ell^S = \varnothing$
7:         Evaluate $\nabla_\Omega \mathcal{L}_\ell(f_\Phi(\mathbf{x}_\ell^S), y_\ell^S), \; \Omega \leftarrow \Omega - \alpha \nabla_\Omega \mathcal{L}_\ell$
8:         Evaluate $\nabla_\Theta \mathcal{L}_\ell(q_\Omega(\mathbf{x}_\ell^S, \mathbf{x}_\ell^Q, y_\ell^S), y_\ell^Q), \; \Theta \leftarrow \Theta - \beta \nabla_\Theta \mathcal{L}_\ell$
9:     **end for**
10:     iter $\leftarrow$ iter $- 1$
11: **end while**
12: % Testing
13: **for** $\hat{\mathcal{D}}_\ell \in \hat{\mathcal{D}}$ **do**
14:     Sample support and query observations $\hat{\mathcal{D}}_\ell^S, \hat{\mathcal{D}}_\ell^Q \sim \hat{\mathcal{D}}_\ell$
15:     $\hat{\mathcal{D}}_\ell^S = \{\mathbf{x}_{\ell,j}^S, y_{\ell,j}^S\}_{j=1}^{N_c}, \hat{\mathcal{D}}_\ell^Q = \{\mathbf{x}_\ell^Q, y_\ell^Q\}, \hat{\mathcal{D}}_\ell^Q \cap \hat{\mathcal{D}}_\ell^S = \varnothing$
16:     Evaluate $\nabla_\Theta \mathcal{L}_\ell(f_\Phi(\mathbf{x}_\ell^S), y_\ell^S), \Theta \leftarrow \Theta - \alpha \nabla_\Theta \mathcal{L}_\ell$
17:     Predict $\tilde{y}^Q = q_\Omega(\mathbf{x}_\ell^S, \mathbf{x}_\ell^Q, y_\ell^S)$
18: **end for**

---

by conducting gradient descent with respect to both base network parameters Φ and meta parameters Θ. This helps finding a coarse initialization points for learning meta parameters Θ, boosting the convergence speed of RF-Net q_Ω. Afterwards (line 8), we meta-train g_Θ conditioned on learned f_Φ using both support and query observations, so that it produces q_Ω that performs well on recognizing query observations.

Now we have achieved a well-trained RF-HAR q_Ω by sequentially training f_Φ and g_Θ. Given a testing dataset $\hat{\mathcal{D}}$, we could evaluate the performance of RF-Net on performing one-shot RF-HAR. We first intend to refine q_Ω with labeled support observations $\hat{\mathcal{D}}^S = \{\mathbf{x}_j^S, y_j^S\}_{j=1}^{N_c}$. To be specific, we adapt the well-trained q_Ω using the same inner training procedure, except that we only fine-tune meta-learning framework parameter Θ (lines 16). Finally, we classify query observations with refined q_Ω and labeled support observations as $\tilde{y}^Q = q_\Omega(\mathbf{x}^S, \mathbf{x}^Q, \mathbf{y}^S)$.

3.3 Dual-Path Base Network

As illustrated in Figure 5b, our base HAR network f_Φ is composed of three main modules to learn RF feature representations:

- Spatial Module f_s: it extracts sensitive spatial features from RF matrices $\mathbf{x}$ in both time and frequency domains.
- Attention-based Temporal Module f_t: it aims to capture long-term temporal features from RF matrices in both time and frequency domains. And it also intends to generate joint temporal representations across two domains.
- Classification Module f_c: this final module predicts activity label, given trained features from f_s and f_t.

For both f_s and f_t, we first compute $\mathbf{x_f} \in \mathbb{R}^{K \times L \times N_r}$ via FFT along the slow time of an RF matrix $\mathbf{x} \in \mathbb{R}^{K \times L \times N_r}$. Then both $\mathbf{x}$ (time domain) and $\mathbf{x_f}$ (frequency domain) are used as input.

3.3.1 Spatial Module. This module leverages f_s to learn the spatial features H^{spat}. Essentially, it extracts sensitive spatial features from RF matrices in both time domain $\mathbf{x}$ and frequency domain $\mathbf{x_f}$. We employ state-of-the-art deep learning model CNNs as the backbone f_b, and we empirically select the most cost-effective backbone in Section 4.3.1. Generally, CNNs are designed for exploring spatial features in image data. Therefore, we regard $\mathbf{x}$ and $\mathbf{x_f}$ as image with N_r tx-rx pair as input channels, L fast time dimension as height of input planes in pixels and K slow time dimension as width in pixels. Moreover, we additionally employ one convolutional layer

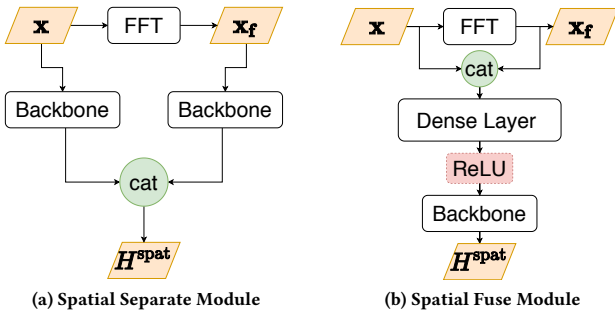(a) Spatial Separate Module
(b) Spatial Fuse Module
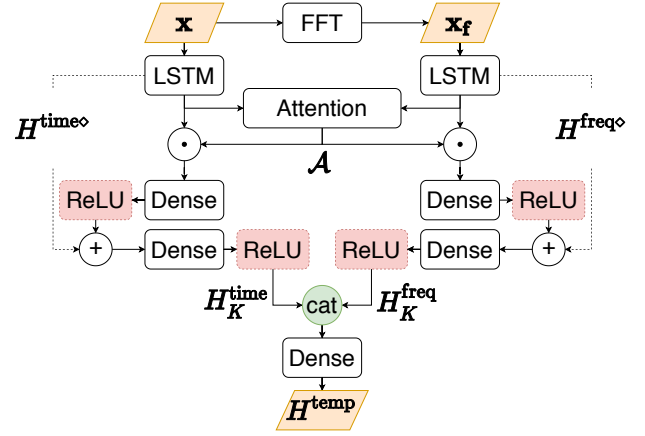
Figure 6: Two alternative spatial modules.

Figure 7: Attention-based temporal module.

f_a to adjust channel dimension into designated dimension declared by the backbone.

We explore two spatial modules, named spatial separate module and spatial fuse module. As depicted in Figure 6a, spatial separate module aims to employ backbone for generating spatial features of $\mathbf{x}$ and $\mathbf{x_f}$ separately: $H^{s,\mathbf{x}} = f_b(f_a(\mathbf{x}))$ and $H^{s,\mathbf{x_f}} = f_b(f_a(\mathbf{x_f}))$; the final spatial features H^{spat} are then produced by concatenating $H^{s,\mathbf{x}}$ and $H^{s,\mathbf{x_f}}$. However, spatial separate module involves a high computation complexity due to repeated backbone usage. Therefore, we further explore spatial fuse module illustrated in Figure 6b. As a simplification, we concatenate $\mathbf{x}$ and $\mathbf{x_f}$ at the initial stage $\mathbf{x_c} = [\mathbf{x}, \mathbf{x_f}] \in \mathbb{R}^{K \times L \times 2N_r}$. Afterwards, we intend to learn a composite representation from $\mathbf{x_c}$. We first reshape $\mathbf{x_c} \in \mathbb{R}^{K \times 2N_r L}$ and then we employ a *Rectified Linear Unit* (ReLU) activated dense layer to generate a composite representation $H^c \in \mathbb{R}^{K \times \alpha \times 2}$, where α denotes the hidden dimension. Afterwards, we employ backbone f_b to directly extract $H^{\text{spat}} = f_b(f_a(H^c))$. The corresponding performance comparisons between these two modules, along with the choice of α and activation function, are conducted in Section 4.

3.3.2 Attention-Based Temporal Module. In order to generate long-term temporal features from RF matrices in both time and frequency domains, we utilize f_t to learn attended time features H^{time} and attended frequency features H^{freq}. As illustrated in Figure 7, we first extract initial time features $H^{\text{time}\diamond}$ and frequency features $H^{\text{freq}\diamond}$, leveraging Long Short Term Memory (LSTM) to avoid gradient vanishing problems. Therefore, we reshape $\mathbf{x} \in \mathbb{R}^{K \times (L \times N_r)}$ and $\mathbf{x_f} \in \mathbb{R}^{K \times (L \times N_r)}$, regarding the slow time dimension as step index with cardinality K. Then, we extract $H^{\text{time}\diamond} \in \mathbb{R}^{K \times \alpha}$ and $H^{\text{freq}\diamond} \in \mathbb{R}^{K \times \alpha}$ by passing $\mathbf{x}$ and $\mathbf{x_f}$ into LSTM separately. After that, we focus on employing attention mechanism [29] to learn richer joint representations between initial feature representations, aiming to generate the attention map $\mathcal{A} \in \mathbb{R}^{K \times K}$ containing the joint information of every step between $H^{\text{time}\diamond}$ and $H^{\text{freq}\diamond}$. To be specific, ReLU activated dense layers are first employed to map $H^{\text{time}\diamond}$, $H^{\text{freq}\diamond}$ into $H^{\text{time}\dagger}$, $H^{\text{freq}\dagger} \in \mathbb{R}^{K \times \iota}$, where ι denotes hidden dimension. Then the attention map is generated as $\mathcal{A} := \text{softmax}(W \circ H^{\text{time}\dagger}(H^{\text{freq}\dagger})^T)$, where $\circ$ stands for Hadamard product, T denotes transpose, and $W \in \mathbb{R}^{K \times \iota}$ denotes weight matrix.

Combing the attention map $\mathcal{A}$ with the initial features produces joint representations $((H^{\text{time}\diamond})^T\mathcal{A})^T$ and $((H^{\text{freq}\diamond})^T\mathcal{A})^T$, and these joint representations are passed through ReLU activated dense layers for adding non-linearity. Especially, we separately incorporate these mapped joint representations into $H^{\text{time}\diamond}$, $H^{\text{freq}\diamond}$ via a residual connection to generate $H^{\text{time}} \in \mathbb{R}^{K\times\alpha}$, $H^{\text{freq}} \in \mathbb{R}^{K\times\alpha}$, thus it is able to solve vanishing gradients issue and achieve optimal identity mapping [18]. Subsequently, we choose final step (K-th) of attended time and frequency features, i.e., H_K^{time} and H_K^{freq}, for producing a composed temporal features H^{temp}. To be specific, we first employ two ReLU activated dense layers to generate H_K^{time} and H_K^{freq}. Then, we stack those mapped features followed by a dense layer to compute $H^{\text{temp}} \in \mathbb{R}^{1\times 2\alpha}$.

3.3.3 Classification Module. We have employed spatial module f_s and attention-based temporal module f_t to generate sensitive spatial features H^{spat} and attended temporal features H^{temp} from both time and frequency domains. In order to efficiently predict an activity label, we further derive an integrated feature by fusing both spatial and temporal representations. In this paper, we directly add spatial features H^{spat} and temporal features H^{temp} as $H^{\text{fuse}} = H^{\text{temp}} + H^{\text{spat}}$. To enable this integration, we set the hidden dimension α properly so as to equalize the dimensions of H^{spat} and H^{temp}. Finally, we utilize a dense layer to predict activity label $\hat{y} = H^{\text{fuse}}W_1$, where $W_1 \in \mathbb{R}^{2\alpha\times N_c}$ is weight matrix.

Remark: We choose the final step (K-th) of the attended temporal features H_K^{time} and H_K^{freq}, along with the fused features H^{fuse}, as $\{H_m^{\text{feat}}\}_{m=1}^M$ in RF-Net (see Section 3.2.1).

4 EXPERIMENTS

In this section, we conduct extensive experiments on all three RF sensing techniques, i.e., Wi-Fi, FMCW, and impulse radio (IR), aiming to demonstrate the efficacy of the dual-path base network for activity recognition and to evaluate the overall performance of RF-Net on performing one-shot RF-HAR. In particular, we report experiments on the following aspects: i) hyperparameter searching of network setting, ii) superiority of RF-Net and dual-path base network over baselines, and iii) efficacy of RF-Net and dual-path base network on various RF sensing datasets.

4.1 Datasets

As described in Section 2.1, RF signal in each sensing technique could be formulated as matrix $\mathbf{x} \in \mathbb{R}^{K\times L\times N_r}$. In this section, we first elaborate RF signal matrix of these RF sensing techniques. Then we describe environment information and human activities provided in each dataset. Given the variety of testing conditions, the activities of testing subjects may vary across different datasets. Details of each dataset is summarized in Table 1: for each RF sensing technology, we specify the number of environments involved, the number of observations per environment taken, and the number of activities tested. To generate different environments, we first select a few rooms with distinct sizes, also involve different testing subjects. Within the same room, we change positions of the furniture and appliances, as well as the location of the subject. In order to artificially create "differences", we make sure that at least

Table 1: Datasets information.

Sensing	Environments #	Observations #	Activities #
Wi-Fi	80	25	6
Wi-Fi	100	20	6
Wi-Fi	120	16	6
FMCW	10	17	6
IR	50	16	6

five objects (including the subject) have their position changed in a room when generating a new environment. Due to the popularity of Wi-Fi sensing, we choose to employ three datasets for it, but only one dataset for FMCW and IR.

4.1.1 Wi-Fi. We exploit CSI of 30 OFDM subcarriers with 2 tx-rx pairs to record six human activities: wiping, walking, moving, rotating, sitting, and standing up. CSI information is sampled at 100 Hz and conducted window slicing size in 5.12 s. Therefore, each signal matrix is with a total $K = 512$ received packets, $L = 30$ subcarriers, and $N_r = 2$ tx-rx pairs. We employ 11 subjects and record from 6 different rooms. In the following experiments, we intend to investigate the impact of the number of environments as well. Therefore, we split these environments data into three datasets including 80 environments, 100 environments, and 120 environments. As described in Table 1, these datasets have 25, 20, and 16 observations per environment per activity, respectively.

4.1.2 FMCW. We utilize FMCW radar device with 1 tx-rx pair to collect data. To be specific, we collect each frame with 253 frequency components every 67 ms. Then, we stack 100 frames for covering a 6.7 s interval. Therefore, each FMCW matrix $\mathbf{x} \in \mathbb{R}^{K\times L\times N_r}$ is with a total $K = 100$ frames, $L = 253$ frequency components, and $N_r = 1$ tx-rx pair. These matrices depict that 9 subjects in 2 different rooms perform six activities: standing up, sitting down, going out, entering room, putting on clothes, and putting off clothes. We aim to keep consistency on the number of observations per environment per activity of all datasets. Therefore, we split observations into 10 different environments.

4.1.3 IR. We employ IR device with 1 tx-rx pair to transmit pulse signal. Specifically, we collect each frame with 138 time components every 2.5 ms and we stack 400 frames together. Accordingly, each IR matrix $\mathbf{x} \in \mathbb{R}^{K\times L\times N_r}$ is with total $K = 400$ frames, $L = 138$ samples of a pulse, and $N_r = 1$ tx-rx pair. We employ 20 subjects in 3 rooms to perform six activities : sitting down, standing up, walking, falling, bending, and lying. We split all the observations into 50 different environments.

4.2 Baselines & Implementation

We present several baseline schemes against which our RF-Net and its base network will be compared. We also briefly explain the implementation of our experiments.

4.2.1 Baselines For Dual-path Base Network. We compare our proposed base network with the following baseline networks. We aim to demonstrate the efficacy of our base network on extracting features from RF signal matrix for HAR.

- TIME: the proposed spatial module with only time domain data. It is used to evaluate network with spatial features in time domain.
- FREQ: the proposed spatial module with only frequency domain data. It is used to investigate the performance of network with spatial features in frequency domain.
- SPATIAL-SEP: the proposed spatial separate module. It is used to investigate the performance of network with spatial features in both time and frequency domain.
- SPATIAL-FUSE: the proposed spatial fuse module. It is used to examine network with spatial features in both time and frequency domain.
- DUAL-wo-ATT: the proposed dual-path base network without attention mechanism. It is used to demonstrate the efficacy of the attention mechanism.
- DUAL-PATH: the proposed dual-path base network.

Note that, as SPATIAL-SEP and SPATIAL-FUSE are two alternative components, DUAL-wo-ATT and DUAL-PATH will choose to involve whichever performs better.

4.2.2 Baselines For RF-Net. We evaluate the performance of one-shot human activity recognition when RF-Net is adapted to new environments. Therefore, we mainly compare our network with the following meta-learning baselines and fine-tuning baseline.

- Fine-Tuning (FT): FT firstly trains dual-path base network with the training set. When FT adapts the network on the test set, the parameters of classification module would be fine-tuned with support observations on the test set.
- MAML: a state-of-the-art optimization-based meta-learning baseline. This scheme has been adopted by Metasense [14] to enable adaptive wearable sensing system. MAML relies on meta-optimization through gradient descent in a model agnostic way. It expects to learn an initial representation that can be fine-tuned efficiently in a few steps.
- Prototypical Network (PN): a state-of-the-art metric-based meta-learning baseline. Given few observations, PN generates prototypes (feature representations) for each class and it uses the Euclidean distance metric to search the closest prototype as predicted class.
- RF-Net*: our proposed meta-learning framework without the RF metric module.
- RF-Net: our proposal. The comparison between RF-Net and RF-Net* highlights the impact of RF metric module.

4.2.3 Implementation. We first apply data normalization to all datasets. Afterwards, we sample 80% environments of each dataset into training dataset $\mathcal{D}$ and the remaining into testing dataset $\hat{\mathcal{D}}$, and we perform a 10-fold cross-validation for each experiment. We report the average accuracy evaluated on the testing dataset under 1-, 2-, 3-shot scenarios, i.e., $N_s = \{1, 2, 3\}$. We use regularization technique batch normalization [23] to avoid overfitting. The following experiments are all programmed with PyTorch [45] and run on NVIDIA TESLA V100 with 16GB memory.

4.3 Hyperparameter Searching

We first conduct hyperparameter searching of dual-path base network, and thus we could employ the final hyperparameters in the

following experiments for the fairness comparison. To be specific, we explore the following aspects: selection of CNN backbone, hidden dimension size, as well as activation function selection. In this section, we explore the aspects on the dual-path base network on the Wi-Fi dataset with 100 environments.

4.3.1 Backbone Selection. Since our base network leverages CNN backbone f_b for extracting spatial features, the selection of CNN-backbone is essential for building a good base network. Therefore, we evaluate multiple state-of-the-art CNN backbones with various model complexity implanted into our base network. As depicts in Figure 8, these include two shallow networks: 5 layers CNN (cnn5) (with 3 convolutional layers and 2 fully connected layers) and 8 layers AlexNet (alexnet8) [32], two medium deep networks: 18 layers ResNet (resnet18) [18] and 19 layers VGG (vgg19) [49], as well as two deep networks: 44 layers ShuffleNet (shufflenet44) [71] and 152 layers ResNet (resnet152) [18]. We observe that our choice
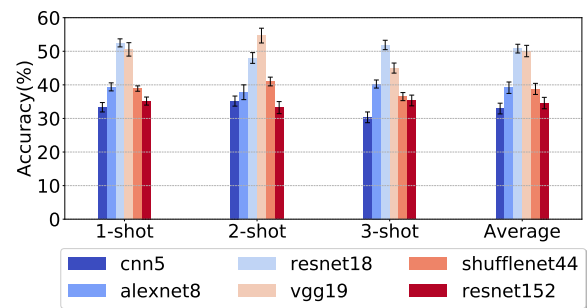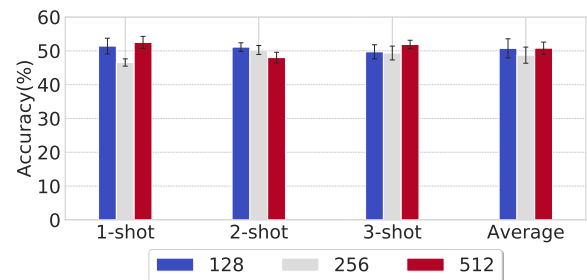

Figure 8: CNN backbone selection.

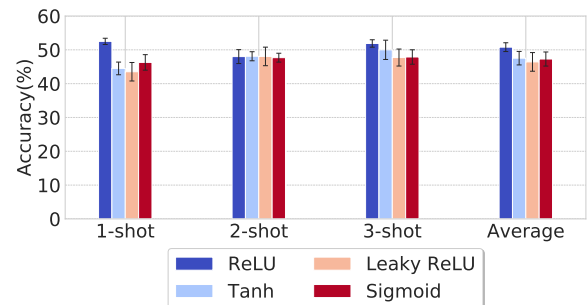**Figure 9: Hidden dimension size.**

Figure 10: Activation function selection.

resnet18 achieves the best overall accuracy compared with other backbones, particularly for one-shot HAR. We further study the model overhead (complexity, inference time, and memory usage) of resnet. As the complexity of resnet18 takes up around 85% of the whole system, its overhead can be roughly regarded as the system overhead. According to a recent benchmark [5], resnet18, with a low model complexity and memory usage, achieves a low inference time, leading to an excellent real-time performance. Moreover, compared with other light-weight model such as shufflenet44, resnet18 achieves a comparable performance when it is deployed to commercial edge devices. Therefore, we select pre-trained ResNet 18 layers network as the CNN backbone for RF-Net.

4.3.2 Hidden Dimension Size & Activation Function Selection. As described in Section 3.3, we use a hidden variable α to compose our network. Therefore, we aim to search a good dimension size. We separately explore three settings $\alpha = 128$, $\alpha = 256$, and $\alpha = 512$, and report the accuracy in Figure 9. Although $\alpha = 128$ is compared well with $\alpha = 512$ in average, $\alpha = 512$ outperforms $\alpha = 128$ on one-shot HAR. In this paper, one-shot HAR takes precedence, and thus we set α to 512. Afterwards, we also search the best activation function in our network. We explore ReLU, Leaky ReLU, Sigmoid, and Tanh. Figure 10 plots the performance. We observe that ReLU outperforms other function and we set all the activation function to ReLU. Therefore, we will employ pre-trained ResNet with 18 layers as our CNN backbone, set hidden dimension to 512 and select ReLU as activation function in the following experiments.

4.4 Superiority of Proposed Network

In this section, we will investigate the superiority of dual-path base network and RF-Net over baselines on various RF sensing datasets. Specifically, we set following parameters: M' to 12, batch size to 3, learning rate to 0.001, the number of epochs to 20 and we employ Adam optimizer to train networks.

4.4.1 Superiority of Dual-path base Network over Baselines. Figure 11, Figure 12, and Figure 13 separately report the accuracy of dual-path base network and the baselines on Wi-Fi, FMCW, and IR datasets. In all the cases, our base network outperforms all the baselines. It shows the effectiveness of both spatial and temporal features extracted by the dual-path base network for the purpose of HAR. To be specific, the performance of SPATIAL-FUSE outperforms TIME and FREQ remarkably. It validates our motivation that the utilization of RF signal matrix in dual-path (both time and frequency domain) is beneficial for HAR. In our experiments, we find that SPATIAL-FUSE is far better than SPATIAL-SEP when accuracy and time complexity are taken into account. One possible reason is that using the CNN backbone twice doubles the number of parameters in the network, and makes SPATIAL-SEP relatively harder to train and converge.

In most cases, we observe that the performance of DUAL-wo-ATT is still unacceptable in spite of already involving temporal features. By contrast, the outstanding performance of dual-path base network demonstrates the efficacy of attended temporal features. We further plot attention map $\mathcal{A}$ of two different activities on FMCW datasets in Figure 14. This showcases that the attention mechanism used in the temporal module is able to discover a rich
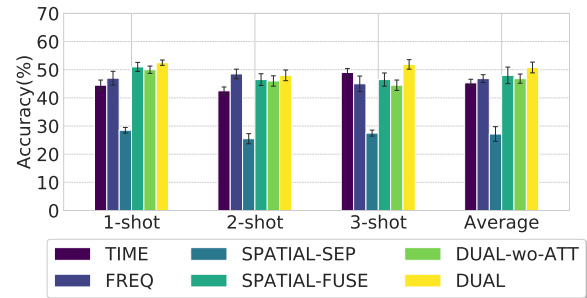

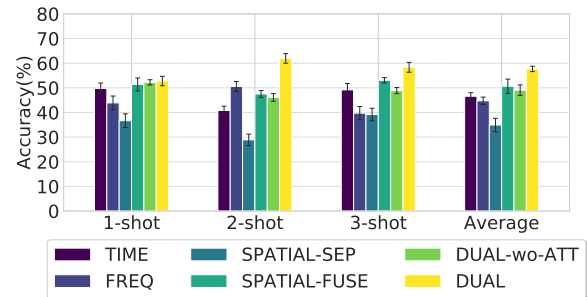**Figure 11: Overall comparison results for dual-path base network evaluation on Wi-Fi dataset.**

Figure 12: Overall comparison results for dual-path base network evaluation on FMCW dataset.
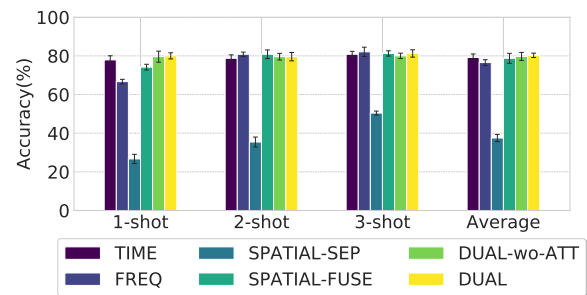

Figure 13: Overall comparison results for dual-path base network evaluation on IR dataset.

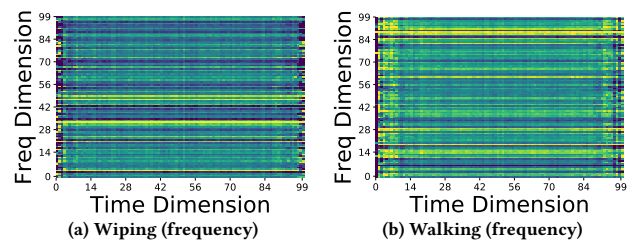(a) Wiping (frequency)   (b) Walking (frequency)

Figure 14: FMCW attention map $\mathcal{A}$ of two different activities: walking and sitting down.

joint temporal representations across two domains, thus improving the performance of HAR. In general, the accuracy of our dual-path base network outperforms DUAL-wo-ATT by around 4.2%.
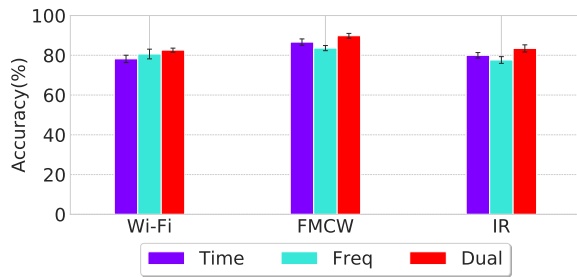
Figure 15: Overall comparison results for dual-path base network evaluation on Wi-Fi, FMCW, IR datasets in terms of general classification.

Remark: One may intuitively expect a monotonic increase in accuracy given more shots. However, as the base network is designed for RF-HAR in better exploiting the input of high-dimensional RF signal matrices, its performance is optimized for general classification to fit into the overall meta learning framework. Therefore, the fluctuation in accuracy across a few shots is reasonable and can be expected. We further demonstrate the efficacy of dual-path base network for general classification. As shown in Figure 15, dual-path base network performs still outstanding consistently in general classification on all RF datasets. Figure 15 (albeit not a fair comparison due to the separated datasets for different RF sensing techniques) also sheds a light on the intrinsic differences among the three RF signals. Wi-Fi with narrowband cannot provide a high resolution in time domain, but Doppler shift can still be retained in frequency domain. Therefore, the overall performance of Wi-Fi is worse than the others, but the accuracy in frequency domain is better than that in time domain. Resorting to the power of wideband, FMCW and IR both achieve better performance due to a higher time resolution, yet FMCW slightly outperforms IR thanks to a higher tx power.

4.4.2 Superiority of RF-Net over Baselines. The comparison results for RF-Net on Wi-Fi, FMCW, and IR are listed in Figures 16, 17, and 18. We observe that our proposed RF-Net network has proved the superiority over baselines on all three RF sensing techniques. In particular, compared RF-Net with RF-Net*, overall performance is enhanced by RF metric module and relatively steady. It highlights the efficacy of RF metric module. This performance improvement gives credit to the trainable linear mapping of RF metric module. As demonstrated in Section 3.2.1, its convex loss function should allow the RF metric module to learn an optimal metric for RF-Net. One may argue that all three RF techniques achieve a rather low accuracy, but this is the best one may obtain so far when adapting to new environments with very few labelled observations; in fact, our performance is already better than that for image recognition [8], possibly thanks to the depth information obtained by RF sensing.

In Figure 19, we further plot the accuracy density plot of RF-Net along with the dual-path base network and all the baselines. The best performance is highlighted in bold. The density plot depicts the accuracy distribution on all three RF datasets with $N_s = \{1, 2, 3\}$. Intuitively, dual-path base network notably outperforms its baselines. However, it is still far below the expectation. Compared with dual-path base network, the density plot of RF-Net along with its meta-learning baselines and fine-tuning baselines exhibit some notable
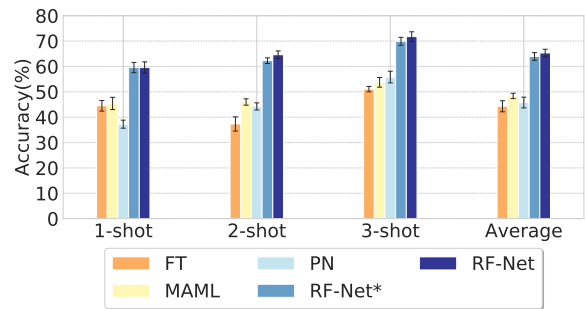

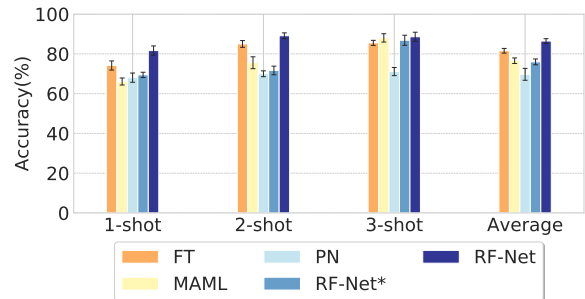**Figure 16: Overall comparison results for RF-Net evaluation on Wi-Fi dataset.**

Figure 17: Overall comparison results for RF-Net evaluation on FMCW dataset.
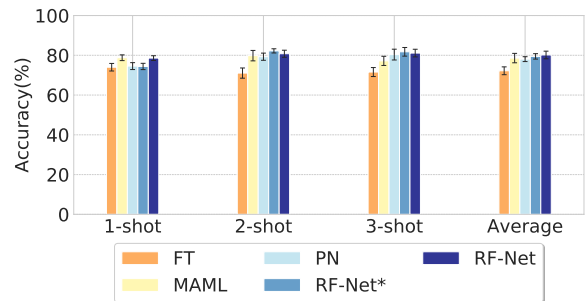

Figure 18: Overall comparison results for RF-Net evaluation on IR dataset.

right skew. It indicates that RF-Net and its baselines as expected entail significantly performance improvement of the implanted base network. Nevertheless, RF-Net boosts the performance maximally among them: it has greatly helped the base network adapting to new environment with even a single observation.

4.4.3 Wi-Fi vs. IR vs. FMCW. To further evaluate the performance on each RF sensing dataset, we first plot the performance of the proposed networks and their baselines on Wi-Fi, IR, and FMCW in Figure 20. Meanwhile, we further generate a more representative performance heatmap in Figure 21 to help visualizing varying degrees of improved performance by RF-Net in each RF sensing dataset. From these two figures, we observe that RF-Net achieves the highest overall performance in all datasets, and the performance of all concerned networks differ significantly across different RF
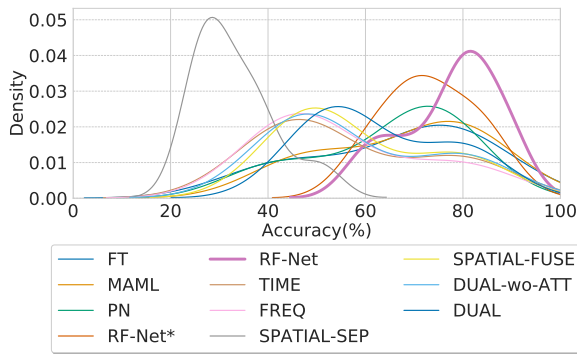
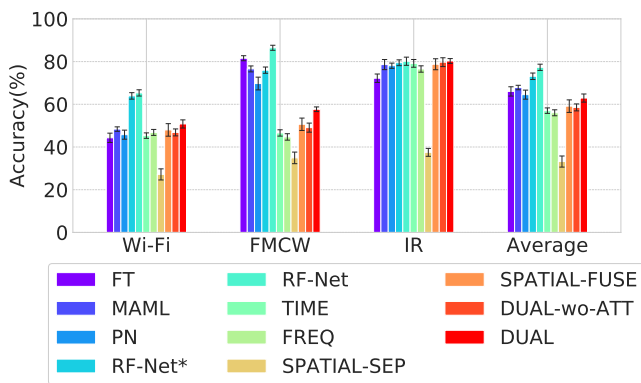Figure 19: Accuracy density plot of all evaluated networks.

Figure 20: Performance on RF sensing types.

sensing datasets. One curious observation on the evaluation over IR dataset is that all networks seem to achieve similar performances with and without meta-learning. The reason is that the intrinsic advantage of IR being not sensitive to environment impact (thus allowing the environment-free features to be captured via base network), yet RF-Net still performs the best among all networks. To summarize, based on the average accuracy shown in Figures 20 and 21, our proposed RF-Net are demonstrated to offer the most
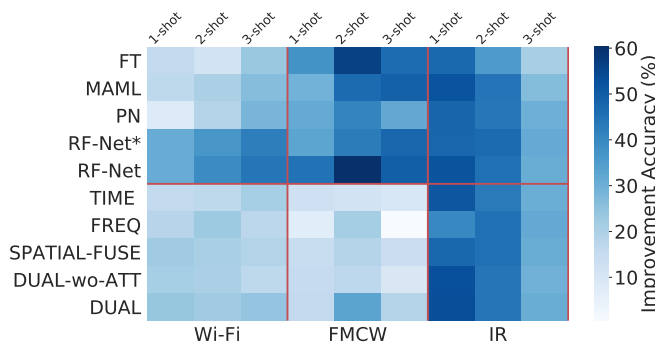

Figure 21: Performance heatmap: accuracy improvement from SPATIAL-SEP, while a darker color denotes a higher accuracy improvement.
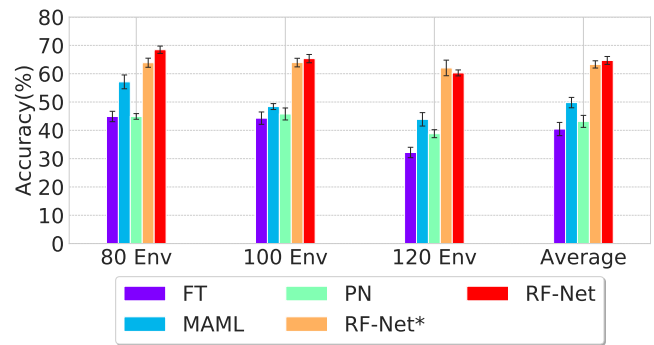

Figure 22: Impact of number of environments.

robust generalization that may fit all data intrinsic properties of all RF signals.

4.4.4 Impact of Number of Environments. We examine networks from the perspective of environmental diversity as well. The motivation is to explore how many environments could be handled for each network. We report average accuracy of RF-Net and its baselines on Wi-Fi dataset with 80, 100 and 120 environments in Figure 22. With the number of environments increasing, we observe that the performance of both fine-tuning and meta-learning baselines (MAML and PN) have clearly dropped. Our RF-Net show more steady trends. This performance indicates that RF-Net can deal with more environments without sacrificing performance severely, i.e., it is a more robust approach.

5 CONCLUSION

Based on the study of representative RF sensing techniques along with major meta-learning approaches, we have proposed RF-Net, a meta-learning based neural network for one-shot RF-based HAR; it contributes to the capability of being fast adapted to new environments with a single observation. RF-Net consists of a meta-learning framework that involves a parametric RF-specific module for training a powerful distance metric, and a dual-path base network that fully exploits the high-dimensional features contained in the RF signal matrix. We have conducted extensive experiments on all three RF sensing techniques. These experiments have demonstrated the efficacy of our proposed RF-Net and dual-path base network.

As a potential future direction, we are looking into extending our meta-learning frameworks to various sensing applications that can be heavily affected by environment changes. These applications include, among others, radio frequency sensing for, e.g., vibration detection and indoor localization [9, 61, 68], acoustic sensing for similar purposes [6, 7, 40], and visible light sensing for, e.g., information decoding [64, 67] and occupancy inference [65, 66]. As these applications may involve rather different sensing methodologies and modalities, both the base network and the meta-learning framework will need to be substantially revamped.

6 ACKNOWLEDGEMENTS

REFERENCES

[1] Abdelnasser, H., Harras, K., and Youssef, M. A Ubiquitous WiFi-Based Fine-Grained Gesture Recognition System. *IEEE Transactions on Mobile Computing 18*, 11 (2019), 2474–2487.

[2] Adib, F., Kabelac, Z., and Katabi, D. Multi-Person Localization via RF Body Reflections. In *Proc. of the 11st USENIX NSDI* (2015), p. 279–292.

[3] Ali, K., Liu, A. X., Wang, W., and Shahzad, M. Keystroke Recognition Using WiFi Signals. In *Proc. of the 21st ACM MobiCom* (2015), pp. 90–102.

[4] Bi, C., Xing, G., Hao, T., Huh-Yoo, J., Peng, W., Ma, M., and Chang, X. Family-Log: Monitoring Family Mealtime Activities by Mobile Devices. *IEEE Transactions on Mobile Computing 19*, 8 (2020), 1818–1830.

[5] Bianco, S., Cadene, R., Celona, L., and Napoletano, P. Benchmark Analysis of Representative Deep Neural Network Architectures. *IEEE Access 6* (2018), 64270–64277.

[6] Cai, C., Chen, Z., Pu, H., Ye, L., Hu, M., and Luo, J. AcuTe: Acoustic Thermometer Empowered by a Single Smartphone. In *Proc. of the 18th ACM SenSys* (2020), pp. 1–14. https://doi.org/10.1145/3384419.3430714.

[7] Cai, C., Pu, H., Hu, M., Zheng, R., and Luo, J. SST: Software Sonic Thermometer on Acoustic-enabled IoT Devices. *IEEE Transactions on Mobile Computing* (2020), 1–14.

[8] Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C. F., and Huang, J.-B. A Closer Look at Few-shot Classification. In *Proc. of the 7th ICLR* (2019), pp. 1–16.

[9] Chen, Z., Li, Z., Zhang, X., Zhu, G., Xu, Y., Xiong, J., and Wang, X. AWL: Turning Spatial Aliasing From Foe to Friend for Accurate WiFi Localization. In *Proc. of the 13th ACM CoNEXT* (2017), pp. 238–250.

[10] Chen, Z., Zhang, L., Jiang, C., Cao, Z., and Cui, W. WiFi CSI Based Passive Human Activity Recognition Using Attention Based BLSTM. *IEEE Transaction on Mobile Computing 18*, 11 (2018), 2714–2724.

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

[12] Finn, C., Abbeel, P., and Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proc. of the 34th ICML* (2017), pp. 1126–1135.

[13] Ghose, A., Chakravarty, K., Agrawal, A. K., and Ahmed, N. Unobtrusive Indoor Surveillance of Patients at Home Using Multiple Kinect Sensors. In *Proc. of the 11st ACM SenSys* (2013), pp. 1–2.

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

[15] Halperin, D., Hu, W., Sheth, A., and Wetherall, D. Tool Release: Gathering 802.11n Traces with Channel State Information. *ACM SIGCOMM Computer Communication Review 41*, 1 (2011), 53–53.

[16] Hao, T., Bi, C., Xing, G., Chan, R., and Tu, L. MindfulWatch: A Smartwatch-Based System For Real-Time Respiration Monitoring During Meditation. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1*, 3 (Sept. 2017).

[17] Hao, T., Xing, G., and Zhou, G. iSleep: Unobtrusive Sleep Quality Monitoring Using Smartphones. In *Proc. of the 11st ACM SenSys* (2013), pp. 1–14.

[18] He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. In *Proc. of the 29th IEEE CVPR* (2016), pp. 770–778.

[19] He, Y., Liang, J., and Liu, Y. Pervasive Floorplan Generation Based on Only Inertial Sensing: Feasibility, Design, and Implementation. *IEEE Journal on Selected Areas in Communications 35*, 5 (2017), 1132–1140.

[20] Hnat, T. W., Srinivasan, V., Lu, J., Sookoor, T. I., Dawson, R., Stankovic, J., and Whitehouse, K. The Hitchhiker's Guide to Successful Residential Sensing Deployments. In *Proc. of the 9th ACM SenSys* (2011), pp. 232–245.

[21] Hu, J.-F., Zheng, W.-S., Lai, J., and Zhang, J. Jointly Learning Heterogeneous Features for RGB-D Activity Recognition. In *Proc. of the 28th IEEE CVPR* (2015), pp. 5344–5352.

[22] Humenberger, M., Schraml, S., Sulzbachner, C., Belbachir, A. N., Srp, A., and Vajda, F. Embedded Fall Detection with A Neural Network and Bio-Inspired Stereo Vision. In *Proc. of the 25th IEEE CVPR Workshops* (2012), pp. 60–67.

[23] Ioffe, S., and Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proc. of the 32th ICML Workshop* (2015), pp. 448–456.

[24] Jalal, A., Uddin, M. Z., and Kim, T.-S. Depth Video-Based Human Activity Recognition System Using Translation and Scaling Invariant Features for Life Logging at Smart Home. *IEEE Transaction on Consumer Electronics 58*, 3 (2012), 863–871.

[25] Jamal, M. A., and Qi, G.-J. Task Agnostic Meta-Learning for Few-shot Learning. In *Proc. of the 32nd IEEE CVPR* (2019), pp. 11719–11727.

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

[27] Keally, M., Zhou, G., Xing, G., Wu, J., and Pyles, A. PBN: Towards Practical Activity Recognition Using Smartphone-Based Body Sensor Networks. In *Proc. of the 9th ACM SenSys* (2011), pp. 246–259.

[28] Kempe, V. *Inertial MEMS: Principles and Practice*. Cambridge University Press, 2011.

[29] Kim, J.-H., Jun, J., and Zhang, B.-T. Bilinear Attention Networks. In *Proc. of the 32nd NIPS* (2018), pp. 1564–1574.

[30] Kim, S. Y., Han, H. G., Kim, J. W., Lee, S., and Kim, T. W. A Hand Gesture Recognition Sensor Using Reflected Impulses. *IEEE Sensors Journal 17*, 10 (2017), 2975–2976.

[31] Koch, G., Zemel, R., and Salakhutdinov, R. Siamese Neural Networks for One-Shot Image Recognition. In *Proc. of the 32nd ICML Workshop* (2015), vol. 2.

[32] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet Classification with Deep Convolutional Neural Networks. In *Proc. of the 26th NIPS* (2012), pp. 1097–1105.

[33] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. of the IEEE 86*, 11 (1998), 2278–2324.

[34] Li, Z., Zhou, F., Chen, F., and Li, H. Meta-SGD: Learning to Learn Quickly for Few-shot Learning. *arXiv preprint arXiv:1707.09835* (2017).

[35] Lin, J., Gan, C., and Han, S. TSM: Temporal Shift Module for Efficient Video Understanding. In *Proc. of the 33rd IEEE ICCV* (2019), pp. 7083–7093.

[36] Liu, X., Ghosh, P., Ulutan, O., Manjunath, B., Chan, K., and Govindan, R. Caesar: Cross-Camera Complex Activity Recognition. In *Proc. of the 17th ACM SenSys* (2019), pp. 232–244.

[37] Ma, M., Fan, H., and Kitani, K. M. Going Deeper into First-Person Activity Recognition. In *Proc. of the 29th IEEE CVPR* (2016), pp. 1894–1903.

[38] Ma, Y., Zhou, G., Wang, S., Zhao, H., and Jung, W. SignFi: Sign Language Recognition Using WiFi. In *Proc. of the 18th ACM UbiComp* (2018), pp. 23:1–21.

[39] Mahafza, B. R. *Radar Systems Analysis and Design Using MATLAB*. CRC press, 2002.

[40] Mao, W., Zhang, Z., Qiu, L., He, J., Cui, Y., and Yun, S. Indoor Follow Me Drone. In *Proc. of the 15th ACM MobiSys* (2017), pp. 345–358.

[41] McIntosh, J., Marzo, A., Fraser, M., and Phillips, C. EchoFlex: Hand Gesture Recognition Using Ultrasound Imaging. In *Proc. of the 29th ACM CHI* (2017), pp. 1923–1934.

[42] Munkhdalai, T., and Yu, H. Meta Networks. In *Proc. of the 34th ICML* (2017), pp. 2554–2563.

[43] Nichol, A., and Schulman, J. Reptile: A Scalable Metalearning Algorithm. *arXiv preprint arXiv:1803.02999 2* (2018), 2.

[44] Park, J., and Cho, S. H. IR-UWB Radar Sensor for Human Gesture Recognition by Using Machine Learning. In *Proc. of IEEE HPCC-SmartCity-DSS* (2016), pp. 1246–1249.

[45] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Proc. of the 32nd NIPS* (2019), pp. 8024–8035.

[46] Rusu, A. A., Rao, D., Sygnowski, J., Vinyals, O., Pascanu, R., Osindero, S., and Hadsell, R. Meta-Learning with Latent Embedding Optimization. *Proc. of the 7th ICLR* (2019).

[47] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., and Lillicrap, T. Meta-Learning with Memory-Augmented Neural Networks. In *Proc. of the 33rd ICML* (2016), pp. 1842–1850.

[48] Sigg, S., Scholz, M., Shi, S., Ji, Y., and Beigl, M. RF-Sensing of Activities from Non-Cooperative Subjects in Device-Free Recognition Systems Using Ambient and Local Signals. *IEEE Transaction on Mobile Computing 13*, 4 (2013), 907–920.

[49] Simonyan, K., and Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proc. of the 3rd ICLR* (2015).

[50] Snell, J., Swersky, K., and Zemel, R. Prototypical Networks for Few-shot Learning. In *Proc. of the 31st NIPS* (2017).

[51] Sun, Q., Liu, Y., Chua, T.-S., and Schiele, B. Meta-Transfer Learning for Few-shot Learning. In *Proc. of the 32nd IEEE CVPR* (2019), pp. 403–412.

[52] Tian, Y., Lee, G.-H., He, H., Hsu, C.-Y., and Katabi, D. RF-Based Fall Monitoring Using Convolutional Neural Networks. In *Proc. of the 18th ACM UbiComp* (2018), pp. 137:1–24.

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

[54] Tse, D., and Viswanath, P. *Fundamentals of Wireless Communication*. Cambridge University Press, USA, 2005.

[55] Venkatnarayan, R. H., Page, G., and Shahzad, M. Multi-User Gesture Recognition Using WiFi. In *Proc. of the 16th ACM MobiSys* (2018), pp. 401–413.

[56] Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., and Wierstra, D. Matching Networks for One Shot Learning. In *Proc. of the 30th NIPS* (2016), pp. 3630–3638.

[57] Virmani, A., and Shahzad, M. Position and Orientation Agnostic Gesture Recognition Using WiFi. In *Proc. of the 15th ACM MobiSys* (2017), pp. 252–264.

[58] Wang, A., and Gollakota, S. Millisonic: Pushing the Limits of Acoustic Motion Tracking. In *Proc. of the 31st ACM CHI* (2019), pp. 1–11.

[59] Wang, W., Liu, A. X., Shahzad, M., Ling, K., and Lu, S. Understanding and

Modeling of WiFi Signal Based Human Activity Recognition. In *Proc. of the 21st ACM MobiCom* (2015), pp. 65–76.

[60] Wang, Y., Wu, K., and Ni, L. M. WiFall: Device-Free Fall Detection by Wireless Networks. In *Proc. of the 33rd IEEE INFOCOM* (2014), pp. 271–279.

[61] Wang, Z., Chen, Z., Singh, A., Garcia, L., Luo, J., and Srivastava, M. UWHear: Through-wall Extraction and Separation of Audio Vibrations Using Wireless Signals. In *Proc. of the 18th ACM SenSys* (2020), pp. 1–14. https://doi.org/10.1145/3384419.3430772.

[62] Wu, D., Zhang, D., Xu, C., Wang, H., and Li, X. Device-Free WiFi Human Sensing: From Pattern-Based to Model-Based Approaches. *IEEE Communications Magazine 55*, 10 (2017), 91–97.

[63] Xu, X., Yu, J., Chen, Y., Zhu, Y., Kong, L., and Li, M. BreathListener: Fine-Grained Breathing Monitoring in Driving Environments Utilizing Acoustic Signals. In *Proc. of the 17th ACM MobiSys* (2019), pp. 54–66.

[64] Yang, Y., Hao, J., Luo, J., and Pan, S. CeilingCast: Energy Efficient and Location-Bound Broadcast Through LED-Camera Communication. In *Proc. of the 35th IEEE INFOCOM* (2016), pp. 1–9.

[65] Yang, Y., Hao, J., Luo, J., and Pan, S. J. CeilingSee: Device-Free Occupancy Inference through Lighting Infrastructure based LED Sensing. In *Proc. of the 15th IEEE PerCom* (2017), p. 247–256.

[66] Yang, Y., Luo, J., Hao, J., and Pan, S. J. Counting via LED Sensing: Inferring Occupancy Using Lighting Infrastructure. *Elsevier Pervasive and Mobile Computing 45* (2018), 35 – 54.

[67] Yang, Y., Nie, J., and Luo, J. ReflexCode: Coding with Superposed Reflection Light for LED-Camera Communication. In *Proc. of the 23rd ACM MobiCom* (2017), p. 193–205.

[68] Zhang, C., Li, F., Luo, J., and He, Y. iLocScan: Harnessing Multipath for Simultaneous Indoor Source Localization and Space Scanning. In *Proc. of the 12th ACM SenSys* (2014), p. 91–104.

[69] Zhang, J., Tang, Z., Li, M., Fang, D., Nurmi, P., and Wang, Z. CrossSense: Towards Cross-Site and Large-Scale WiFi Sensing. In *Proc. of the 24th ACM MobiCom* (2018), pp. 305–320.

[70] Zhang, X., Yao, L., Huang, C., Wang, S., Tan, M., Long, G., and Wang, C. Multi-Modality Sensor Data Classification with Selective Attention. In *Proc. of the 27th IJCAI* (2018), pp. 3111–3117.

[71] Zhang, X., Zhou, X., Lin, M., and Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In *Proc. of the 31st IEEE CVPR* (2018), pp. 6848–6856.

[72] Zhao, M., Li, T., Abu Alsheikh, M., Tian, Y., Zhao, H., Torralba, A., and Katabi, D. Through-Wall Human Pose Estimation Using Radio Signals. In *Proc. of the 31st IEEE CVPR* (2018), pp. 7356–7365.

[73] Zhao, M., Tian, Y., Zhao, H., Alsheikh, M. A., Li, T., Hristov, R., Kabelac, Z., Katabi, D., and Torralba, A. RF-Based 3D Skeletons. In *Proc. of ACM SIGCOMM* (2018), p. 267–281.

[74] Zheng, T., Chen, Z., Cai, C., Luo, J., and Zhang, X. V^{2}iFi: in-Vehicle Vital Sign Monitoring via Compact RF Sensing. In *Proc. of the 20th ACM UbiComp* (2020), pp. 70:1–27.

[75] Zheng, Y., Zhang, Y., Qian, K., Zhang, G., Liu, Y., Wu, C., and Yang, Z. Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi. In *Proc. of the 17th ACM MobiSys* (2019), pp. 313–−325.

[76] Zhou, X., Huang, Q., Sun, X., Xue, X., and Wei, Y. Towards 3D Human Pose Construction in the Wild: a Weakly-Supervised Approach. In *Proc. of the 26th ACM MobiCom* (2020).