MobHAR: Source-free Knowledge Transfer for Human Activity Recognition on Mobile Devices

MENG XUE, The Hong Kong University of Science and Technology, Hong Kong SAR
YINAN ZHU, The Hong Kong University of Science and Technology, Hong Kong SAR
WENTAO XIE, The Hong Kong University of Science and Technology, Hong Kong SAR
ZHIXIAN WANG, Wuhan University, China
YANJIAO CHEN, Zhejiang University, China
KUI JIANG, Harbin Institute of Technology, China
QIAN ZHANG*, The Hong Kong University of Science and Technology, Hong Kong SAR

Human Activity Recognition (HAR) faces significant challenges when deployed in real-world scenarios due to non-independent and identically distributed (non-IID) data distributions. While existing domain adaptation (DA) approaches attempt to address this issue, they either require access to source data or struggle with large domain shifts. This paper presents a novel source-free domain adaptation framework for HAR that effectively handles substantial domain discrepancies across different datasets. Our approach introduces two key innovations: (1) a Discriminative Information Gramian (DIG) method that quantifies the relationship between target-domain samples and the source domain without requiring access to source data, and (2) an unsupervised domain generalization technique that ensures consistent feature extraction across augmented data samples, enhancing the model's effectiveness in the target domain. We evaluate our framework on five diverse HAR datasets comprising 87 users with varying demographics, devices, and environmental conditions. In single-source scenarios, our method achieves 76.77% accuracy and 67.03% F1-score, surpassing state-of-the-art solutions by 9.77% and 17.43%, respectively. For multi-source scenarios, we attain 85.33% accuracy and 79.55% F1-score, exceeding existing methods by at least 8.8% and 14.8%, respectively. This work represents the first successful attempt at dataset-level HAR domain adaptation without access to source data, marking a significant advancement in practical HAR applications.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**.

Additional Key Words and Phrases: Human activity recognition, Source-free knowledge transfer, Discriminative information gramian

ACM Reference Format:
Meng Xue, Yinan Zhu, Wentao Xie, Zhixian Wang, Yanjiao Chen, Kui Jiang, and Qian Zhang. 2025. *MobHAR*: Source-free Knowledge Transfer for Human Activity Recognition on Mobile Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 1, Article 22 (March 2025), 24 pages. https://doi.org/10.1145/3712620

*Corresponding author.

Authors' Contact Information: Meng Xue, csexuemeng@ust.hk, The Hong Kong University of Science and Technology, Hong Kong SAR; Yinan Zhu, yzhudf@cse.ust.hk, The Hong Kong University of Science and Technology, Hong Kong SAR; Wentao Xie, wentaox@ust.hk, The Hong Kong University of Science and Technology, Hong Kong SAR; Zhixian Wang, hirowong@whu.edu.cn, Wuhan University, China; Yanjiao Chen, chenyanjiao@zju.edu.cn, Zhejiang University, China; Kui Jiang, jiangkui@hit.edu.cn, Harbin Institute of Technology, China; Qian Zhang, qianzh@cse.ust.hk, The Hong Kong University of Science and Technology, Hong Kong SAR.

ACM 2474-9567/2025/3-ART22
https://doi.org/10.1145/3712620

1 INTRODUCTION

Human activity recognition (HAR) is an increasingly important field with significant implications for applications such as healthcare, smart home, and entertainment [3, 14, 15, 21, 26, 27, 40, 57]. The primary goal of HAR is to monitor physical activity accurately. However, a significant problem for traditional HAR approaches is the non-independent and identically distributed (non-IID) data caused by real-world factors, such as different user groups, devices and environments. This will greatly limit the performance of HAR tasks when facing unseen data distributions in real-world deployment [50].

In response, a variety of domain adaption (DA) approaches have been proposed to enhance the performance of HAR models when facing unseen data distributions. Existing DA approaches can be divided into two groups. The first group of methods, such as UniHAR [55] and CrossHAR [18], directly leverage the source data (data for training the original model) to adapt the HAR model to new users. They generally use alignment-based methods to keep the features extracted from the new user's data close to that of the source data for the adaptation. However, these methods are impractical because the source data are often confidential and inaccessible to the customers. To address this issue, the second group of methods attempts to achieve model adaptation without source data. These methods, such as OFTTA [53] and SF-Adapter [23], typically use self-supervised learning methods to force the model to give predictions that can be clustered into distinctive groups with the same structure as the ground truth labels. Although these methods are proven to be capable of adapting a HAR model to unseen users, they have poor performance when facing data with larger domain shifts. For example, these methods cannot work when adapting a model to a completely new dataset with a much larger domain discrepancy due to factors such as different user demographics, device specifications, or environmental conditions. This reveals a fundamental contradiction between achieving dataset-level HAR model adaptation and maintaining performance without access to source data.

In this research, we wish to bridge this gap and design a DA system that can work in dataset-level domain adaptation with no source data prerequisites. We observe that previous successful dataset-level domain adaptation methods generally select the target-domain data samples that are close to the source domain to facilitate the adaptation [18, 55]. However, this selection process relies on access to source data, as it is essential for measuring which target-domain samples align with the source domain's data distribution. Therefore, the biggest challenge is to design a method to measure the distance between the target-domain data and the source domain without access to source data. Our rationale for tackling this challenge is that if a data sample is far from the source domain, it will give predictions with low confidence using the original HAR model [2]. Consequently, minor alterations to the raw data, such as applying data augmentation techniques, can significantly impact prediction results, leading to ambiguity. Conversely, if a data sample is close to the source domain, its augmentations and itself will give consistent predictions. Therefore, to select the target-domain data samples that are close to the source domain to facilitate DA, we first perform data augmentation and select the authentic-virtual data pairs that are close to each other in the feature domain. Notably, when computing the distance between these feature pairs, we design a method called DIG to consider both the distance of the features and their correlation with the source domain. We first extract the last-layer output of the mode as a **D**iscriminative **I**nformation of the data sample, process them with the **G**ramian matrix of the model parameters, and then calculate the distance between these two. This DIG measurement not only characterizes the distance between two features but also encodes their relationship with the source domain because the Gramian matrix characterizes the model parameters, which are optimized with the source data. As we will see in Sec. 4.3.2, this design significantly boosts the performance.

The second challenge is that even if we can select a good amount of high-quality target-domain samples that are close to the source domain, it is still hard to fine-tune the HAR model so that it can be generalized to all target-domain data due to the lack of target-domain labels. To resolve this challenge, we design a domain generalization pipeline that updates the HAR model in a way that the features extracted from the selected

Table 1. A summary of existing human activity recognition approaches and our work.

Approach	Without source data	Source model		Target unlabeled data	Non-IID in real world	
		Cloud	Edge		New user	New user group
UniHAR [55]	⊙	●	●	●	●	●
CrossHAR [18]	⊙	⊙	●	●	●	●
OFTTA [53]	●	⊙	●	●	●	⊙
SF-Adapter [23]	●	⊙	●	●	●	⊙
DDLearn [38]	●	⊙	●	●	●	⊙
DI2SDiff [59]	●	⊙	●	●	●	⊙
MobHAR (our work)	●	⊙	●	●	●	●

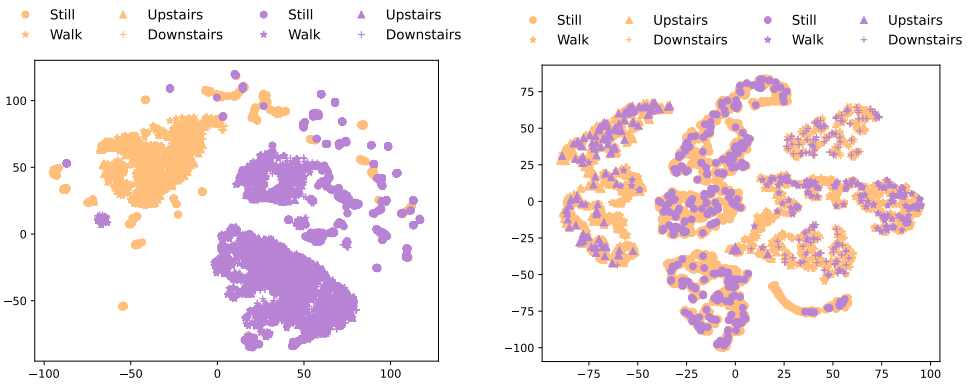
⊙ is not supported by the approach; ● is supported by the approach.

target-domain data and their augmentations are consistent with each other. The rationale behind this design is that we can increase the generalizability of the HAR model by making it invariant to different data augmentation techniques and, further, to all the target domain data. Specifically, we calculate the Wasserstein distance [1] between the distributions of features extracted from two data samples, an authentic data sample and its augmented variant, to characterize their distance [7]. We then use the Nuclear norm $\|\cdot\|_*$ to compute the loss function and use this loss function to enforce feature extraction consistency.

With the above design considerations, we present *MobHAR* to achieve source-free and dataset-level domain adaptation. We evaluate our approach's capabilities on five common HAR datasets. These datasets include different physical activities characterized by the inertial measurement unit (IMU) data on smartphones. These datasets have fundamentally different data collection setups, including completely different user groups, devices, environments, etc. Specifically, a total number of 87 users are included in these five datasets. The experiment results demonstrate that *MobHAR* attains an average accuracy of 76.77% and an F1-score of 67.03% in the single-source scenario (model trained on one dataset and tested on the other four), surpassing state-of-the-art (SOTA) solutions by a minimum of 17.43% in terms of F1-score and 9.77% in terms of accuracy. Additionally, in the multi-source scenario, *MobHAR* achieves an average accuracy of 85.33% and an F1-score of 79.55%, outperforming SOTA solutions by at least 14.8% in terms of F1-score and 8.8% in terms of accuracy. We outline the key contributions of this paper as follows.

- We design *MobHAR*, a source-free domain adaptation framework for HAR models, which is designed to handle large domain discrepancies. To the best of our knowledge, this is the first work to achieve dataset-level HAR domain adaptation without the need for source-domain data.
- We propose a few methods to solve a series of technical challenges, including a DIG method to measure the relationship between target-domain data samples and the source domain, and an unsupervised domain generalization technique to generalize the model from the source domain to the target domain.
- We locally prototype *MobHAR* on the user side and rigorously evaluate its domain adaptation ability with a variety of datasets. Our results show that *MobHAR* exceeds state-of-the-art solutions, achieving at least a 17.43% improvement in F1-score and a 9.77% increase in accuracy[1].

[1]We release the source code of *MobHAR*: https://github.com/xmyun/Phar



(a) Distribution shift

(b) Aligned by *MobHAR*

Fig. 1. Impact of real-world distribution shift. (a) The orange and purple colors denote samples from the Motion and Shoaib datasets, respectively. (b) Distribution shift removed by *MobHAR*.

2 PRELIMINARIES AND RATIONALES

2.1 Source-free HAR

In a practical setting, users download a HAR model from an app store to monitor their activities. As shown in Table 1, we categorize existing HAR approaches [18, 23, 38, 53, 55, 59] into two groups: (i) *Source data-based* method denotes those that are requiring the training (source) data of the downloaded HAR model [18, 55] and (ii) *Source-free* HAR denotes those that solving non-IID do not need source-domain data [23, 38, 53, 59].

Source data-based methods are more effective than source-free approaches. However, source data-based methods require training data for a downloaded HAR model, which is often challenging for consumers to obtain. Source-free HAR approaches do not need source data, requiring high-quality unlabeled new user data. Therefore, there is currently no accurate activity monitor approach that meets the needs of real-world environments.

2.2 Non-IID Data in Real-world Environments

Existing HAR models such as [17, 19, 21, 27, 40, 45, 46, 56, 57] commonly assume that data adheres to the concept of independent and identically distributed (IID) [50]. However, individual differences like age, body type, and behavior deviate from this assumption, preventing the model from being generalized to other domains [53]. Specifically, the performance of HAR models may decline when the data distribution of a new user differs from that of the training datasets. For instance, ASTTL, as noted in [36], achieves only an average accuracy of 66.3% when transferred between datasets, indicating substantial performance degradation due to significant distribution shifts (inter-dataset). Utilizing t-SNE [48] to visualize distribution shifts between the Motion [30] and Shoaib [42] datasets, as shown in Figure 1(a) and detailed in Section 4.1, reveals significant discrepancies in the inertial measurement unit (IMU) data for the same activity type across these datasets. Therefore, there is a significant gap in addressing large data distribution shifts when implementing HAR in practice.

To roughly assess *MobHAR*'s capability in handling distribution shifts, we visualize t-SNE of sample features post-transfer from Shoaib to Motion, as detailed in Section 4.1. As illustrated in Figure 1(b), *MobHAR* effectively manages the distribution shift.

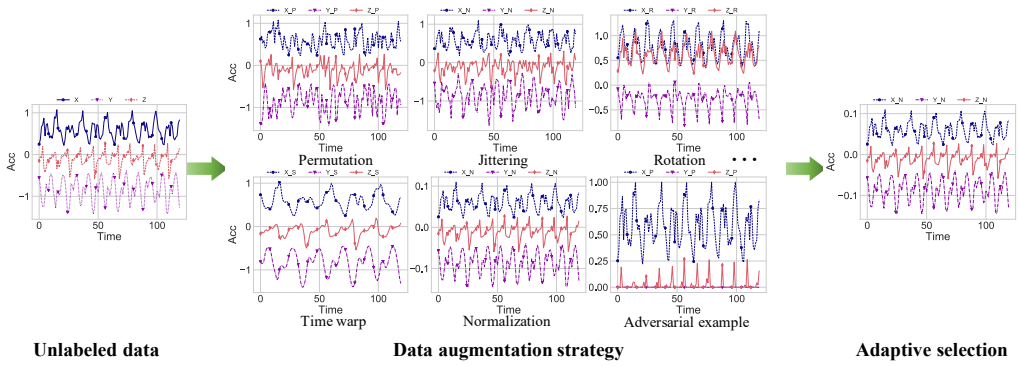


Fig. 2. High-quality virtual data selection based on augmentation strategy.

2.3 Problem Definition

Our HAR framework aims to tailor personalized models locally from pre-trained ones for new, unseen users. As illustrated in Figure 4, the pre-trained model is trained on initial datasets, defined as the source domain $\mathcal{D}_S = \{x_i^s, y_i^s\}_{i=1}^{n_s}$, where $x \in \mathbf{R}^d$ denotes the d-dimensional sensor input and $y \in Y$ signifies the corresponding activity label. New users download pre-trained HAR models, and their local, unlabeled datasets collected from one or more mobile devices are defined as the target domain $\mathcal{D}_T = \{x_i^t\}_{i=1}^{n_t}$. Significant shifts in data distribution between the target $\mathcal{D}_T$ and source $\mathcal{D}_S$ domains are likely due to factors such as *ages, devices, on-body positions, body types, and behavioral habits*. The framework's ultimate objective is to deliver a high-performance activity recognition model that respects user privacy for these new, unseen users.

2.4 Motivation

Test-time adaptation (TTA) [32, 49, 53] can improve a model's generalization by using pseudo labels from unlabeled data. However, there are unavoidable errors between pseudo and real labels, resulting in noise gradients and incorrect model learning. Moreover, noisy data can worsen these errors, potentially disrupting the model and severely degrading its predictive ability. Thus, a natural question emerges: *can we leverage high-quality data to fine-tune the model?* This question is nearly impossible to answer, as evaluating the quality of unlabeled data is quite challenging for different new users.

To tackle this challenge, we introduce the DIG method to quantify the relationship between the target domain and the source domain samples. According to [2], if a data sample is far from the source domain, the original HAR model will produce low-confidence predictions. If a data sample is close to the source domain, both the sample and its augmentations yield consistent predictions. Therefore, as shown in Figure 2, for each unlabeled data, we first augment unlabeled data using six kinds of strategies [6, 35, 40, 46, 47, 51, 55, 56], including normalization, rotation, wrapping, permutation, adversarial examples, and jittering. Then, using DIG, we get the optimal authentic-virtual data pairs from the six kinds of strategies. Finally, we select the top-ten percent of authentic-virtual data pairs in each batch, which are optimized with the source data. Therefore, as shown in Figure 3, we achieve the goal of selecting high-quality unlabeled data by quantifying the relationship between the target domain and the source domain samples.

3 DESIGN OF *MOBHAR*

Figure 4 shows that *MobHAR* comprises two parts: measuring DIG and crafting personalized models.

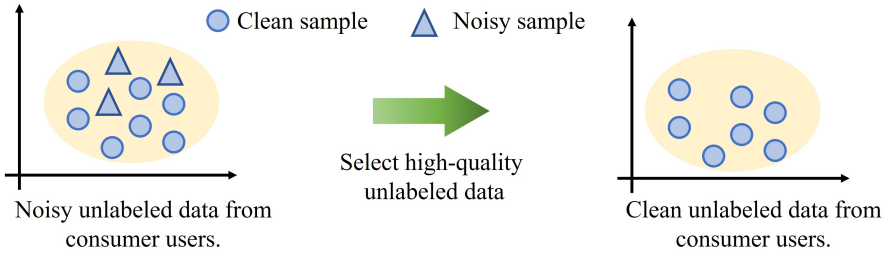


Fig. 3. Select high-quality unlabeled data.

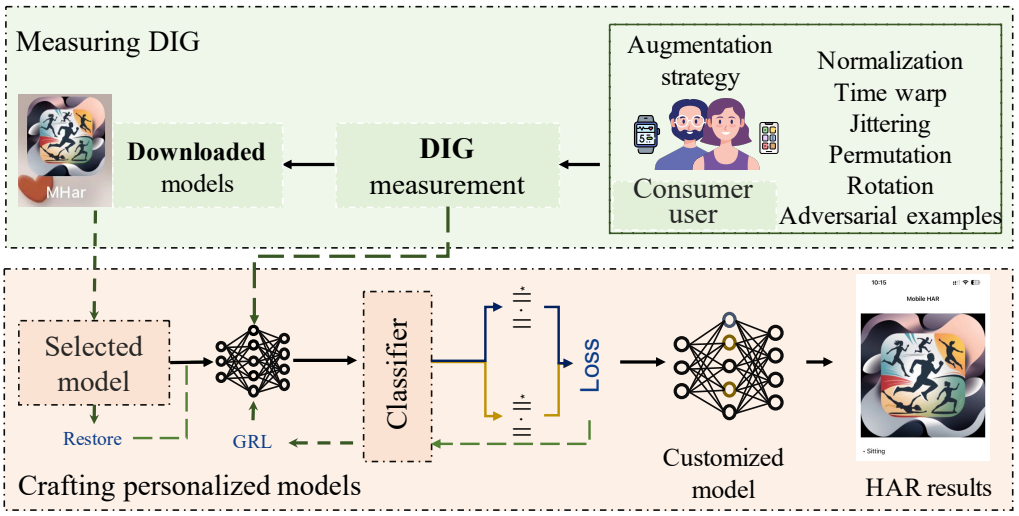


Fig. 4. Architecture of *MobHAR*, which has two parts: measuring DIG and crafting personalized models.

Measuring DIG: In this section, we design the DIG in the context of feature spaces. The DIG matrix quantifies the discrepancy between a new user's data and its augmented variations. By utilizing the DIG, we can assess high-quality data to customize a source-free HAR model. Consequently, the DIG matrix plays a crucial role in selecting appropriate pre-trained models, facilitating source-free knowledge transfer, and ensuring timely stopping to prevent overfitting.

Crafting Personalized Models: In this section, we present a methodology for utilizing DIG to customize a personalized model. Specifically, we design an adversarial mechanism to facilitate source-free knowledge transfer, tailoring a personalized HAR model for new users. First, we employ minimum DIG to select high-quality unlabeled data, leveraging multiple data augmentation strategies. Next, we conclude the fine-tuning of the HAR model at the epoch corresponding to the maximum DIG.

3.1 Maximum Mean Discrepancy Information using Gramian Matrix

Source-free HAR method under non-IID data across diverse user groups from real-world variations requires high-quality pseudo labels from unlabeled data for fine-tuning. In order to ensure pseudo-labels quality as much as possible, it is better to remove the distribution discrepancy between the training data of the consumer-downloaded HAR model and the consumer user's unlabeled data. However, meeting this challenge is difficult due to the

non-IID nature of data across diverse user groups, influenced by real-world variations such as different devices, usage patterns, and environments.

To get high-quality pseudo labels from unlabeled data, we design the maximum mean discrepancy information using gramian matrix of feature spaces to select high-quality samples. Specifically, we first use Gramian information to calculate the inner products of a collection of features extracted from layers of the HAR model. Then, we create gramian representation in the feature space using unlabeled data and its augmentation. Finally, we devise discriminative information from the statistical moments and calculate the median and median absolute deviation. Based on the DIG of feature spaces, we quantify the distance between unlabeled data and its variations. Moreover, a lower DIG indicates higher quality pseudo labels from unlabeled data.

3.1.1 Gram Matrix in HAR Model. Gramian information is commonly employed in style transfer learning [12], where it calculates the inner products of a collection of m-dimensional vectors. These vectors can represent features extracted from layers of the HAR model. Formally, we define a set of m-dimensional layers of a HAR model as:

$$\mathbf{L} := \{l_n | l_n \in \mathbb{R}^m\}_{n=1}^N, \tag{1}$$

where l_n denotes the feature representation of the input sample x at the l-th layer of the HAR model, N denotes the number of channels at layer l, and m signifies the height multiplied by the width of each feature map. Therefore, we define the Gram matrix of the HAR model as: $G_{ij} = \sum_n^m l_{in} \cdot l_{jn}^T$, where G_{ij} represents the Gramian information between each layer in an inner product space of dimensions $N \times N$. The entries of $p-$th order of the Gram matrix is defined as:

$$G_{ij}^p = (l_i^p l_j^{pT})^{1/p}, \tag{2}$$

where p denotes the exponent. Since the matrix G_{ij}^p is symmetric, we form a $N(N+1)/2$-dimensional vector, denoted as G_d^p, where d denotes the dimension of $N(N+1)/2$.

Given the upper limit of an order P, we can derive G_d^p for every order $p \in \{1...P\}$. For the input sample x, we create a new representation $r = [G_d^1, G_d^2, ..., G_d^P] \in \mathbb{R}^{dP}$ by concatenating all the G_d^p. Let x_t denote the data samples of an unseen target user, and x_{ta} denote its augmentation. Consequently, for a given HAR model, we obtain a feature representation as described in Equation 1, along with a concatenated set of Gramian information $R_t = \{r_i, i \in 1, 2, ..., |x_t|\}$. Similarly, we derive $R_{ta} = \{r_i, i \in 1, 2, ..., |x_{ta}|\}$ for the augmentation data samples. Therefore, the relationship between the model and the new user's unlabeled data can be expressed by measuring the distance between R_t and R_{ta}.

3.1.2 Formulating the DIG. Let $x_t^p \sim P_{t,p}$ and $x_{ta}^p \sim P_{ta,p}$ denote the p-th order representation derived from the unseen target user distribution $P_{t,p}$ and the augmented distributions $P_{ta,p}$ of the user, respectively. *MobHAR* calculates representations in feature space without making any assumptions about $P_{t,p}$ and $P_{ta,p}$. Specifically, *MobHAR* derives discriminative information from the statistical moments of $P_{t,p}$ and $P_{ta,p}$ [4]. Referring to [29], we can define the DIG as a view of statistical moments:

$$\begin{aligned} DIG\,(P_t, P_{ta}) &= E_p[E(G_{x_t^p}) - E(G_{x_{ta}^p})] \\ &= E_p[\mu_1^p \mu_1^{pT} - \mu_2^p \mu_2^{pT} + \sigma_1^p - \sigma_2^p], \end{aligned} \tag{3}$$

where μ_1^p and σ_1^p denote the mean vector and the covariance matrix of x_t^p, μ_2^p and σ_2^p denote the mean vector and the covariance matrix of x_{ta}^p, P_t denotes a collection of the representations from the unseen target user distribution with elements of different powers, and P_{ta} denotes that of the augmented distributions of these users. We provide a detailed derivation in the Appendix 8.1.

ALGORITHM 1: Customizing a HAR model for a new user.

Input: The HAR models downloaded from the App store.
Phase 1: A new user downloads the HAR models.
Phase 2: *MobHAR* selects the best model parameters for the new user using DIG (Equation 5).
while $i \leftarrow 1$ *to* 30 **do**
    Conduct data augmentation: normalization, rotation, wrapping, permutation, adversarial examples, and jittering.
    Select the optimal data augmentation strategy using DIG.
    for *batch in epoch i* **do**
        Compute the distribution shift loss by Equation 14.
        Adapt the HAR model using the distribution shift.
        Stochastically restore model by Equation 7.
    end
    Calculate the DIG (Equation 5) sum.
end
Output the best HAR model by DIG.

3.1.3 Calculating the DIG. Utilizing the Gramian information $R_t = \{r_i, i \in 1, 2, ..., |x_t|\}$ from the unseen target user's data x_t, we calculate the median and Median Absolute Deviation (MAD) [25] as $\tilde{r}_j = Me(r_{ij}, i \in |x_t|)$ and $MAD_j = Me(|r_{ij} - \tilde{r}_j|, i \in |x_t|)$. We calculate the distance of feature representation $\hat{r}_j$ from the candidate point $\tilde{r}$ as:

$$\delta_j(\hat{r}_j) = \begin{cases} 0 & if \quad b_l \leq \hat{r}_j \leq b_u, \\ \frac{b_l - \hat{r}_j}{b_l}, & if \quad \hat{r}_j \leq b_l, \\ \frac{\hat{r}_j - b_u}{b_u}, & if \quad b_u \leq \hat{r}_j, \end{cases} \tag{4}$$

where $b_l = \tilde{r}_j - 10 \times MAD_j$ and $b_u = \tilde{r}_j + 10 \times MAD_j$ denote the lower and upper bounds, respectively. Therefore, the distance of the candidate point $\tilde{r}$ can be calculated by summing all $\hat{r}_j$ as: $\delta = \frac{1}{dP} \sum_{j=1}^{dP} \delta_j$.

As the distributions of the unseen target user and the generalized distributions may not exhibit significant differences, it is essential to employ a distribution measurement method without assumptions about the distributions being tested [41] when evaluating the distance between R_t and R_{ta}. Drawing from advancements in distance measurement methodologies [9, 16, 31], we can reformulate Equation 9 using RMMD as:

$$DIG(P_t, P_{ta}) = \|\mu_{P_t} - \mu_{P_{ta}}\|^2_{\mathcal{H}} - \lambda_{P_t}\|\mu_{P_t}\|^2_{\mathcal{H}} - \lambda_{P_{ta}}\|\mu_{P_{ta}}\|^2_{\mathcal{H}}, \tag{5}$$

where P_t and P_{ta} denote the distribution of representations obtained from the deviation measurement for the two subgroups (unseen target user and its augmentation).

3.2 Crafting Personalized Models

After evaluating the disparity between a general model and personalized models, we can leverage this information to customize personalized models via source-free knowledge transfer. As shown in Figure 5, we design an adversarial mechanism to facilitate source-free knowledge transfer. At the minimum stage, we choose high-quality unlabeled data to fine-tune the HAR model based on minimum DIG. At the maximum stage, we assess whether the high-quality data is sufficient to conclude the fine-tuning using maximum DIG. The minimum stage contains *Measuring Disparity to Select High-Quality Data* and *Eliminating Disparity*. The maximum stage contains *Preventing Overfitting*.

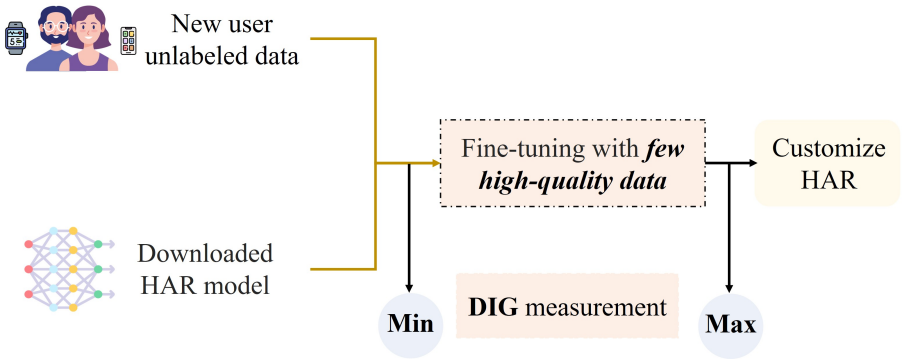


Fig. 5. Adversarial mechanism to facilitate source-free knowledge transfer. 'Min': choose high-quality unlabeled data to fine-tune the HAR model based on minimum DIG. 'Max': assess whether the high-quality data is sufficient to conclude the fine-tuning using maximum DIG.

3.2.1 Measuring Disparity to Select High-Quality Data. Since consumer users can only access unlabeled data from their own devices and a downloaded HAR model, assessing the quality of this unlabeled data is significantly challenging. Although some test-time adaptation methods [20, 32, 49, 53] can generate pseudo labels using both unlabeled data and a downloaded HAR model, there are inherent errors between the pseudo and actual labels. Additionally, noisy data can exacerbate these errors, making it less practical to rely on pseudo labels for evaluating unlabeled data.

To solve this problem, we utilize the DIG to quantify the distance between unlabeled data and its augmentation variations, ensuring *MobHAR* can effectively select high-quality samples. This is essential because the inclusion of unfavorable samples can adversely affect the model's performance [43]. To achieve this, we first perform data augmentation on each sample, employing techniques such as normalization, wrapping, rotation, adversarial examples, permutation, and jittering. The reason why we utilize multiple data augmentation strategies is that augmentation operations might alter key features of the data. Moreover, since consumer users lack the training data of a downloaded HAR model, irrelevant variations from augmentation operations on unlabeled data can increase distribution discrepancies in non-IID environments. Subsequently, we calculate the distance between each augmented sample and the original sample using DIG. Finally, we obtain the new user's data and its corresponding variants according to the smallest DIG values.

To conduct source-free knowledge transfer based on the new user's data and its corresponding variants, we resort to the predicted discriminative information from the HAR model. Specifically, we calculate the Wasserstein distance [1] between the distributions of two features, P_t and P_{ta}. Due to the complexity of the distribution of a new user's data, the Wasserstein distance often mistakenly emphasizes domain-level feature confusion, which consequently undermines the user's activity information. Based on [22], the main diagonal elements in the self-correlation matrix of the HAR model's prediction denote the intra-class correlation I_i, while the off-diagonal elements denote the inter-class confusion I_o. For an unseen target user, if the distribution gap between the new user and pretrained model's training users is large, then the HAR model's prediction generally yields a small I_i and large I_o. However, if the distribution gap is small, the HAR model's prediction generally yields a large I_i and small I_o. Thus, I_i - I_o can be used to denote the discrepancy while eliminating the distribution shift.

3.2.2 Eliminating Disparity. To reduce the disparity between a general model and personalized models, we develop a structure based on the pre-trained HAR model f and a softmax layer as the classifier S. Since the discrepancy between I_i and I_o, as measured by the Frobenius-norm Wasserstein distance, may overlook the classes

with small samples [8], we use the nuclear norm $\|\cdot\|_*$ to enhance prediction diversity [7, 8]. To mitigate the need for repetitive alternating updates in the HAR model, we employ a gradient reverse layer (GRL) [11], which facilitates updates within a single backpropagation step. As a result, *MobHAR* conducts source-free knowledge transfer as follows:

$$\mathcal{L}_{\theta_f} = min\{\frac{1}{N_u}\sum_{i=1}^{N_u}\left(\left\|S(f(x_i^t))\right\|_* - \left\|S(f(x_i^{ta}))\right\|_*\right)\}, \tag{6}$$

where N_u denotes the number of samples from the unseen target user, x^t denotes the sample from the unseen target user t, x^{ta} denotes the augmentation data samples of t. We provide a detailed derivation in the Appendix 8.2.

3.2.3 Preventing Overfitting. The HAR model may face significant challenges in recovering its performance after encountering challenging samples during adaptation, which can eventually lead to catastrophic forgetting. Therefore, a source-free HAR approach requires an adaptive stop fine-tuning method, improving adaptation efficiency and reducing unnecessary computational costs. To address this problem, we leverage the DIG to quantify the stop time with the largest DIG. Moreover, we utilize a stochastic restoration method that restores weight from the pre-trained model to support long-time adaptation.

Firstly, we utilize a stochastic restoration method [52] that restores weight from the pre-trained model:

$$W_{t+1} = M \odot W_0 + (1 - M) \odot W_{t+1}, \tag{7}$$

where $M \sim \text{Bernoulli}(p)$, $\odot$ denotes element-wise multiplication, W_0 denotes the source weight, W_{t+1} denotes the trainable parameters, and M is a mask tensor. To maximize the model's utilization of high-quality unlabeled data from new users, we stop fine-tuning at the maximum value after removing a significant outlier from the DIG distance. As shown in Algorithm 1, combining the discrepancy information by Gramian with adapting source model leads to *MobHAR* framework.

4 EVALUATION

To evaluate whether *MobHAR* can customize personalized models on mobile devices while preserving user privacy, we tested it on non-independent and identically distributed (non-IID) data distributions. As shown in Table 2, we summarize the factors contributing to the distribution shift based on five diverse datasets: HHAR [44], UCI [39], Motion [30], Shoaib [42], and USC [60]. These factors are categorized into four user characteristics: *position, location, age, and devices*. The term position refers to the placement of the device on the body, while location indicates that participants exhibit various body types and behavioral patterns. This diversity allows us to thoroughly evaluate the adaptability of *MobHAR* across different user activities and settings.

4.1 Experiment Information

4.1.1 Datasets. **HHAR** [44]. This dataset includes accelerometer and gyroscope measurements from 9 users performing 6 different activities (sitting, standing, walking, upstairs, downstairs, biking) with 4 types of mobile phones (3 Samsung Galaxy models, 1 LG model), worn around their waists.

UCI [39]. Comprising raw accelerometer and gyroscope data, this dataset was collected from 30 Italian volunteers, aged 19 to 48, capturing six basic activities (standing, sitting, lying, walking, downstairs, upstairs) at 50 Hz with a Samsung Galaxy S II on the waist.

Motion [30]. Utilizing an iPhone 6, this dataset collected time-series accelerometer and gyroscope data at 50 Hz from 24 UK participants engaged in six activities (sitting, standing, walking, upstairs, downstairs, jogging), with the device in front pockets.

Table 2. Summary of experiment setups.

Dataset	Factors resulting in the distribution shift				Time per user (minute)
	Position	Location	Age (year)	Device	
HHAR [44]	Waist	Denmark	25∼30	LG Nexus 4, Galaxy -S3/ -S3 mini/ -S plus	25
UCI [39]	Waist	Italy	19∼48	Galaxy S II	4
Motion [30]	Front pocket	UK	18∼46	iPhone 6S	18.5∼29
Shoaib [42]	Right/left pocket, belt, right arm, right wrist	Netherlands	25∼30	Samsung Galaxy S II	15∼20
USC [60]	Front right hip	USA (California)	21∼49	MotionNode	240

Shoaib [42]. This dataset features 10 male participants performing activities like sitting, standing, walking, upstairs, downstairs, jogging, and biking, with five Samsung Galaxy SII phones placed at various body locations (right pocket, left pocket, belt, upper arm, wrist), recording at 50 Hz.

USC [60]. Capturing data from 12 activities (walking, running, jumping, sitting, standing, sleeping, elevator), this dataset involves 14 diverse participants, with accelerometer and gyroscope readings taken at 100 Hz.

To demonstrate the versatility of *MobHAR*, we choose four prevalent activities (still, walk, upstairs, downstairs) from five diverse datasets, echoing UniHAR's approach [55]. This unified dataset showcases a broad spectrum of factors—ages, devices, body positions, types, habits—and adopts unified labels for these activities, revealing varying distributions across datasets.

4.1.2 Metrics. We evaluate the performance of *MobHAR* with average F1-score and accuracy of users in the target domain, which are defined as $\overline{F1} = \sum F1_i$, $\overline{Acc} = \sum Acc_i$, s.t. $i \in \mathcal{D}_t$, where $F1_i$ and Acc_i are the activity classification F1-score and accuracy of the i-th user, respectively. Due to the fact that the F1-score, being the harmonic mean of precision and recall, offers a more comprehensive assessment of the model's performance, particularly effective in dealing with imbalanced class distributions, we choose to utilize the F1-score as the primary metric for evaluation.

4.1.3 Baseline Methods. To assess the performance of *MobHAR* in customizing personalized models for new users, we benchmark it against five pertinent baseline methods:

OFTTA [53] uses conventional and test-time batch normalization (TBN) to extract stable features across domains, with TBN effectiveness diminishing progressively at deeper network layers.

TAST [20] leverages multiple adaptation modules, each randomly initialized, to harness useful information from nearest neighbors for classifying test data experiencing domain shifts.

TENT [49] fine-tunes pre-trained models by dynamically adjusting normalization statistics and channel-wise affine transformations based on test batch entropy minimization.

SAR [32] introduces a dual approach to enhance test-time adaptation: first by filtering out noisy samples with high gradient magnitudes, and then by nudging the model towards flat minima to ensure stability against residual noise.

UniHAR [55] embeds a range of data augmentation strategies within a self-supervised learning framework to tackle data diversity issues, necessitating that new users upload their features or raw data to a cloud server.

Table 3. Performance comparison: Multi-source to single-target. (The two numbers in each cell are accuracy and F1-score.)

Source data	Method	HHAR		UCI		Motion		Shoaib		Average
		Acc, F1		Acc, F1		Acc, F1		Acc, F1		Acc, F1
With	UniHAR [55]	73.8	64.5	89.0	81.9	79.2	68.3	77.7	72.3	79.93, 71.75
	CrossHAR [55]	76.19	-	88.68	-	78.26	-	73.67	-	79.2, -
Without	OFTTA [53]	62.9	50.2	86.4	76.6	79.8	57.9	69.3	61.0	74.6, 61.43
	TENT [49]	64.0	51.6	89.2	81.8	82.1	61.7	70.8	63.9	76.53, 64.75
	SAR [32]	63.6	51.8	86.1	77.1	**82.8**	62.6	65.9	56.1	74.6, 61.9
	TAST [20]	60.4	48.6	87.8	79.4	81.6	61.8	69.8	60.3	74.9, 64.53
	MobHAR (**Ours**)	**85.4**	**79.5**	**96.5**	**94.2**	81.2	**71.0**	**78.2**	**73.5**	**85.33, 79.55**

CrossHAR [18] initially builds a pre-trained model using a self-supervised approach with augmented data, and subsequently fine-tunes the model with partially labeled data from the source-domain dataset.

To evaluate *MobHAR*'s performance under different discrepancies between general and personalized models, we conduct cross-dataset evaluations using both single-source and multi-source pre-trained models. In the single-source setup, characterized by significant distributional disparity, *MobHAR* transfers models from the HHAR dataset to other unlabeled datasets including UCI, Motion, Shoaib, and USC. Conversely, in the multi-source scenario, designed to reduce the gap between general and personalized models, *MobHAR* utilizes models trained on a combination of UCI, Motion, and Shoaib datasets adapted to the HHAR dataset. For both strategies, 10% of users in the source domain are selected as a validation set based on training loss, helping to optimize the choice of pre-trained models. Users in the target domain, possessing only unlabeled IMU data, download pre-trained models to customize their own HAR model.

4.2 *MobHAR* for Human Activity Recognition

4.2.1 Cross-dataset Evaluation of Multi-source Single-target. Table 3 compares the performance of *MobHAR* and other baseline models in the multi-source setup. The two numbers in each cell denote average accuracy and F1-score, respectively. We divide our baselines into two groups based on the presence of source domain data: source data-based and source-free. Source data-based methods contain UniHAR[55] and CrossHAR [18]. Source-free methods contain OFTTA [53], TAST [20], SAR [32], and TENT [49]. In the multi-source to single-target scenario, the pre-trained model is built on multiple datasets, omitting both the target dataset and USC. This method preserves a setting similar to CrossHAR, facilitating a fair comparison. As shown in Table 3, the source data-based methods have higher performance than SOTA source-free methods. This is because source data-based methods include source domain data, which helps facilitate the alignment between source and target domain data when addressing the distribution shift. Comparing with source data-based methods, *MobHAR* outperforms the best of baselines by 5.4% in terms of accuracy and 7.8%. For the source-free methods, the elimination of distribution shift can only rely on the unlabeled target domain users. However, the source-free method is more practical, as source domain data is often unavailable to end users. *MobHAR* achieves 85.33% average accuracy and 79.55% F1-score, outperforming the best of the four baselines. Although SAR delivers the best performance in the case of transferring from *HHAR, UCI, and SHOAIB* to MOTION, it is shown to be sensitive to the source domain datasets and performs much lower in other cases. For example, SAR achieves only 63.6% in terms of accuracy

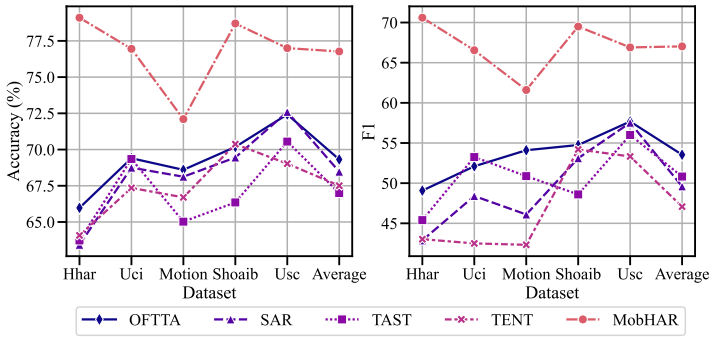


Fig. 6. Overall performance of each method on different datasets comparison: Single-source to single-target. (The *'Average' is the average of a specific method across four datasets.*)

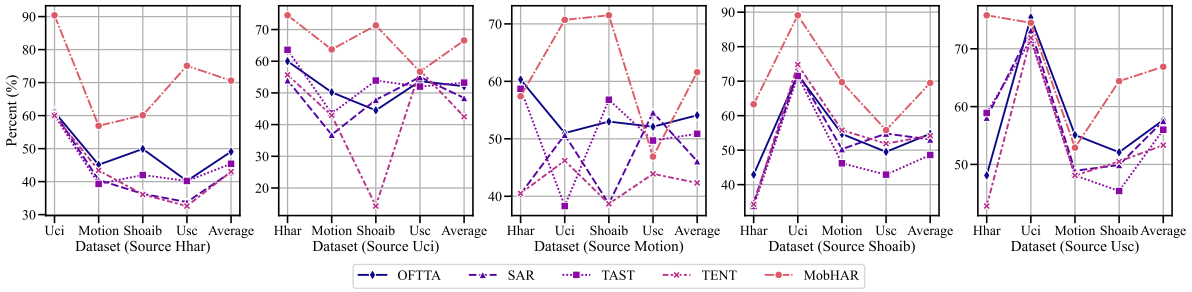


Fig. 7. Performance in F1-score comparison: Single-source to single-target.

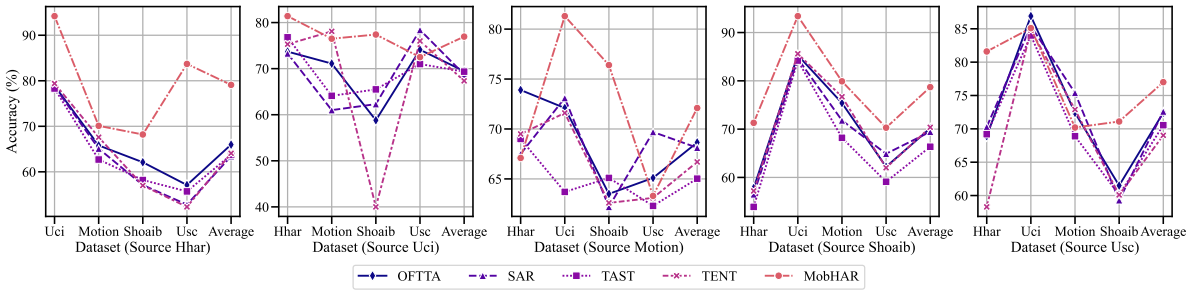


Fig. 8. Performance in accuracy comparison: Single-source to single-target.

when transferring from *UCI, MOTION, and SHOAIB* to HHAR. Comparing with source-free methods, *MobHAR* outperforms the best of baseline by 14.76 % in terms of F1-score and 8.8% in terms of accuracy.

In summary, *MobHAR* can customize personalized models in a source-free manner.

4.2.2 Cross-dataset Evaluation of Single-source Single-target. Figure 6 compares the performance of *MobHAR* and other baseline models (Sec.4.1) in the single-source setup. We leverage four source-free methods as our baseline, *i.e.*, OFTTA [53], TAST [20], SAR [32], and TENT [49]. As shown in Figure 6, OFTTA has higher performance

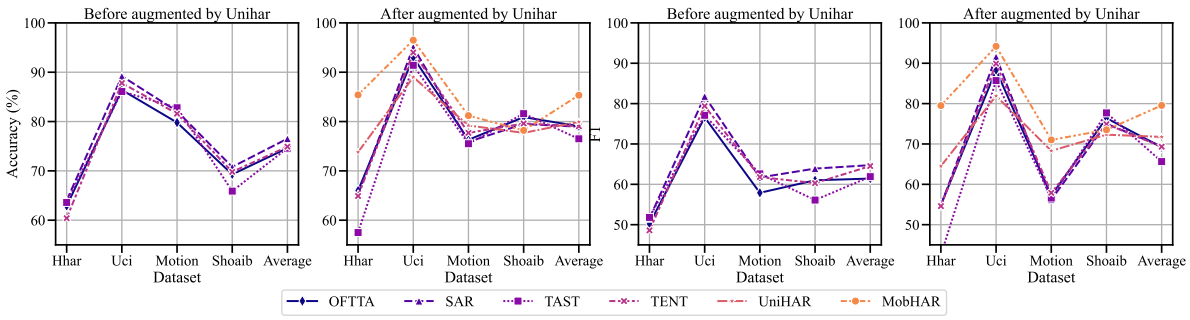


Fig. 9. Impact of data augmentation.

than other source-free methods, *i.e.*, TAST, SAR, and TENT. This is because the OFTTA is specifically designed for IMU data while other source-free methods are designed for image data. In contrast, *MobHAR* achieves 76.77% average accuracy and 67.03% F1-score, outperforming the best of four baselines by at least 17.43% in terms of F1-score and 9.77% in terms of accuracy. Figures 8 and 7 present the performance of *MobHAR* and other baseline models (see Sec. 4.1) in the single-source setup, detailing accuracy and F1-score, respectively. As shown in Figure 8 and 7, we can see that when the source domain is HHAR, *MobHAR* can get the best performance than other source domains. This is because *MobHAR* leverages the high quality of target unlabeled data to mitigate the distribution shift caused by unseen new user data. Consequently, when the pre-trained model is trained on cleaner data, it becomes more effective at customizing a source-free HAR model. Although SAR delivers the best performance in the case of transferring from MOTION to USC, it is shown to be sensitive to other cases and performs much lower in them. Similarly, OFTTA can also achieve the best performance in the case of transferring from USC to UCI. In summary, the results demonstrate the remarkable performance of *MobHAR* in mitigating significant domain shifts, enabling the customization of personalized models in a source-free way.

4.3 Sensitivity, Model Size and Latency, and System Overhead

4.3.1 Sensitivity: Impact of Data Augmentation. Since *MobHAR* utilizes different data augmentation strategies to assess the quality of unlabeled data. Thus, in this section, we assess how various data augmentation methods affect the customization of personalized models under four distinct target domains: HHAR, UCI, MOTION, and SHOAIB. In addition, the corresponding source pre-trained model draws from different datasets without the target dataset and USC, which is because of fair comparison with the baselines. Specifically, we evaluate the impact of data augmentation by comparing *MobHAR* with five baseline methods, *i.e.*, OFTTA, SAR, TAST, TENT, UniHAR.

Figure 9 showcases the impact of different data augmentation strategies on *MobHAR* 's performance. Results clearly demonstrate that data augmentations contribute differently to the efficacy of *MobHAR*. Comparing the performance between after and before data augmentation of UniHAR, there is a slight improvement for SOTA test-time adaptation methods, *i.e.*, OFTTA, SAR, TAST, and TENT. As shown in Figure 9, the UniHAR has a higher performance than OFTTA. This is because UniHAR can get the source domain data to improve the HAR model's performance. Without acquiring the source domain data, *MobHAR* gets a high performance under source-free scenarios with only high-quality data from them.

4.3.2 Sensitivity: Impact of Selecting High-Quality Data. As described in Section 2.4, *MobHAR* leverage DIG to quantify the relationship between the target domain and the source domain samples, achieving the goal of

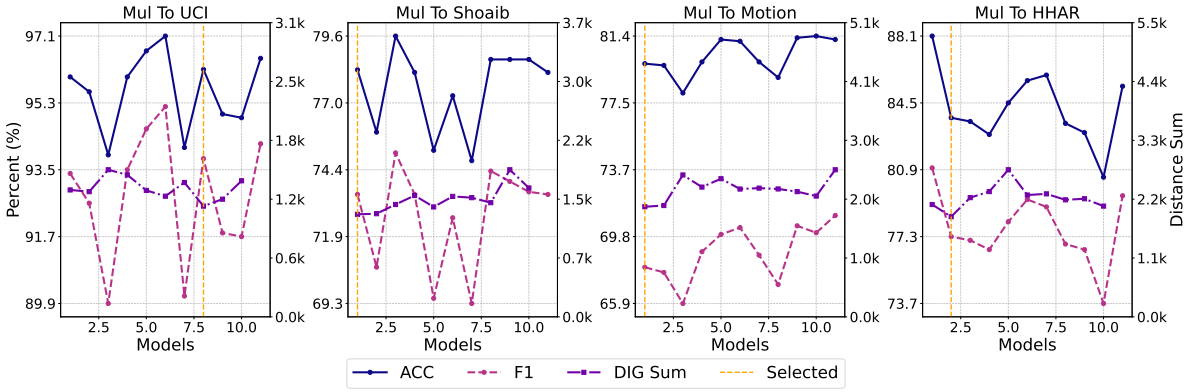


Fig. 10. Impact of selecting the appropriate pre-trained model.

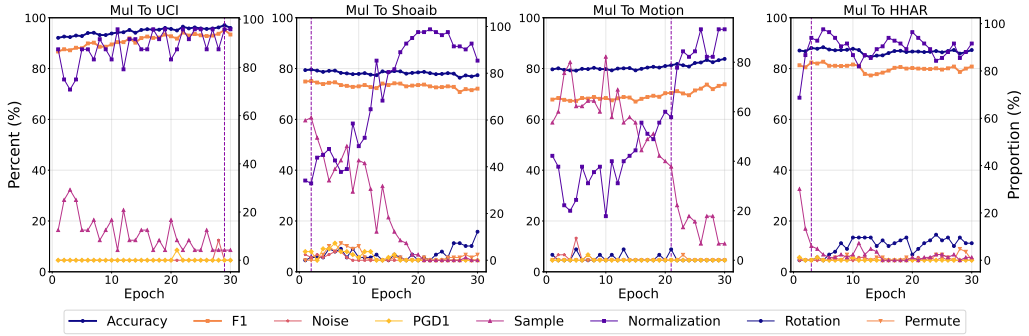


Fig. 11. The data augmentation strategy ratio at different fine-tuning epochs.

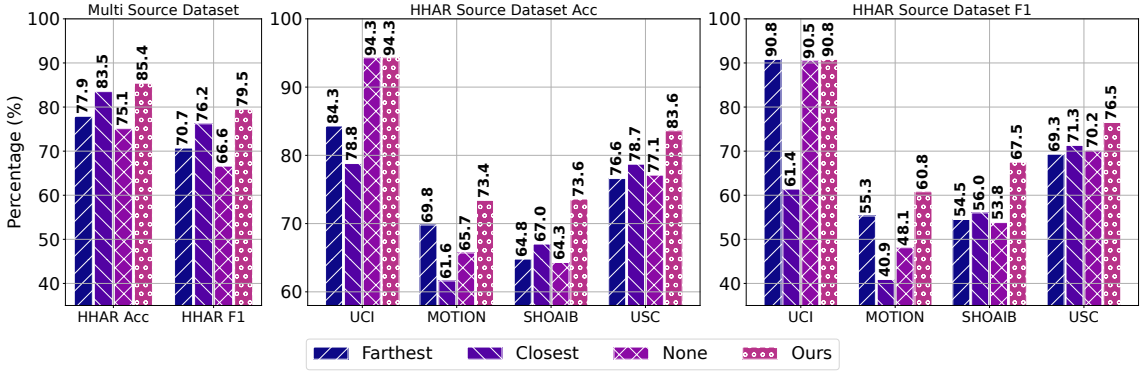


Fig. 12. Impact of data quality.

selecting high-quality unlabeled data. Thus, in this section, we will show the impact of selecting high-quality unlabeled data into two parts:

(i) selecting optimal authentic-virtual data pairs,
(ii) selecting the top ten percent of authentic-virtual data pairs in each batch.

To evaluate (i), we leverage multiple-source datasets to a single-target dataset. For instance, as shown in the 'Mul To UCI' subfigure of Figure 11, the source domain is a combination of the Shoaib, Motion, and HHAR datasets. The target domain is the UCI dataset. Specifically, *MobHAR* first augments unlabeled data and then selects the most appropriate variants by assessing their minimum DIG values. Therefore, we demonstrate both the ratio of each augmentation strategy and the performance of *MobHAR* across different fine-tuning epochs. As depicted in Figure 11, *MobHAR* can select different augmentation methods to generate variations, which vary across fine-tuning epochs. Moreover, as we can see, the normalization method has the highest ratio among these augmentation strategies, followed by sample warping. Additionally, the ratios of other methods are similar.

To assess point (ii), whether high-quality unlabeled data identified by *MobHAR* enhances the generalization of the HAR model, we evaluate various levels of data quality within non-IID settings of HAR, including single-source to single-target datasets and multiple-source to single-target datasets. The data for the new user is categorized into four types: fully optimal authentic-virtual data (Closest), fully suboptimal authentic-virtual data (Farthest), minimal authentic-virtual data (None), and partial authentic-virtual data (Ours). As shown in Figure 12, high-quality data within non-IID settings of HAR demonstrates the highest performance in terms of both accuracy and F1-score. This means that *MobHAR* can effectively evaluate the quality of unlabeled data for various new users.

4.3.3 Sensitivity: Impact of Selecting the Suitable Pre-trained Model. This section assesses the impact of choosing the right pre-trained models. Pre-trained models are typically halted at the minimum validation loss point [55]. Nonetheless, there is a prevailing view that extended training could undermine generalization even though it improves performance on a specific validation dataset [58]. A generic pre-trained model may not fit every new individual user's needs. To assess the impact of pre-trained models, we generate customized models from those exhibiting varying validation loss levels. Recognizing that various model parameters yield different results for unseen users, we maintain a selection of multiple pre-trained models to enhance the generalization ability of our HAR system. *MobHAR* employs the DIG to identify the optimal model parameters from a pool of candidates tailored for each new user. Specifically, *MobHAR* adopts the DIG method to identify the appropriate pre-trained model based on the minimum distance. Figure 10 illustrates how the performance of customized personalized models can differ based on the choice of pre-trained models. As shown in Figure 10, *MobHAR* effectively identifies and selects a good-fitting pre-trained model, leading to significant performance improvements. For instance, as depicted in Figure 10, the accuracy gap among different pre-trained models can be larger than 7.5% and 7.2% with respect to F1-score. The experiments demonstrate that DIG in *MobHAR* serves as an effective metric for quantifying the compatibility between a HAR model and an unseen user's data.

4.3.4 Sensitivity: Impact of Model Architecture. This section explores the impact of model architecture, *i.e.*, DCNN, OFTTA, and MobHAR. As depicted in Figure 13, we evaluate different model architectures from multi-source to single-target domain. Compared to raw DCNN performance as reported in [55], *MobHAR* improves the accuracy from 59.4% to 75.8%, F1-score from 43.8% to 65.875%. It is clear that DCNN has the lowest improvement, and the performance is near in average of OFTTA and LIMU-BERT in accuracy. Moreover, in terms of F1-score, LIMU-BERT is slightly better than OFTTA. Therefore, we select LIMU-BERT as our basic model architecture.

4.3.5 Sensitivity: Impact of Stopping to Prevent Overfitting. This section explores the efficacy of implementing timely stopping strategies to curb overfitting, a phenomenon where the HAR model overlearns from unlabeled data. As depicted in Figure 14, prolonged adaptation leads to a decline in model performance. To counteract this, we implement two approaches: stochastic parameter restoration from the initial pre-trained model and strategic, timely stopping, further explained in Sec. 3.2.3. The essence of this strategy revolves around pinpointing the precise moment to cease the adaptation process. We leverage the DIG to gauge the correlation between the

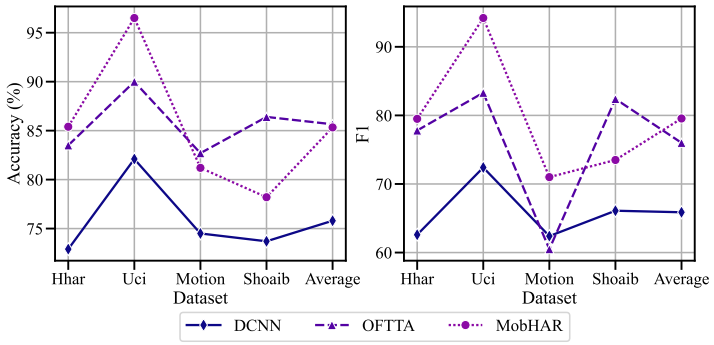


Fig. 13. Performance of each method on different mode architectures.

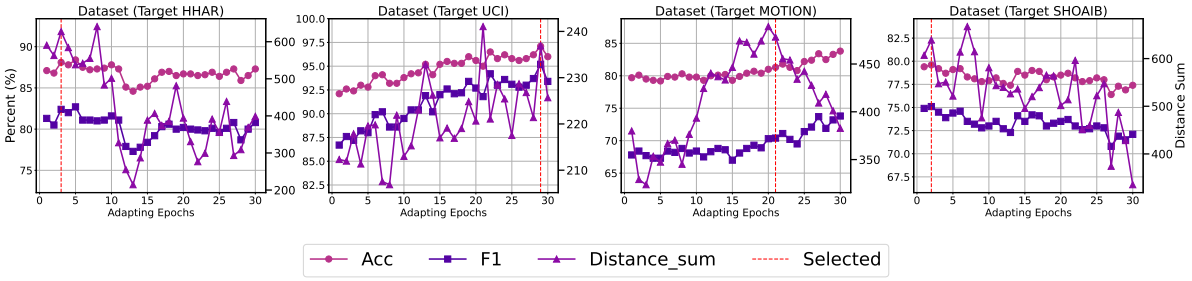


Fig. 14. Adaptivly stop based on DIG.

Table 4. System overhead (**Iphone14-Pro**, CPU, T4 GPU).

Model		Pre-trained	Adapt	Infer	Iphone14-Pro
Latency	User	/	9.05 s	36.6 ms	**13.75ms**
	Server (T4)	157 ms	/	/	/
Size		311KB × 11	311KB	311KB	311KB
New user	CPU	/	41.27 %	30.29 %	**11.6 %**
	Memory	/	736.82 MB	333.58 MB	**47.5MB**

adapted model and the unlabeled data effectively. Figure 14 confirms that *MobHAR* successfully identifies the optimal stopping point across various scenarios listed in Table 3.

4.3.6 System Overhead. To evaluate system overhead, we utilize a smartphone, the **Iphone14-Pro**, and a laptop powered by a *12th Gen Intel(R) Core(TM) i7-1260P 2.10 GHz* CPU.

Latency. Table 4 details the latency experienced by *MobHAR* on the user's side. The adaptation phase for the pre-trained model, sourced from a server, requires about 9.05 seconds. In contrast, performing inference on a batch of HAR model samples is notably quicker, at just 36.6 milliseconds. Notably, this pre-trained model

Table 5. Efficiency comparison (T4 GPU).

Model	Para.	FLOP	Size	Time	
				Train	Infer
DCNN	16.59 K	334.72 K	73 KB	4.1 ms	0.7 ms
LIMU	60.74 K	3.16 M	247 KB	17.3, 18.7 ms	5.2 ms
UniHAR	16.4 K	272.16 K	75 KB	8.5, 9.3 ms	2.6 ms
OFTTA	615.7 K	30.13 M	311 KB	23.9, 2.6ms	2.4 ms
TENT	562.18 K	60.64 M	311 KB	23.9, 29.5 ms	15.7 ms
SAR	562.18 K	60.64 M	311 KB	23.9, 49.1 ms	11.6 ms
TAST	617.49 K	30.48 M	311 KB	23.9, 1.5 ms	1.4 ms
MobHAR	47.65 K	3.43 M	255 KB	157, 2473 ms	5.1 ms

undergoes a training session for a single batch on an NVIDIA T4 Tensor Core GPU, clocking a training duration of 157 milliseconds.

Computation. Computational tasks are executed entirely on the user's device. Initially, the user downloads pre-trained models, with each model sized at 311 KB, cumulatively amounting to 3.34 MB for eleven models. Subsequently, the user undertakes the local customization of a HAR model. This process, encompassing both adaptation and inference, imposes a CPU load of approximately 41.27% and utilizes around 736.82 MB of RAM. Following this, ongoing inferences by the tailored HAR model demand about 30.29% CPU usage, while memory consumption drops to roughly 333.58 MB.

4.3.7 Model Size and Latency. Table 5 offers a detailed comparison of *MobHAR* against baseline models, evaluating parameters such as the number of parameters, model size, floating-point operations per second (FLOPs), training time, and inference time. Training time is defined as the duration required by the server to train a mini-batch of 64 samples, and inference time is the time taken to process a single IMU sample on the NVIDIA T4 Tensor Core GPU. Notably, *MobHAR* 's training time is linked to the pre-training done by the server.

Illustrated in Table 5, *MobHAR* demonstrates a higher parameter count compared to server-based solutions like DCNN and UniHAR. However, against user-side models like OFTTA, TENT, SAR, and TAST, *MobHAR* stands out with the smallest footprint—47.65K parameters, 3.43M FLOPs, and a model size of 255K. Despite *MobHAR* requiring more time to tailor a HAR model for a specific user, it impressively processes inferences in just 5.1ms, thanks to an architecture that mirrors those detailed in [55, 56].

5 RELATED WORK

5.1 Human Activity Recognition

The human activity recognition [14, 15, 27, 57] has significant application prospects in various fields, such as healthcare [54], smart home [3], and spots training [33]. Current HAR models [21, 45] assume that data conforms to the concept of being independent and identically distributed [50]. However, individual differences (such as age, body type, and behavioral habits) lead to data from different individuals not meeting the IID assumption, also known as domain shift, significantly undermining the generalizability of the HAR model. We categorize the current solution for improving the HAR model's generalizability into *solving IID with source domain data* and *solving IID without source domain data.*

Table 6. Impact of target users wearing position ("→" denotes from 'source' to 'target' domain).

Scene1: front right hip to waist				*Scene2*: waist to waist			
USC→HHAR (20 min)		USC→UCI (4 min)		UCI→HHAR		HHAR→UCI	
F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score	Accuracy
75.8%	81.6%	74.5%	85.1%	74.5%	81.4%	90.4%	94.2%

Scene3: front pocket to right/left pocket, right arm/wrist, leg, belt		*Scene4*: front pocket, waist-HHAR, waist-UCI to right/left pocket, right arm/wrist, leg, belt	
Motion→Shoaib		HHAR, UCI, Motion →Shoaib	
F1-score	Accuracy	F1-score	Accuracy
71.5%	76.4%	73.5%	78.2%

5.2 Solving Non-IID with Source Domain Data

To improve the HAR model's generalizability, a straightforward idea is transferring source-domain knowledge to the target domain in the model training phase. Domain generalization (DG) involves learning a generalizable model on the source domains and directly applying it to new users [28, 34, 37]. GILE [34] eliminates domain-specific representations using DG to enhance the performance of the HAR model. AFFAR [37] combines domain-specific and domain-invariant representations to enhance the generalization of the HAR model.

To enhance the practicality of eliminating domain shift, domain adaptation (DA) incorporates target domain data during the training phase [6, 24]. HDCNN [24] dynamically adjusts weights from various distributions. Chang et al. [6] propose aligning the features of both the source and target domains to enhance the robustness against diverse wearables. UniHAR [55] combines self-supervised and supervised training to address heterogeneity in IMU data. Self-supervised training utilizes both source and target domains on the server side, while supervised training focuses on the source domain. CrossHAR [18] solves the IID using partial labeled data from the source-domain dataset.

However, acquiring data from the target domain can be challenging during the training phase due to the complexities of downstream tasks and privacy concerns. As mentioned in [10, 61], if the HAR model does not utilize target data, it may be overly optimistic to expect the model to generalize effectively to an unseen distribution.

5.3 Solving Non-IID without Source Domain Data

Since obtaining data from the target domain poses challenges due to the complexity of downstream tasks or privacy considerations, many researchers are now exploring a challenging and practical solution that adapting the source-domain trained model with unlabeled target domain data [13, 23, 32, 49, 53]. In this paradigm, also known as test-time adaptation, the source-domain knowledge is transferred to the target domain without accessing the source-domain data during the inference phase. SAR [32] employs reliable entropy minimization and sharpness-aware minimization to achieve more stable adaptation. TAST [20] leverages useful information from nearest neighbors for classifying test data experiencing domain shifts. Tent [49] enhances the model's confidence by reducing prediction entropy and dynamically updating normalization statistics on a per-batch basis. OFTTA [53] adjusts exponential decay test-time normalization and linear classifier simultaneously to adapt for cross-person activity recognition.

6 DISCUSSION

Impact of a Suitable Pre-trained Model. The fundamental concept of DIG is to evaluate the distance between general and personalized models. By reducing this distance through adjustments using the user's unlabeled data, we can mitigate the distribution shift caused by user variances. However, if the pre-trained model is significantly inadequate, reflecting a large domain disparity, our ability for improvement is constrained, revealing limitations in handling certain scenarios.

Impact of Stopping Adaptation. As depicted in Figure 14, while our stopping criteria can result in suboptimal outcomes, attaining the optimal result remains challenging. For instance, optimal outcomes are achievable in both the UCI and SHOAIB target domains but not in the MOTION domain. Extended adjustments lead to the model internalizing our decision criteria, reducing domain disparities but diminishing returns, which can ultimately become detrimental.

Impact of Target Users Wearing Position. In reality, users may change device positions during use. As shown in Table 6, we evaluate the performance of *MobHAR* in three scenes about different positions. According to the results in *scene3*, after adaptation, the HAR model can work on multiple different positions compared with the position from source domain, reaching 71.5% F1-score and 76.4% accuracy. This is because *MobHAR* can enable source-free knowledge transfer from the pretrained HAR model, resulting in implicit generalization in target user data. Based on the results of *scene4*, the pretrained HAR model with multiple positions can enhance the generalization in target user data. Moreover, according to the results in *scene2*, the same device location has good performance since the distribution gap is lower in domain shifts. Thus, users can change their wearing position during use.

Impact of Minimum Amount of Initial Data. As shown in *scene1* of Table 6, for the same wearing position of target users, the performance of *MobHAR* after adaptation is similar for both UCI and HHAR. Since the average time of each user is 4 minutes for UCI and 25 minutes for HHAR, therefore, the minimum initial data suitable for a new user lasting 4 minutes. The model can continuously update if new data becomes available according to the DIG values. However, since there are no fixed DIG values for all users to stop updating the HAR model, currently, *MobHAR* cannot trigger the update process. A potential trigger can be a fixed time, *i.e.*, one week, one month, or one year, which can be evaluated using larger and more complex in-the-wild activity datasets [5].

7 CONCLUSION

To achieve personalized models on mobile devices with user privacy intact, we introduce *MobHAR*, a pioneering source-free HAR customization framework based on an adversarial mechanism that enables source-free knowledge transfer. We design a unique auxiliary tool called DIG to evaluate the quality of unlabeled data. Building on DIG, we select high-quality unlabeled data for fine-tuning the HAR model using minimum DIG and determine whether this data is adequate to conclude the fine-tuning process using maximum DIG. Comprehensive evaluation shows that *MobHAR* enables the customization of personalized models on mobile devices while maintaining stringent privacy safeguards.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their constructive comments. This research is supported in part by RGC under Contract CERG 16206122, 16204523, AoE/E-601/22-R, R6021-20, and Contract R8015.

REFERENCES

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine learning* 79 (2010), 151–175.

[3] Valentina Bianchi, Marco Bassoli, Gianfranco Lombardo, Paolo Fornacciari, Monica Mordonini, and Ilaria De Munari. 2019. IoT wearable sensor and deep learning: An integrated approach for personalized human activity recognition in a smart home environment. *IEEE Internet of Things Journal* 6, 5 (2019), 8553–8562.
[4] Christopher M Bishop. 2006. Pattern recognition and machine learning. *Springer google schola* 2 (2006), 5–43.
[5] Marius Bock, Hilde Kuehne, Kristof Van Laerhoven, and Michael Moeller. 2024. Wear: An outdoor sports dataset for wearable and egocentric activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–21.
[6] Youngjae Chang, Akhil Mathur, Anton Isopoussu, Junehwa Song, and Fahim Kawsar. 2020. A systematic study of unsupervised domain adaptation for robust human-activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–30.
[7] Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. 2022. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7181–7190.
[8] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. 2021. Fast batch nuclear-norm maximization and minimization for robust domain adaptation. *arXiv preprint arXiv:2107.06154* (2021).
[9] Somayeh Danafar, Paola Rancoita, Tobias Glasmachers, Kevin Whittingstall, and Juergen Schmidhuber. 2013. Testing hypotheses by regularized maximum mean discrepancy. *arXiv preprint arXiv:1305.0423* (2013).
[10] Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. 2021. Adaptive methods for real-world domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14340–14349.
[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research* 17, 59 (2016), 1–35.
[12] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.
[13] Taesik Gong, Yewon Kim, Adiba Orzikulova, Yunxin Liu, Sung Ju Hwang, Jinwoo Shin, and Sung-Ju Lee. 2023. DAPPER: Label-Free Performance Estimation after Personalization for Heterogeneous Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–27.
[14] Taesik Gong, Yeonsu Kim, Jinwoo Shin, and Sung-Ju Lee. 2019. Metasense: few-shot adaptation to untrained conditions in deep mobile sensing. In *Proceedings of the 17th Conference on Embedded Networked Sensor Systems*. 110–123.
[15] Andreas Grammenos, Cecilia Mascolo, and Jon Crowcroft. 2018. You are sensing, but are you biased? a user unaided sensor calibration approach for mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.
[16] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.
[17] Harish Haresamudram, Irfan Essa, and Thomas Plötz. 2021. Contrastive predictive coding for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (2021), 1–26.
[18] Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. 2024. CrossHAR: Generalizing Cross-dataset Human Activity Recognition via Hierarchical Self-Supervised Pretraining. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–26.
[19] Yash Jain, Chi Ian Tang, Chulhong Min, Fahim Kawsar, and Akhil Mathur. 2022. Collossl: Collaborative self-supervised learning for human activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–28.
[20] Minguk Jang, Sae-Young Chung, and Hye Won Chung. 2023. Test-time adaptation via self-training with nearest neighbor information. *Proceedings of the International Conference on Learning Representations (ICLR)* (2023).
[21] Wenchao Jiang and Zhaozheng Yin. 2015. Human activity recognition using wearable sensors by deep convolutional neural networks. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1307–1310.
[22] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. [n. d.]. Minimum class confusion for versatile domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 464–480.
[23] Hua Kang, Qingyong Hu, and Qian Zhang. 2024. SF-Adapter: Computational-Efficient Source-Free Domain Adaptation for Human Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–23.
[24] Md Abdullah Al Hafiz Khan, Nirmalya Roy, and Archan Misra. 2018. Scaling human activity recognition via deep learning-based domain adaptation. In *2018 IEEE international conference on pervasive computing and communications (PerCom)*. 1–9.
[25] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology* 49, 4 (2013), 764–766.
[26] Huining Li, Huan Chen, Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Xiaoye Qian, Dongmei Li, Ming-chun Huang, and Wenyao Xu. 2023. Neuralgait: Assessing brain health using your smartphone. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–28.
[27] Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. 2020. Giobalfusion: A global attentional deep learning framework for multisensor information fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable*

and Ubiquitous Technologies 4, 1 (2020), 1–27.
[28] Wang Lu, Jindong Wang, Yiqiang Chen, Sinno Jialin Pan, Chunyu Hu, and Xin Qin. 2022. Semantic-discriminative mixup for generalizable sensor-based cross-domain activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–19.
[29] Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. 2022. The" Beatrix"Resurrections: Robust Backdoor Detection via Gram Matrices. *arXiv preprint arXiv:2209.11715* (2022).
[30] Mohammad Malekzadeh, Richard G Clegg, Andrea Cavallaro, and Hamed Haddadi. 2019. Mobile sensor data anonymization. In *Proceedings of the international conference on internet of things design and implementation*. 49–58.
[31] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. 2017. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning* 10, 1-2 (2017), 1–141.
[32] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards stable test-time adaptation in dynamic wild world. *In 11th International Conference on Learning Representations* (2023).
[33] Preksha Pareek and Ankit Thakkar. 2021. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review* 54 (2021), 2259–2322.
[34] Hangwei Qian, Sinno Jialin Pan, and Chunyan Miao. 2021. Latent independent excitation for generalizable sensor-based cross-person activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11921–11929.
[35] Hangwei Qian, Tian Tian, and Chunyan Miao. 2022. What makes good contrastive learning on small-scale wearable-based tasks?. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3761–3771.
[36] Xin Qin, Yiqiang Chen, Jindong Wang, and Chaohui Yu. 2019. Cross-dataset activity recognition via adaptive spatial-temporal transfer learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–25.
[37] Xin Qin, Jindong Wang, Yiqiang Chen, Wang Lu, and Xinlong Jiang. 2022. Domain generalization for activity recognition via adaptive feature fusion. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (2022), 1–21.
[38] Xin Qin, Jindong Wang, Shuo Ma, Wang Lu, Yongchun Zhu, Xing Xie, and Yiqiang Chen. 2023. Generalizable low-resource activity recognition with diverse and discriminative representation learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1943–1953.
[39] Jorge-L Reyes-Ortiz, Luca Oneto, Albert Samà, Xavier Parra, and Davide Anguita. 2016. Transition-aware human activity recognition using smartphones. *Neurocomputing* 171 (2016), 754–767.
[40] Aaqib Saeed, Tanir Ozcelebi, and Johan Lukkien. 2019. Multi-task self-supervised learning for human activity detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–30.
[41] Meyer Scetbon and Gaël Varoquaux. 2019. Comparing distributions: $l1$ geometry improves kernel two-sample testing. In *NeurIPS 2019-33th Conference on Neural Information Processing Systems*.
[42] Muhammad Shoaib, Stephan Bosch, Ozlem Durmaz Incel, Hans Scholten, and Paul JM Havinga. 2014. Fusion of smartphone motion sensors for physical activity recognition. *Sensors* 14, 6 (2014), 10146–10176.
[43] Samarth Sinha, Zhengli Zhao, Anirudh Goyal ALIAS PARTH GOYAL, Colin A Raffel, and Augustus Odena. 2020. Top-k training of gans: Improving gan performance by throwing away bad samples. *Advances in Neural Information Processing Systems* 33 (2020), 14638–14649.
[44] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*. 127–140.
[45] Scott Sun, Dennis Melamed, and Kris Kitani. 2021. IDOL: Inertial deep orientation-estimation and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6128–6137.
[46] Chi Ian Tang, Ignacio Perez-Pozuelo, Dimitris Spathis, Soren Brage, Nick Wareham, and Cecilia Mascolo. 2021. Selfhar: Improving human activity recognition through self-training with unlabeled data. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 1 (2021), 1–30.
[47] Terry T Um, Franz MJ Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. 2017. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 216–220.
[48] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[49] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully test-time adaptation by entropy minimization. *In 9th International Conference on Learning Representations* (2021).
[50] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering* (2022).
[51] Jinqiang Wang, Tao Zhu, Jingyuan Gan, Liming Luke Chen, Huansheng Ning, and Yaping Wan. 2022. Sensor data augmentation by resampling in contrastive learning for human activity recognition. *IEEE Sensors Journal* 22, 23 (2022), 22994–23008.
[52] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. 2022. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7201–7211.

[53] Shuoyuan Wang, Jindong Wang, Huajun Xi, Bob Zhang, Lei Zhang, and Hongxin Wei. 2024. Optimization-Free Test-Time Adaptation for Cross-Person Activity Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–27.
[54] Yan Wang, Shuang Cang, and Hongnian Yu. 2019. A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications* 137 (2019), 167–190.
[55] Huatao Xu, Pengfei Zhou, Rui Tan, and Mo Li. 2023. Practically Adopting Human Activity Recognition. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.
[56] Huatao Xu, Pengfei Zhou, Rui Tan, Mo Li, and Guobin Shen. 2021. Limu-bert: Unleashing the potential of unlabeled data for imu sensing applications. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 220–233.
[57] Shuochao Yao, Shaohan Hu, Yiran Zhao, Aston Zhang, and Tarek Abdelzaher. 2017. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th international conference on world wide web*. 351–360.
[58] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sy8gdB9xx
[59] Junru Zhang, Lang Feng, Zhidan Liu, Yuhan Wu, Yang He, Yabo Dong, and Duanqing Xu. 2024. Diverse Intra-and Inter-Domain Activity Style Fusion for Cross-Person Generalization in Activity Recognition. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
[60] Mi Zhang and Alexander A. Sawchuk. 2012. USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors. In *ACM International Conference on Ubiquitous Computing (Ubicomp) Workshop on Situation, Activity and Goal Awareness (SAGAware)*. Pittsburgh, Pennsylvania, USA.
[61] Yi-Fan Zhang, Zhang Zhang, Da Li, Zhen Jia, Liang Wang, and Tieniu Tan. 2022. Learning domain invariant representations for generalizable person re-identification. *IEEE Transactions on Image Processing* 32 (2022), 509–523.

8 APPENDIX

8.1 A Detailed Derivation of the DIG Formulation

Let $x_t^p \sim P_{t,p}$ and $x_{ta}^p \sim P_{ta,p}$ denote the p-th order representation derived from the unseen target user distribution $P_{t,p}$ and the augmented distributions $P_{ta,p}$ of the user, respectively. *MobHAR* calculates representations in feature space without making any assumptions about $P_{t,p}$ and $P_{ta,p}$. Specifically, *MobHAR* derives discriminative information from the statistical moments of $P_{t,p}$ and $P_{ta,p}$ [4]. Based on x_t^p and $P_{t,p}$, we can calculate the second raw moment as follows:

$$\begin{aligned} E\left(x_t^p x_t^{pT}\right) &= E(x_t^p)E(x_t^p)^T + E[(x_t^p - \mu_1^p)(x_t^p - \mu_1^p)^T] \\ &= \mu_1^p \mu_1^{pT} + \sigma_1^p, \end{aligned} \tag{8}$$

where μ_1^p and σ_1^p denote the mean vector and the covariance matrix of x_t^p. According to Equation 2, we can obtain the Gramian information of x_t^p as $G_{x_t^p} = x_t^p x_t^{pT}$. Therefore, combining with Equation 8, we can derive that $E(G_{x_t^p}) = E(x_t^p x_t^{pT}) = \mu_1^p \mu_1^{pT} + \sigma_1^p$. Similarly, we can obtain $E(G_{x_{ta}^p}) = \mu_2^p \mu_2^{pT} + \sigma_2^p$, where μ_2^p and σ_2^p denote the mean vector and the covariance matrix of x_{ta}^p. After this, we can define the DIG as a view of statistical moments:

$$\begin{aligned} DIG\,(P_t, P_{ta}) &= E_p[E(G_{x_t^p}) - E(G_{x_{ta}^p})] \\ &= E_p[\mu_1^p \mu_1^{pT} - \mu_2^p \mu_2^{pT} + \sigma_1^p - \sigma_2^p], \end{aligned} \tag{9}$$

where P_t denotes a collection of the representations from the unseen target user distribution with elements of different powers, and P_{ta} denotes that of the augmented distributions of these users.

8.2 A Detailed Derivation of Source-free Knowledge Transfer

To reduce the disparity between a general model and personalized models, we develop a structure based on the pre-trained HAR model f and a softmax layer as the classifier S. Given Y human activity categories and N_u samples from the unseen target user, the classifier S can produce a prediction matrix $H \in \mathbb{R}^{Y \times N_u}$. For the

self-correlation matrix of the HAR model's prediction $R = H^T H$, H satisfies

$$\sum_{y=1}^{Y} H_{i,y} = 1, \forall i \in 1 \dots N_u, \text{and}$$
$$H_{i,y} \geq 0, \forall i \in 1 \dots N_u, y \in 1 \dots Y. \tag{10}$$

Therefore, we denote the I_i as $\sum_{y=1}^{Y} R_{i,y}$ and the I_o as $\sum_{y \neq i}^{Y} R_{i,y}$. According to Equation 10, I_i and I_o satisfy $I_i + I_o = N_u$. Moreover, I_i is the same as Frobenius-norm [7] of the HAR model's prediction H, *i.e.*, $\|H\|_F$. Therefore, we can get $I_i - I_o = 2 \times \|H\|_F - N_u$. Since the classifier S produces H, we can derive a loss function as

$$\mathcal{L}_{raw-F} = 2 \times \|S\|_F - N_u, \tag{11}$$

which can give high scores for the samples with low distribution gap between the new user and pretrained model's training users and low scores for the samples with large distribution gap. According to Equation 11, both the coefficients of the polynomial 2 and constant term N_u are constant. Thus, we can simplify the Equation 11 as $\mathcal{L}_{raw-F} = \|S\|_F$.

Inspired by [1], to get high scores for the samples with a low distribution gap between the new user and pretrained model's training users and low scores for the samples with a large distribution gap, an intuitive method is to learn a K-Lipschitz loss function h. Using the 1-Wasserstein distance, we can get the distribution gap among the unseen target user distribution $P_{t,p}$ and the augmented distributions $P_{t_a,p}$ of the user by

$$W_1(P_{t,p}, P_{t_a,p}) = \mathbb{E}_{f \sim P_{t,p}}[h(f)] - \mathbb{E}_{f \sim P_{t_a,p}}[h(f)]. \tag{12}$$

According to $\mathcal{L}_{raw-F} = \|S\|_F$, $\|S\|_F$ can serve as a loss function. Therefore, the distribution gap between $P_{t,p}$ and $P_{t_a,p}$ can be written as:

$$W_w(P_{t,p}, P_{t_a,p}) = \mathbb{E}_{f \sim P_{t,p}}[\|S\|_F] - \mathbb{E}_{f \sim P_{t_a,p}}[\|S\|_F]. \tag{13}$$

Since the discrepancy between I_i and I_o, as measured by the Frobenius-norm Wasserstein distance, may overlook the classes with small samples [8], we use the Nuclear norm $\|\cdot\|_*$ to enhance prediction diversity [7, 8]. To mitigate the need for repetitive alternating updates in the HAR model, we employ a gradient reverse layer (GRL) [11], which facilitates updates within a single backpropagation step. As a result, *MobHAR* conducts source-free knowledge transfer as follows:

$$\mathcal{L}_{\theta_f} = min\{\frac{1}{N_u} \sum_{i=1}^{N_u} \left(\left\| S(f(x_i^t)) \right\|_* - \left\| S(f(x_i^{ta})) \right\|_* \right)\}, \tag{14}$$

where f denotes the pre-trained HAR model, N_u denotes the number of samples from the unseen target user, x^t denotes the sample from the unseen target user t, x^{ta} denotes the augmentation data samples of t, and S denotes the softmax layer.