ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240535

Class-Aware Sample Weight Learning for Cross-Modal Unsupervised Domain Adaptation in Cross-User Wearable Human Activity Recognition

Tong Wu[a,*], **Yanbin Liu**[a] **and Sira Yongchareon**[a]

[a]School of Engineering, Computer and Mathematical Sciences, Auckland University of Technology, New Zealand
ORCID (Tong Wu): https://orcid.org/0000-0002-2888-275X, ORCID (Yanbin Liu): https://orcid.org/0000-0003-4724-8065, ORCID (Sira Yongchareon): https://orcid.org/0000-0002-2880-6618

Abstract. Existing unsupervised domain adaption approaches for cross-user Wearable Human Activity Recognition (WHAR) typically assume that users utilize the uni-modal sensor deployment configuration and cannot transfer across different sensor modalities. In this paper, we consider the more realistic cross-modal wearable human activity recognition setting to investigate the unsupervised domain adaptation task. This new context presents two formidable challenges: (1) how to alleviate modality heterogeneity across users, and (2) how to explore cross-modal domain correlation for better unsupervised domain adaptation. We propose a cross-modal unsupervised domain Adaptation model with Class-Aware Sample Weight Learning (**CASWL-Adapt**) to address both challenges. First, a *spherical modality discriminator* is designed to capture modal-specific discriminative features of each user during domain adaptation, thus achieving a reduction of sample variance caused by modal heterogeneity. Given a user-specific modal, modality-independent domain-invariant features can be efficiently generated by the well-developed modality discrimination loss and adversarial training. Second, a *class-aware weight network* is devised to calculate sample weights through classification loss and activity class similarity for each sample. Furthermore, the network leverage end-to-end learning and meta-optimization update rules to explore inter-domain correlations. Cross-modal activity classes are expected to adaptively implement different weighting schemes based on their intrinsic bias characteristics to select the most appropriate samples for domain knowledge transfer. We demonstrate that CASWL-Adapt achieves state-of-the-art results on three challenging benchmarks: Epic-Kitchens, Multimodal-EA and RealWorld, especially effective for new users of unseen modality.

1 Introduction

Wearable devices based on sensors can be used to collect data through sensing devices mounted on the subject, which in turn can be used to perform Wearable Human Activity Recognition (WHAR). Since such devices can protect user privacy in a non-intrusive way and can obtain information from multiple sensors installed in different parts of the human body without affecting the user's normal life and work, this makes sensor-based HAR research more realistic [33]. Although WHAR has been widely used in many applications [5], the real-world application still faces the challenge of high-precision recognition triggered by the scarcity of data annotation.

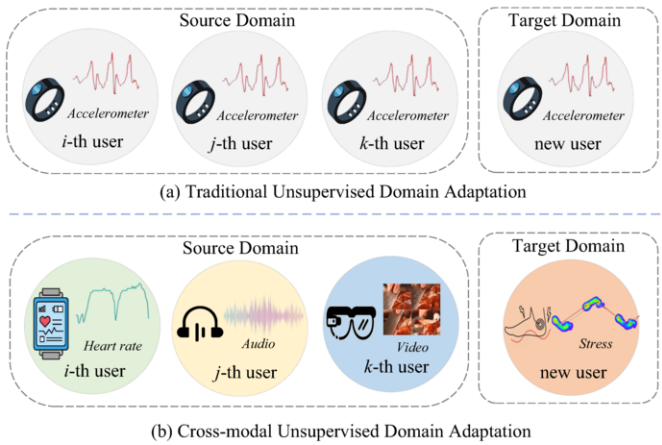

Figure 1. (a) Traditional UDA methods assume the same uni-modal sensor deployment across different users. (b) Our method deals with a more practical cross-modal UDA setting (CMA-WHAR) in real-world scenarios.

Unsupervised Domain Adaptation (UDA) aims to train a model that can migrate knowledge learned in the source domain with labeled data to the target domain with unlabeled data, thus alleviating the problem of scarcity of user data annotations in real-world scenarios. Recent progress has been made with domain difference metrics [13, 15, 23] and adversarial learning [42, 4, 37] domain invariant representations: instead of learning any training classes explicitly, these approaches utilize the training classes by attempting to learn transfer through finding commonalities between the source and target domains. Excellent results are achieved on common benchmarks (*e.g.*, RealWorld [34]) by a series of methods [40, 9, 21]. Despite their success, most of them are trained and evaluated only on sensor datasets of the same type (*i.e., uni-modal*), and fail to learn generalized models across different types of sensors (*i.e., cross-modal*). In practical applications, the need for cross-modal adaptation is imperative as it will facilitate the deployment of models on large-scale user wearables [18, 14]. For example, we would like models developed using accelerometer sensors to be adaptable to visual sensors without the necessity of gathering extra target training examples (Figure 1).

* Corresponding Author. Email: tong.wu@autuni.ac.nz.

To break the limitations of existing UDA methods and benchmarks, we propose a new **Cross-Modal** unsupervised domain **Adaptation** task for cross-user **Wearable Human Activity Recognition** (CMA-WHAR), which focuses on adapting to new users of unseen modalities using models learned from user data of different modalities as shown in Figure 1. It differs from traditional unimodal unsupervised domain adaptation in two aspects: (1) it contains different modal data from multiple different users for training and unseen modal data for testing; and (2) there is a large modal heterogeneity between the source and the target domains, and also potential domain correlation across modalities.

These differences pose two challenges for Cross-User Wearable Human Activity Recognition. **(1)** How to alleviate modality heterogeneity across users. In traditional UDA settings, where user data comes from the same type of sensor, the heterogeneity of the sample distribution mainly originates from the different types of activities of different users [26, 2]. However, in the CMA-WHAR task, there are different feature dimensions and modality heterogeneity between data from different users, a property that leads to greater distributional heterogeneity. **(2)** How to explore cross-modal domain correlation for better unsupervised domain adaptation. Traditional UDA tasks only consider user data from the same type of modality, which makes the knowledge learned on the source domain user samples always shareable by the target domain. In contrast, for the CMA-WHAR task, different recognition patterns are required for different activity classes due to the huge heterogeneity among user data. For example, the ability to recognize activity classes using visual data is ambiguous for those activity classes only using accelerometer or gyroscope signals [39].

To address the above challenges of the CMA-WHAR task, we propose a cross-modal unsupervised domain Adaptation model with Class-Aware Sample Weight Learning **(CASWL-Adapt)** for cross-user WHAR. First, a *spherical modality discriminator* is designed to capture modal-specific discriminative features of each user during domain adaptation, thus achieving a reduction of sample variance caused by modal heterogeneity. Given a user-specific modal, modality-independent domain-invariant features can be efficiently generated by the well-developed modality discrimination loss and adversarial training.

Second, to explore the cross-modal domain correlation, we devise a novel *class-aware weight network*, which treats each user activity class as an individual learning task (taking sample loss along with class/task features as input and the sample weights as output). Specifically, we conduct the K-means clustering of all activity classes to generate task clusters of different sample sizes. These clusters are combined with a sample weighting network to compose the class-aware weight network. Then, with the class-aware weight network, per sample weight is computed from the classification loss and activity class similarity of each sample. Moreover, the class-aware weight network leverages end-to-end learning and meta-optimization update rules to explore inter-domain correlations. Intuitively, the cross-modal activity classes are expected to adaptively implement different weighting schemes based on their intrinsic bias characteristics to select the most appropriate samples for domain knowledge transfer.

Experimentally, we show the effectiveness of each devised component, and explain how the class-aware weight network works on cross-modal domain correlation through a modality-strictly-differentiated setup. In summary, our contributions are threefold:

- We propose a new cross-modal UDA task (CMA-WHAR) for cross-user wearable human activity recognition, which is important for deploying HAR models with different data modalities on large-scale user wearable devices.
- We propose a cross-modal unsupervised domain Adaptation method with Class-Aware Sample Weight Learning (CASWL-Adapt). It can effectively alleviate cross-modal heterogeneity and explore cross-modal domain correlation.
- We achieve state-of-the-art performance on the challenging Epic-Kitchens, Multimodal-EA and RealWorld benchmarks, especially effective for new users of unseen modality.

2 Related Work

2.1 Unsupervised Domain Adaptation

The difference in domain distribution leads to the poor performance of deep network models when trained on the source domain and applied to the target domain. And these models typically face the problem of scarcity of data labeling in practical applications, which prevents them from being used in real-world scenarios [36]. To alleviate this problem, existing methods generally use unsupervised domain adaptation (UDA) to guide the human action recognition (HAR) model. The prevailing inference of this approach is to reduce inter-domain distributional differences by aligning human activity-related features in the source/target domains [25, 20]. In addition, the more popular approach [10, 38] exploits the confounding properties of generative adversarial networks to learn domain-invariant feature representations of domain discriminators inversely. Alternatively, the recognition error between different domains can be reduced by sample mapping [9, 40].

2.2 Sample Weight Learning in UDA

Inspired by the development of meta-learning [12], some recent work [35, 23, 24] attempts to build models that can adaptively learn the weights of data samples to make learning more automatic, which improve the reliability in some approaches using UDA for human activity recognition. In addition, some work [31, 41, 28] attempts to align the differences between different samples in particular scenarios, allowing some of the negative samples to be selectively discarded during the domain alignment process. The significant limitations of these UDA models are that they all require manual pre-specification of the form of the sample weighting function and complex hyper-parameter settings that rely on a priori knowledge, which results in a model with limited flexibility to adapt to user data that have an interclass bias-heterogeneous distribution [30].

2.3 UDA for Cross-user WHAR

UDA has made some progress on cross-user WHAR [4, 27, 42]. With the rise of deep learning, the use of deep neural networks [6, 3, 11] as reliable feature extractors has become a consensus. It is worth noting that all of these approaches require the data acquisition source to be controlled within a single sensor, which means that the HAR model is not designed to handle cross-modal differences in user data. Therefore, when applied to cross-user WHAR tasks in real-life scenarios where data modalities are heterogeneous, it may show degraded performance [5]. In addition, these methods require retraining of the model when it is tested with datasets of different modalities. Unlike the previous methods, we proposed a novel cross-modal unsupervised domain adaptation model with class-aware sample weight learning, which is both effective and free from re-training for multi-modal datasets.

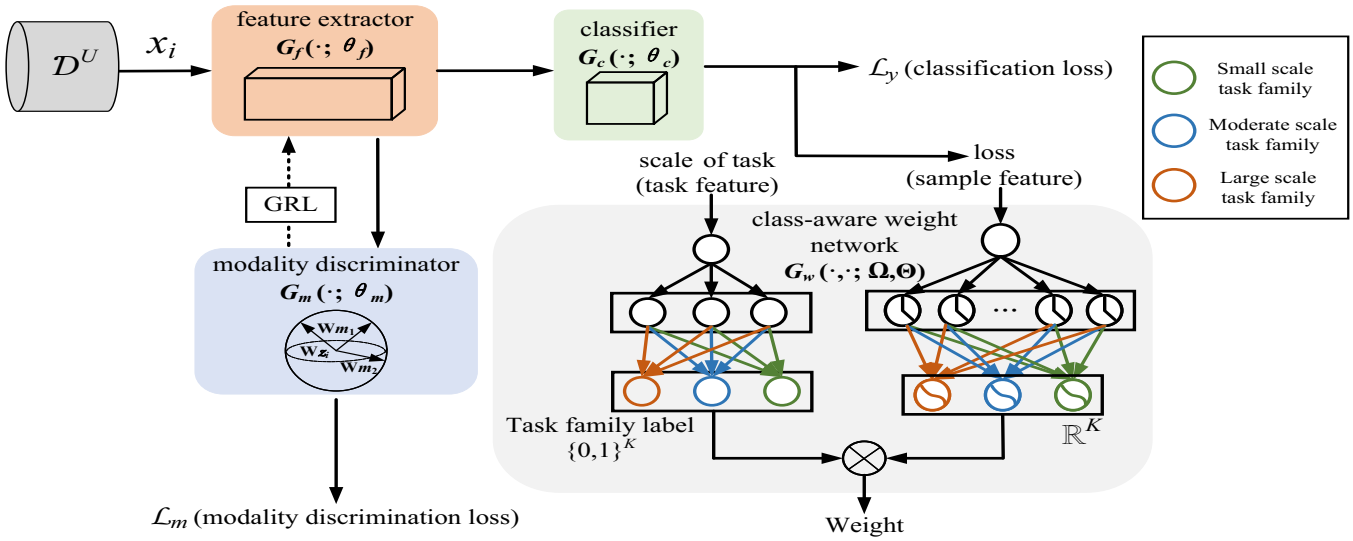

Figure 2. The proposed **CASWL-Adapt** framework for cross-user wearable human activity recognition. At first, we build the basic activity recognition network by employing the feature extractor and the classifier. Then, we design two novel components for cross-modal cross-user recognition. **(1)** The *spherical modality discriminator* is adversarially trained on user data from source and target domains through a Gradient Reversal Layer (GRL) to obtain cross-modal domain invariant features. **(2)** In the *class-aware weighting network*, samples from the source domain are guided by samples from the target domain, and adaptive weighting is implemented to achieve cross-modal domain correlation.

3 Methodology

3.1 Problem Definition

For the CMA-WHAR task, we treat it as a multimodal multi-classification task with data bias. The difficulty of this task is mainly due to the collection of multimodal user data in real-life scenarios and the imbalance of user sample classes across modalities. Formally, We denote a set of labeled data with several training users as the source dataset $\mathcal{D}^S = \{(x_i, y_i, z_i)\}_{i=1}^{n_S}$, where x_i is the sample, y_i is the activity class label, z_i is the modality label and n_S is the number of samples from source domain users. Meanwhile, we also require the new user data (target dataset) $\mathcal{D}^{\mathrm{T}} = \{\mathbf{x}_i, \mathbf{z}_i\}_{i=n_{\mathrm{S}}+1}^{n_{\mathrm{S}}+n_{\mathrm{T}}}$ with samples $\mathbf{x}_i$ and modality labels $\mathbf{z}_i$ from the target domain, where n_T is the number of samples from target domain users. The total number of samples is $N = n_S + n_T$, and we use $\mathcal{D}^U = \{(x_i, z_i, d_i)\}_{i=1}^{N}$ to denote all samples with domain label d_i. Here, $d_i = 0$ for source domain samples and $d_i = 1$ for target domain samples. We assume the source domain and target domain (new users) share the same activity set. The new user's modality label ($\mathbf{z}_i$) is known, and only the activity class label is unlabeled.

3.2 Overall Framework of CASWL-Adapt

Our overall framework of the proposed CASWL-Adapt is shown in Figure 2, which consists of four subnetworks, *i.e.*, feature extractor $G_f(\cdot;\theta_f)$, classifier $G_c(\cdot;\theta_c)$, spherical modality discriminator $G_m(\cdot;\theta_m)$ and class-aware weight network $G_w(,;\Theta,\Omega)$ with learnable parameters θ_f, θ_c, θ_m, Θ and Ω, respectively.

In current WHAR applications, the feature extractor and classifier are usually composed using convolutional neural networks and deep neural networks. Similarly, our activity recognition network also adopts this standard feed-forward architecture. However, we designed several novel modules to address the cross-user cross-modal action recognition setting. Specifically, in our model, the *spherical modality discriminator* is connected to the feature extractor via a Gradient Reversal Layer (GRL) to achieve cross-modal unsupervised domain adaptation. In the subsequent processing, we pass the sample and class/task features together to the *class-aware weight network* to generate per sample weight, which is then multiplied by the outputs. Finally, using the target domain samples as guidance, the results of the class-aware weighting network will be used to perform class-aware adaptive reweighting on the source domain samples.

3.3 Model Details and Training Losses

Next, we illustrate the important model components: the spherical modality discriminator used to constrain CASWL-Adapt, the class-aware weight network used for adaptive re-weighting of source domain samples, as well as more details on the model loss function.

3.3.1 Modality Discriminator and Discrimination Loss

Similar to the traditional domain classifier setting, we expect to train the spherical modality discriminator through an adversarial approach [10] to make it indistinguishable between samples from the source and target domains and to assume that the features of the two domains are already aligned by default.

To make it easier to explain, we first represent the spherical modality discriminator as $G_m(g;\theta_m) = Softmax(W^T\nu(g))$, where $g = G_f(x;\theta_f)$ denote the deep features from the feature extractor $G_f(\cdot;\theta_f)$ and ν denote a single-layer perception with the ReLU activation function. W is the weight matrix without bias terms in the modality discriminator. Since each user has only one class of modality and there is an inter-modal similarity between samples from different users, this creates a significant ambiguity in modal decision boundaries when using multi-classification loss function (*e.g.*, cross-entropy loss). To address this issue, we perform a small amount of normalization on each column of W and use additive angular margin

loss [8] to enhance the intra-class tightness of the modality discriminator in the spherical:

$$\mathcal{L}_{am}(g,z) = -\log \frac{e^{\delta\cos(\theta_z+\eta)}}{e^{\delta\cos(\theta_z+\eta)} + \sum_{j=1,j\neq z}^{M} e^{\delta\cos\theta_j}}, \quad (1)$$

where $\cos\theta_j = \arccos\left(\frac{W_j^T}{\|W_j\|}\cdot\frac{g_i}{\|g_i\|}\right)$, and θ_j is the angle between the deep feature g_i of the i-th sample x_i and weight W_j. Similarly, an additive angular margin factor η is added between the sample feature g_i and the target weight W_z to obtain $\cos(\theta_z+\eta)$. Furthermore, δ is a scale factor and z denotes the ground-truth modality label. M denotes the total number of modalities. Based on the above idea, we can derive modality discrimination loss for all user samples as follows:

$$\mathcal{L}_m(\theta_f,\theta_m) = \frac{1}{N}\sum_{i\in\{d_i=0,1\}} \mathcal{L}_{am}(g_i,z_i)\ . \quad (2)$$

3.3.2 Classification Loss and Class-aware Weight Network

Generally, when using the activity recognition network for standard training, we can obtain the classification loss of the CMA-WHAR task on the training samples:

$$\mathcal{L}_y(\theta_f,\theta_c) = \frac{1}{n_S}\sum_{i\in\{d_i=0\}} \ell\left(G_c\left(g_i;\theta_c\right),y_i\right), \quad (3)$$

where $G_c(g_i;\theta_c)$ denotes the network output and ℓ is the cross-entropy loss function. For notation convenience, we denote that $\mathcal{L}_i^y(\theta_c) = \ell\left(G_c\left(g_i;\theta_c\right),y_i\right)$. However, due to the modality heterogeneity between different user data, activity recognition networks with deep neural networks as a classifier can easily overfit biased training data with class imbalance, making it difficult to generalize existing UDA models to real-world scenarios.

Inspired by the explicit class-aware mapping of sample weights [30], we use a class-aware weight network to adaptively learn the explicit weighting of training samples, which is used to overcome the apparent inter-class bias variations in heterogeneous user data. Specifically, we pass the sample and task features together to the class-aware weight network. In our study, we attempt to take the scale level (*i.e.*, the number of training samples) of each training class/task to represent its task feature. Specifically, denote $N_i\,(i=1,...,N;d_i=0)$ as the number of samples contained in activity classes to which the i-th sample x_i belongs and put it as a task feature into the left branch of the network. This branch contains a hidden layer of *K* nodes with a *K*-level scale $\Omega=\{\mu_k\}_{k=1}^K$ in ascending order. The outcome of this branch is a *K*-dimensional one-hot vector (*i.e.*, a task family label) whose 1 element is located in its *K*-th dimension, corresponding to μ_k that is closest to the input N_i. For notation convenience, we denote the left branch as $\mathrm{L}(N_i;\Omega)\in\{0,1\}^K$. Then, we pass the $\mathcal{L}_i^y(\theta_c)$ to the right branch of the network. This branch inputs the classification loss value of the i-th training sample into a multi-layer perceptron (MLP) consisting of a hidden layer and a *K*-dimensional weighted output. For notation convenience, we denote the right branch as $\mathrm{R}(\mathcal{L}_i^y(\theta_c);\Theta)\in[0,1]^K$. Then the class-aware weight network function can be formulated as follows:

$$G_w\left(\mathcal{L}_i^y(\theta_c),N_i;\Theta,\Omega\right) = \mathrm{L}\left(N_i;\Omega\right)\otimes\mathrm{R}\left(\mathcal{L}_i^y(\theta_c);\Theta\right), \quad (4)$$

where $\bigotimes$ denotes the dot product between two vectors. By modulating the high-level task feature information, the class-aware weight network is expected to learn a class-aware function by accumulating class/task with homogeneous bias situations and allowing different classes/tasks to adaptively implement different weighting schemes based on their inherent bias situation. Now, the objective function of the class-aware weight network can be simplified to the following bi-level optimization problem:

$$\{\Theta^*,\Omega^*\} = \arg\min_{\Theta,\Omega}\frac{1}{n_T}\sum_{i\in\{d_i=1\}}\mathcal{L}_i^{meta}(\theta_c^*,\Theta,\Omega)\,,$$
$$\theta_c^* = \arg\min_{\theta_c}\frac{1}{n_S}\sum_{i\in\{d_i=0\}} G_w\left(\mathcal{L}_i^y(\theta_c),N_i;\Theta,\Omega\right)\mathcal{L}_i^y(\theta_c)\,, \quad (5)$$

where $\mathcal{L}^{meta}(\theta_c^*,\Theta,\Omega)$ is the meta classification loss computed using the parameter θ_f and θ_c^* by using the target domain user data. Notably, the class-aware weight network needs to be updated with unlabeled target domain user data for the meta-learning approach. Therefore, the target domain's user data must be pseudo-labeled [18] with the activity recognition network before a new round of model training.

3.4 Optimization

To make the model easy to optimize, we pre-count the number of samples in different activity classes N_i and apply the standard K-means algorithm [1] to the training samples to obtain the cluster centers $\Omega=\{\mu_k\}_{k=1}^K$ sorted in ascending order. Based on the tuning experience of previous work [30], we set $K=3$ to achieve differentiation between small, moderate, and large task families with different user data. At each training step, we sample a small batch of labeled source samples and another small batch of unlabeled target samples. We denote the learning rate of the class-aware weight network as β and the learning rate of the other three sub-networks α.

1) Updating feature extractor and classifier parameters. The first update formulates the learning modality of the activity recognition network so that it can learn category differentiation patterns on labeled source domain samples. The formulation is as follows:

$$\hat{\theta}_f^{(t+1)} = \theta_f^{(t)} - \alpha\nabla_{\theta_f^{(t)}}\mathcal{L}_y(\theta_f^{(t)},\theta_c^{(t)})\,, \quad (6)$$
$$\hat{\theta}_c^{(t+1)} = \theta_c^{(t)} - \alpha G_w\left(\mathcal{L}^y(\theta_c^{(t)});\Theta^{(t)}\right)\nabla_{\theta_c^{(t)}}\mathcal{L}^y(\theta_c^{(t)})\,. \quad (7)$$

2) Updating class-aware weight network parameters. We design a meta-optimization updating rule for the second update and aim at learning an adaptive sample weighting function, which helps to enable the activity recognition network to have the ability to specify an appropriate weighting scheme based on the internal bias characteristics of the different classes/tasks themselves. The formulation is as follows:

$$\Theta^{(t+1)} = \Theta^{(t)} - \beta\nabla_{\Theta^{(t)}}\mathcal{L}^{meta}(\hat{\theta}_c^{(t+1)},\Theta^{(t)})\,, \quad (8)$$
$$\theta_c^{(t+1)} = \theta_c^{(t)} - \alpha G_w\left(\mathcal{L}^y(\theta_c^{(t)});\Theta^{(t+1)}\right)\nabla_{\theta_c^{(t)}}\mathcal{L}^y(\theta_c^{(t)}). \quad (9)$$

3) Updating feature extractor and modality discriminator parameters. The third update enables the spherical modality discriminator to learn the modality discriminative knowledge on labeled source samples and pseudo-labeled target samples. This adversarial

Table 1. Comparison with the state-of-the-art UDA models (mean±std). "$*$" indicates that CASWL-Adapt is statistically superior to the compared model according to the pairwise t-test at a 95% significance level.

Model (Venue)	Epic-Kitchens		Multimodal-EA		RealWorld	
	Acc.	Mac.F1	Acc.	Mac.F1	Acc.	Mac.F1
HDCNN (PerCom 2018)	0.587±0.004*	0.565±0.004*	0.624±0.003*	0.611±0.004*	0.651±0.009*	0.632±0.006*
MMD (TPAMI 2019)	0.582±0.006*	0.567±0.007*	0.623±0.011*	0.607±0.005*	0.645±0.012*	0.629±0.008*
AdvSKM (IJCAI 2021)	0.628±0.003*	0.612±0.008*	0.689±0.006*	0.679±0.004*	0.713±0.013*	0.683±0.012*
DANN (JMLR 2016)	0.639±0.007*	0.621±0.009*	0.708±0.005*	0.685±0.003*	0.738±0.011*	0.712±0.007*
DUA (CVPR 2022)	0.566±0.011*	0.563±0.008*	0.647±0.007*	0.632±0.005*	0.678±0.003*	0.636±0.005*
ETN (CVPR 2019)	0.648±0.005*	0.632±0.009*	0.703±0.004*	0.689±0.008*	0.739±0.012*	0.709±0.027*
TCL (AAAI 2019)	0.652±0.007*	0.639±0.007*	0.719±0.005*	0.705±0.011*	0.741±0.025*	0.705±0.032*
SS-UniDA (AAAI 2021)	0.657±0.009*	0.642±0.002*	0.722±0.001*	0.707±0.008*	0.743±0.031*	0.712±0.039*
SWL-Adapt (AAAI 2023)	0.663±0.006*	0.643±0.003*	0.727±0.007*	0.723±0.006*	0.753±0.038*	0.742±0.048*
CASWL-Adapt	**0.694±0.005**	**0.688±0.004**	**0.748±0.008**	**0.743±0.002**	**0.764±0.039**	**0.756±0.042**

goal is achieved through a gradient reversal layer (GRL), which is updated as follows:

$$\theta_f^{(t+1)} = \hat{\theta}_f^{(t+1)} - \alpha \nabla_{\hat{\theta}_f^{(t+1)}} \mathcal{L}_m(\hat{\theta}_f^{(t+1)}, \theta_m^{(t)}), \tag{10}$$

$$\theta_m^{(t+1)} = \theta_m^{(t)} + \alpha \nabla_{\theta_m^{(t)}} \mathcal{L}_m(\hat{\theta}_f^{(t+1)}, \theta_m^{(t)}). \tag{11}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

Epic-Kitchens. This is the largest public multimodal dataset in egocentric HAR [7]. In Epic-Kitchens, which included 89,977 video clips of human-object interactions captured by 37 participants and 16 participants who also provided audio and sensor data. For the CMA-WHAR task, we used the unique 35 verb labels as activity classes and considered four modalities (i.e., video, optical flow, audio, and sensor) to simulate cross-modal WHAR scenarios in real environments. We randomly select 4 participants for each schema and treat three of them as source-domain users and the remaining one participant as a target-domain user, each of which retains data for only one schema. Finally, we obtain a 16-user 97-class 4-modality CMA-WHAR task with 34018 samples.

Multimodal-EA. This is an early multimodal dataset for self-centered HAR [32] that contains 50 minutes of video and 20 active sensor signals. We randomly split this dataset into four users, with two clients having 100 video modal samples and the other two clients having 100 sensor modal samples. Finally, we obtain a 4-user 20-class 2-modality CMA-WHAR task with 400 samples.

RealWorld. This dataset [34] provides data on the daily activities of a total of 15 users in a realistic environment and under uncontrolled conditions. We randomly select four users, three of which are used as source domain data and the remaining one as target domain data. Finally, we obtain a 15-user 8-class 1-modality HAR task with 36980 samples.

Dataset Splitting. For the above three datasets, we randomly divide the labeled user data of the source domain dataset into training and validation sets according to the ratio of 0.75:0.25, while for the unlabeled user data of the target domain dataset, we randomly divide it into adaptation and test sets according to the ratio of 0.5:0.5. All models are trained on the training set and the adaptation set, tuned on the validation set, and tested on the test set.

Evaluation Metrics. Following the existing UDA-based WHAR approach, we use accuracy as the basic evaluation metric and supplement it with a macro F1 score to observe the model's ability to balance performance on inter-classes [14].

4.2 Implementation Details

The three datasets in our experiments employ user data from up to four different modalities. To provide a fair comparison of existing methods, we use the same dimensional raw features for the different modalities of users' data and incorporate them into our model or other baseline methods to perform the CMA-WHAR task.

The overall framework of our method is implemented with Pytorch [22]. For all baselines, we use the publicly released code and compare it with our proposed model after retraining. Specifically, the activity recognition network is constructed based on the SWL-Adapt model [14] and pseudo-labels the user data in the target domain by similar tactics [18], but with the difference that the target domain data is not mixed with the source domain data for training. The hidden layer of the MLP contains one hundred nodes that constitute a universal approximator for almost any continuous function, allowing our model to fit a wide range of weighting functions [29], including those assumed in traditional sample weighting methods. For the modality discriminator, the output dimension of the single-layer perceptron ν is set to 128. The additive angular margin factor η and scale factor δ of the additive angular margin loss [8] is set to 0.5 and 72. Both our model and baseline were trained using the Adam optimizer [17], and the cosine annealing scheme was applied to the learning rate of the model to progressively reduce it from 1e-3 to 1e-4 and to make the gradient of the class-aware weighting network parameter of the same magnitude as the gradient of the parameters of the other three sub-networks. Training batch size is set to 128 and the total number of epochs is set to 300.

4.3 Comparison with State-of-the-art Methods

Baselines. We compare CASWL-Adapt with the following two categories of state-of-the-art UDA models: UDA models without the differentiation of samples: HDCNN [16], MMD [25], AdvSKM [19],

Table 2. CASWL-Adapt ablation experiments on three datasets (mean±std). "∗" indicates that CASWL-Adapt is statistically superior to the compared model according to pairwise t-test at a 95% significance level.

Model	Epic-Kitchens		Multimodal-EA		RealWorld	
	Acc.	Mac.F1	Acc.	Mac.F1	Acc.	Mac.F1
Base (DANN)	0.639±0.007*	0.621±0.009*	0.708±0.005*	0.685±0.003*	0.738±0.011*	0.712±0.007*
CASWL-M (Only modality discriminator)	0.678±0.004*	0.658±0.005*	0.731±0.009*	0.705±0.006*	0.753±0.037*	0.731±0.035*
CASWL-W (Only class-aware weight network)	0.685±0.008*	0.668±0.003*	0.737±0.003*	0.719±0.005*	0.759±0.034*	0.752±0.038*
CASWL-K (Only simple weight network)	0.675±0.007*	0.653±0.006*	0.729±0.006*	0.713±0.007*	0.756±0.035*	0.745±0.032*
CASWL-Adapt	**0.694±0.005**	**0.688±0.004**	**0.748±0.008**	**0.743±0.002**	**0.764±0.039**	**0.756±0.042**

Table 3. Comparison with state-of-the-art UDA models (mean±std) under the 4-client 97-class 4-modality CMA-WHAR task.

Model (Venue)	Epic-Kitchens	
	Acc.	Mac.F1
HDCNN (PerCom 2018)	0.534±0.005	0.501±0.003
MMD (TPAMI 2019)	0.531±0.003	0.505±0.002
AdvSKM (IJCAI 2021)	0.568±0.004	0.537±0.006
DANN (JMLR 2016)	0.574±0.008	0.551±0.004
DUA (CVPR 2022)	0.513±0.005	0.479±0.008
ETN (CVPR 2019)	0.595±0.011	0.568±0.003
TCL (AAAI 2019)	0.604±0.007	0.565±0.007
SS-UniDA (AAAI 2021)	0.615±0.009	0.585±0.006
SWL-Adapt (AAAI 2023)	0.631±0.007	0.617±0.009
CASWL-Adapt	**0.659±0.008**	**0.642±0.007**

DANN [10], and DUA [21]. UDA models with the differentiation of samples: ETN [2], TCL [31], SSUniDA [18], and SWL-Adapt [14].

Results. Overall, from Table 1, the proposed CASWL-Adapt outperforms all baselines, suggesting that our model can effectively mitigate cross-modal heterogeneity. Specifically, our model improves the Acc and Mac.F1 metrics on the Epic-Kitchens dataset by 3.1% and 4.5% in the case of multimodal and multi-activity classes, respectively. In the two-modal stochastic mixing case, our model rises by more than 2% for each indicator on the Multimodal EA dataset. Even in the simplest single-modality setting, CASWL-Adapt is still more competitive than all baselines, which demonstrates that sample-weighted learning with activity classes/tasks as additional supplementary information is better able to adapt to new users.

To further assess the effectiveness of our model in mitigating cross-modal heterogeneity, we conducted additional experiments under the more challenging 4-client 97-class 4-modality CMA-WHAR task. Specifically, we randomly select a modality for each user and ensure that no two users have the same modality between them. As seen in Table 3, CASWL-Adapt still outperformed all baselines in the strict cross-modality condition. This demonstrates that the modality discriminator is more capable of helping models to improve intra-class compactness than the traditional domain discriminator and thus mitigate data heterogeneity caused by user cross-modality. Furthermore, we can observe a clearer contrast in Figure 3, where CASWL-Adapt already shows the potential to mitigate cross-modal heterogeneity early in the training, and the class-aware weight network can capture sample inter-class differences in the cross-modality case more consistently and efficiently as the training progresses.

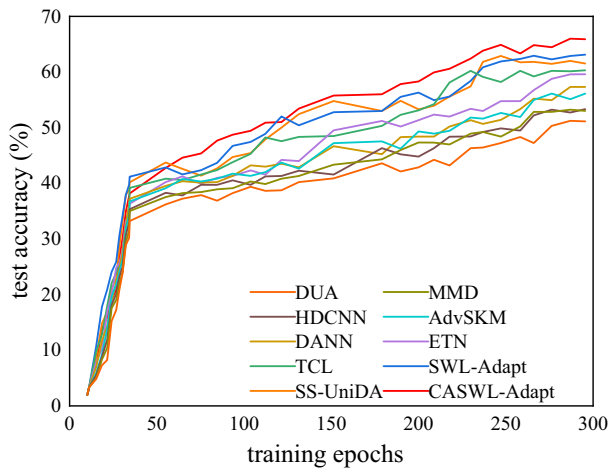

Figure 3. User-averaged test accuracy under strict cross-modality conditions with different training epochs. Our CASWL-Adapt method performs better than all other baselines.

4.4 Ablation Studies

Analysis of several variants of CASWL-Adapt. Table 2 demonstrates ablation experiments on different sub-networks within CASWL-Adapt. Specifically, the classical UDA model **DANN** [10] serves as the Base, which, compared to our CASWL-Adapt, uses the original domain discriminator and no additional sample weighting function. **CASWL-M** is obtained by removing the class-aware weighting network from CASWL-Adapt. **CASWL-W** is obtained by removing the modality discriminator from CASWL-Adapt. **CASWL-K** is obtained by removing the modality discriminator of CASWL-Adapt as well as the class/task input branch in the class-aware weight network.

From the results of ablation comparison experiments, CASWL-Adapt consistently outperforms CASWL-M and CASWL-W, which implies that the modal discriminator and the class-aware weight network, are useful for mitigating cross-modal heterogeneity. Furthermore, CASWL-K also outperforms DANN in terms of performance by simply weighting the samples without considering the additional task-level feature branches and modality discriminator. CASWL-W outperforms CASWL-K on all datasets, demonstrating that the introduction of additional class/task feature branches into the sample weight network is beneficial in helping UDA models using deep neural networks as classifiers to overcome the tendency to overfit when using cross-modal heterogeneous data, and is effective in reducing inter-class variance to increase activity recognition accuracy.

Analysis of Class-aware Sample Weights on Unimodal Setting. From Figure 4, it can be seen that in the unimodal setting,

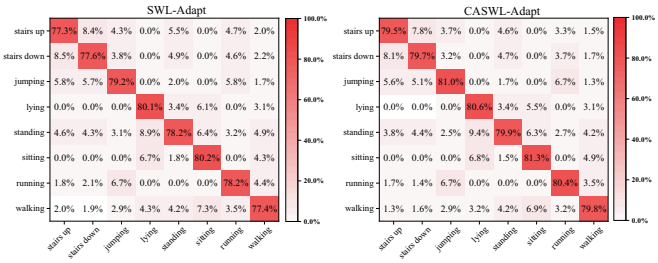

Figure 4. The confusion matrices generated by SWL-Adapt (left) and CASWL-Adapt (right). Training is performed using the RealWorld dataset.

Table 4. Modality discriminator results using different losses. # denotes models to which the sample weighting is simply applied.

Model	Epic-Kitchens	
	Acc.	Mac.F1
CE-Loss	0.682±0.004	0.664±0.005
CASWL-Adapt	**0.694±0.005**	**0.688±0.004**
CE-Loss#	0.657±0.003	0.639±0.006
CASWL-Adapt#	**0.669±0.004**	**0.651±0.007**

CASWL-Adapt degrades the modality discriminator to the same domain discriminator as SWL-Adapt [14]. With the sample-weighted strategy, CASWL-Adapt achieves higher recognition performance and mitigates the differences between user data simply by additional input class-aware features compared to SWL-Adapt using domain/classification feature differentiation.

Analysis of Modality Discrimination Loss. To verify the effect of the modality discrimination loss on the model performance, we re-evaluated the model on the Epic-Kitchens dataset after replacing it with the CE-loss. From Table 4, it can be seen that CE-Loss reduces the model performance by at least 2.00% and performs much worse than CASWL-Adapt# with modality discrimination loss when simple sample weighting [29] is applied.

4.5 Visualization of Categorical Feature Distributions

Figure 5 shows the distribution of classification features of t-SNE for RealWorld user 15. These features are the output of the dense layer of the classifier (before softmax). The target domain samples are usually close to the source domain samples of the same activity class, and CASWL-Adapt employs additional class-aware high-level task features to make multiple small similarity clusters, which suggests that our approach reduces not only the domain differences but also the interclass differences across modal activity types, making the target domain samples more discriminative.

4.6 Recall per User and per Activity Class

Figure 6 shows the recall per user and per activity class on RealWorld dataset, where the recall is calculated from the test set and averaged over 5 repetitions of the experiment. Based on the experimental setup, five users aged 30 years or older were selected as new users. Across new users, user 15, age 30, performs best on average results per activity class, benefiting from cross-modal domain correlation. Moreover, each new user performs well on the lying activity class.

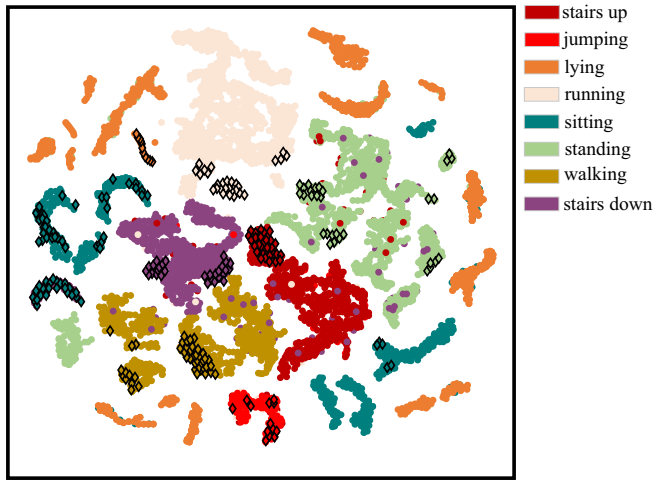

Figure 5. Visualization of categorical feature distributions using t-SNE. Different colors represent different activity classes. Dots represent source samples, and diamonds represent target samples. For clarity, diamonds are framed to distinguish target samples from source samples.

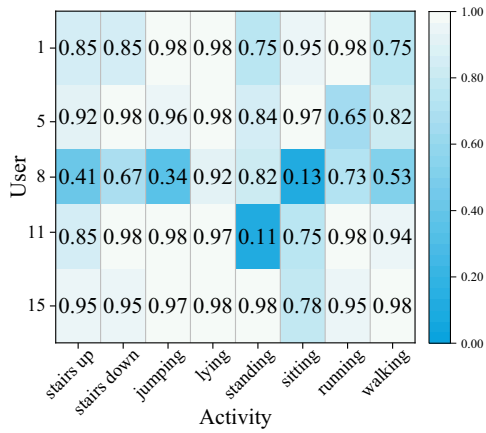

Figure 6. Recall per user and per activity class (mean).

This suggests that the model is better able to transfer knowledge by learning the different activity classes itself during main adaptation.

5 Conclusion

In this paper, we proposed a new cross-modal unsupervised domain adaptation task for cross-user wearable human activity recognition. It differs from traditional unimodal unsupervised domain adaptation in two aspects: (1) it contains different modal data from multiple different users for training and unseen modal data for testing, and (2) there is a large modal heterogeneity between the source and the target domains, and also potential domain correlation across modalities. We proposed a cross-modal unsupervised domain adaptive model named CASWL-Adapt to solve the new task by using class-aware sample weight learning and a spherical modality discriminator with modality discrimination loss. Experiments on three benchmark datasets demonstrated the strength and flexibility of CASWL-Adapt. Moreover, our method is verified to not only reduce the cross-modal heterogeneity but also exploit cross-modal domain correlation for better unsupervised domain adaptation.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Berlin Germany, 2006.

[2] Z. Cao, K. You, M. Long, J. Wang, and Q. Yang. Learning to transfer examples for partial domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2985 – 2994, 2019.

[3] J. Chen, P. Song, and C. Zhao. Multi-scale self-supervised representation learning with temporal alignment for multi-rate time series modeling. *Pattern Recognition*, 2023.

[4] K. Chen, L. Yao, D. Zhang, X. Chang, G. Long, and S. Wang. Distributionally robust semi-supervised learning for people-centric sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3321 – 3328, 2019.

[5] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys*, 54(3):1 – 40, 2021.

[6] L. Chen, R. Hu, M. Wu, and X. Zhou. Hmgan: A hierarchical multimodal generative adversarial network model for wearable human activity recognition. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1 – 27, 2023.

[7] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130:33 – 55, 2022.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[9] S. Elnaz and E. Nazerfard. Cross-subject transfer learning in human activity recognition systems using generative adversarial networks. *Neurocomputing*, 426:26 – 34, 2021.

[10] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the International Conference on International Conference on Machine Learning*, page 1180 – 1189, 2016.

[11] Z. Gao, Y. Wang, J. Chen, J. Xing, S. Patel, X. Liu, and Y. Shi. Mmtsa: Multi-modal temporal segment attention network for efficient human activity recognition. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, pages 1 – 26, 2023.

[12] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5149 – 5169, 2022.

[13] A. Hosseini, D. Zamanzadeh, L. Valencia, R. Habre, A. A. T. Bui, and M. Sarrafzadeh. Domain adaptation in children activity recognition. In *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1725 – 1728, 2019.

[14] R. Hu, L. Chen, S. Miao, and X. Tang. Swl-adapt: An unsupervised domain adaptation model with sample weight learning for cross-user wearable human activity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6012–6020, 2023.

[15] S. Kasim and J. W.Sheppard. Cross-domain similarity in domain adaptation for human activity recognition. In *International Joint Conference on Neural Networks*, pages 1 – 8, 2023.

[16] M. A. A. H. Khan, N. Roy, and A. Misra. Scaling human activity recognition via deep learning-based domain adaptation. In *IEEE International Conference on Pervasive Computing and Communications*, pages 1 – 9, 2018.

[17] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference for Learning Representations*, 2015.

[18] O. Lifshitz and L. Wolf. Sample selection for universal domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8592–8600, 2021.

[19] Q. Liu and H. Xue. Adversarial spectral kernel matching for unsupervised time series domain adaptation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 2744–2750, 2021.

[20] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2208–2217, 2020.

[21] M. J. Mirza, J. Micorek, H. Possegger, and H. Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14765–14775, 2022.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

[23] Prabono, A. Ghora, B. N. Yahya, and S.-L. Lee. Hybrid domain adaptation for sensor-based human activity recognition in a heterogeneous setup with feature commonalities. *Pattern Analysis and Applications*, 24(4):1501 – 1511, 2021.

[24] A. G. Prabono, B. N. Yahya, and S.-L. Lee. Multiple-instance domain adaptation for cost-effective sensor-based human activity recognition. *Future Generation Computer Systems*, 133:114 – 123, 2022.

[25] A. Rozantsev, M. Salzmann, and P. Fua. Beyond sharing weights for deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):801 – 814, 2019.

[26] K. Saito, K. Watanabe, Y. Ushiku, and T. Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3723 – 3732, 2018.

[27] A. R. Sanabria, F. Zambonelli, S. Dobson, and J. Ye. Contrasgan: Unsupervised domain adaptation in human activity recognition via adversarial and contrastive learning. *Pervasive and Mobile Computing*, 78, 2021.

[28] S. Seo, Y. Suh, D. Kim, G. Kim, J. Han, and B. Han. Learning to optimize domain specific normalization for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference*, pages 68 – 83, 2020.

[29] J. Shu, Q. Xie, L. Yi, Q. Zhao, S. Zhou, Z. Xu, and D. Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in neural information processing systems*, 2019.

[30] J. Shu, X. Yuan, D. Meng, and Z. Xu. Cmw-net: Learning a class-aware sample weighting mapping for robust deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11521 – 11539, 2023.

[31] Y. Shu, Z. Cao, M. Long, and J. Wang. Transferable curriculum for weakly-supervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4951 – 4958, 2019.

[32] S. Song, N.-M. Cheung, V. Chandrasekhar, B. Mandal, and J. Liri. Egocentric activity recognition with multimodal fisher vector. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2717–2721, 2016.

[33] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang, and J. Liu. Human action recognition from various data modalities: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3200 – 3225, 2023.

[34] T. Sztyler and H. Stuckenschmidt. On-body localization of wearable devices: An investigation of position-aware activity recognition. In *IEEE International Conference on Pervasive Computing and Communications*, pages 1–9, 2016.

[35] X. Wang, Y. Xu, J. Yang, and K. Mao. Calibrating class weights with multi-modal information for partial video domain adaptation. In *Proceedings of the ACM International Conference on Multimedia*, page 3945 – 3954, 2022.

[36] G. Wilson and D. J. Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology*, 11(5):1 – 46, 2020.

[37] G. Wilson, J. R. Doppa, and D. J.Cook. Multi-source deep domain adaptation with weak supervision for time series sensor data. In *In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1768 – 1778, 2020.

[38] Q. Xu, X. Wei, R. Bai, S. Li, and Z. Meng. Integration of deep adaptation transfer learning and online sequential extreme learning machine for cross-person and cross-position activity recognition. *Expert Systems with Applications*, 212:801 – 814, 2023.

[39] X. Yang, B. Xiong, Y. Huang, and C. Xu. Cross-modal federated human activity recognition via modality-agnostic and modality-specifc representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3063 – 3071, 2022.

[40] X. Yu, Z. Cao, Z. Wu, C. Song, and Z. Xu. Sample intercorrelation-based multidomain fusion network for aquatic human activity recognition using millimeter-wave radar. *IEEE Geoscience and Remote Sensing Letters*, 20, 2023.

[41] Y. Zhang, F. Liu, Z. Fang, B. Yuan, G. Zhang, and J. Lu. Learning from a complementary-label source domain: Theory and algorithms. *IEEE Transactions on Neural Networks and Learning Systems*, 33:7667 – 7681, 2022.

[42] Z. Zhou, Y. Zhang, X. Yu, P. Yang, X.-Y. Li, J. Zhao, and H. Zhou. Xhar: Deep domain adaptation for human activity recognition with smart devices. In *17th Annual IEEE International Conference on Sensing, Communication, and Networking*, pages 1 – 9, 2020.