Domain Generalization for Activity Recognition via Adaptive Feature Fusion

XIN QIN, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Inst. of Computing Tech., CAS, University of Chinese Academy of Sciences, China
JINDONG WANG*, Microsoft Research Asia, China
YIQIANG CHEN*, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Inst. of Computing Tech., CAS, University of Chinese Academy of Sciences, Pengcheng Laboratory, Shenzhen, China
WANG LU, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Inst. of Computing Tech., CAS, University of Chinese Academy of Sciences, China
XINLONG JIANG, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Inst. of Computing Tech., CAS, China

Human activity recognition requires the efforts to build a generalizable model using the training datasets with the hope to achieve good performance in test datasets. However, in real applications, the training and testing datasets may have totally different distributions due to various reasons such as different body shapes, acting styles, and habits, damaging the model's generalization performance. While such a distribution gap can be reduced by existing domain adaptation approaches, they typically assume that the test data can be accessed in the training stage, which is not realistic. In this paper, we consider a more practical and challenging scenario: domain-generalized activity recognition (DGAR) where the test dataset *cannot* be accessed during training. To this end, we propose *Adaptive Feature Fusion for Activity Recognition (AFFAR)*, a domain generalization approach that learns to fuse the domain-invariant and domain-specific representations to improve the model's generalization performance. AFFAR takes the best of both worlds where domain-invariant representations enhance the transferability across domains and domain-specific representations leverage the model discrimination power from each domain. Extensive experiments on three public HAR datasets show its effectiveness. Furthermore, we apply AFFAR to a real application, i.e., the diagnosis of Children's Attention Deficit Hyperactivity Disorder (ADHD), which also demonstrates the superiority of our approach.

CCS Concepts: • **Human-centered computing** → **Ubiquitous computing**; • **Computing methodologies** → **Transfer learning**.

Additional Key Words and Phrases: Human Activity Recognition, Domain Generalization, Transfer Learning

*Correspondence to: J. Wang and Y. Chen.

Authors' addresses: Xin Qin, qinxin18b@ict.ac.cn, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Inst. of Computing Tech., CAS, University of Chinese Academy of Sciences, Beijing, China; Jindong Wang, jindong.wang@microsoft.com, Microsoft Research Asia, Beijing, China; Yiqiang Chen, yqchen@ict.ac.cn, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Inst. of Computing Tech., CAS, University of Chinese Academy of Sciences, Pengcheng Laboratory, Shenzhen, China; Wang Lu, luwang@ict.ac.cn, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Inst. of Computing Tech., CAS, University of Chinese Academy of Sciences, Beijing, China; Xinlong Jiang, jiangxinlong@ict.ac.cn, Beijing Key Lab. of Mobile Computing and Pervasive Devices, Inst. of Computing Tech., CAS, Beijing, China.

2157-6904/2021/5-ART000 $15.00
https://doi.org/00.0000/0000000.0000000

ACM Reference Format:
Xin Qin, Jindong Wang, Yiqiang Chen, Wang Lu, and Xinlong Jiang. 2021. Domain Generalization for Activity Recognition via Adaptive Feature Fusion. *ACM Trans. Intell. Syst. Technol.* 0, 0, Article 000 (May 2021), 21 pages. https://doi.org/00.0000/0000000.0000000

1 INTRODUCTION

Human activity recognition (HAR) is an active research topic in ubiquitous computing. HAR aims at recognizing people's activities by building machine learning models on the activity data. HAR has been wildly applied in smart-home [15], fatigue detection [47], fall detection for the elder [32], attention deficit hyperactivity disorder (ADHD) [11], and other fields. Therefore, accurate HAR is of vital importance in real-world applications. Many machine learning methods have been adopted to improve the performance of HAR, such as Support Vector Machines (SVM), K-Nearest Neighbor (KNN), Random Forest (RF), and the deep learning models including Convolutional Neural Networks (CNN) [25], Long-short Term Memory (LSTM) [21] and others [53].

Despite the great success in the past, one critical challenge is the *generalization* ability of the HAR models, i.e., the performance of applying the trained models to a new, *unseen* dataset. In real applications, the sensor signals are easily influenced by the diverse personalities of end-users such as acting styles, habits, or different body shapes. When testing on a new end-user whose activity data are never seen in the training set, the performance of the model is likely to drop. For instance, Figure 1 shows the sensor readings of two users from DSADS dataset [3] collected using the same device, where the distributions of sensor readings are different, i.e., $P(\mathcal{D}^1) \neq P(\mathcal{D}^2)$. When deployed to an unseen test user $\mathcal{D}^{te}$ whose distribution is different from the training set, the performance of the activity recognition model will be likely to drop. This is due to the domain shift caused by the non-IID (independently and identically distributed) distributions between training and testing datasets [48].

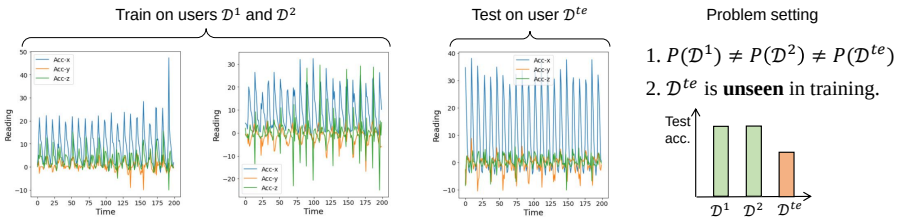


Fig. 1. The distributions of accelerometer readings vary among different users. When the test dataset has different distributions and cannot be accessed during training, the performance of existing methods will drop while our method can reach better performance.

Such a domain shift issue can be mitigated by transfer learning and domain adaptation techniques [10, 36], which have been applied to HAR over the years. Transfer learning first pre-trains a model from the source dataset and then fine-tunes it on the new test data. Domain adaptation performs instance reweighting or feature transformation between the training and testing datasets to learn domain-invariant representations where their distribution gap can be minimized [7, 10, 27, 39, 43, 54]. Unfortunately, both transfer learning and domain adaptation require access to target domain training data, which is often not realistic in real applications as we aim to achieve "Train once, deploy everywhere". In real applications, it is hard to obtain the test data [55]. For example, it is impractical to collect each patient's data for medical healthcare, and to collect data on a variety of fall poses for fall detection in advance.

Domain generalization (DG) [55] is an emerging research topic in recent years. DG focuses on utilizing the knowledge from several different domains to build a model that can generalize well to unknown domains. Many works have been done and make a good performance in the computer vision field. Unfortunately, we cannot directly apply existing DG methods to our problem due to the characteristics of wearable-sensor-based activity data. To the best of our knowledge, few existing works focus on the domain generalization problems in HAR and we refer to this as domain-generalized HAR (DGAR).

In this paper, we propose **Adaptive Feature Fusion for Activity Recognition (AFFAR)** to improve the generalization ability of HAR models (refer to Figure 2). The key of AFFAR is to learn both domain-specific representation and domain-invariant representation and fuse them dynamically in a unified deep neural network. Specifically, domain-invariant representation learning is to capture the general and transferable knowledge from the training domains, and domain-specific representation learning is to learn the specific characteristics of each domain to preserve the diversity of features to improve generalization ability. Our key assumption is that although we cannot get access to the test data and the sensor readings from different persons are different, they still share some similarities that can be utilized to learn domain-invariant representations. Therefore, we can learn transferable knowledge while preserving their diversities for generalization. Our method can be optimized in an end-to-end neural network. We show the superiority of AFFAR by experimenting on both public HAR datasets and a real application to the diagnosis of attention deficit hyperactivity disorder (ADHD). Experiments demonstrate that our method significantly outperforms the comparison methods.

The main contributions of this paper are four-fold:

(1) We propose and study a more practical and challenging problem scenario: domain-generalized activity recognition (DGAR), for robust and generalized activity recognition. We thoroughly analyze the reason for this problem, indicating a new research direction.
(2) To solve the DGAR problem, we propose a novel algorithm: Adaptive Feature Fusion for Activity Recognition (AFFAR), to learn both domain-invariant and domain-specific deep representations to enhance the generalization capability of the model to unseen datasets.
(3) We evaluate AFFAR on three public HAR datasets. Experiments on cross-person activity recognition demonstrate that the proposed AFFAR can significantly outperform the comparison methods.
(4) Finally, we apply our AFFAR algorithm to a real-world ADHD problem where it also achieves the best performance.

2 RELATED WORK

2.1 Human Activity Recognition

Human activity recognition (HAR) [40] aims at recognizing the activities of people by training machine learning models on the data collected during performing some specific activities. In HAR, the wearable sensor-based human activity recognition has occupied an important position as it is superior in pervasiveness, computational consumption, and privacy preservation with wearable sensors as interface [9]. So in this paper, we mainly focus on wearable sensor-based activity recognition. Over these years, many efforts have been done to achieve accurate and robust activity recognition including traditional machine learning methods[28] combined with feature extraction, and deep learning methods [8, 53]. However, the conventional methods usually focus on the i.i.d. data situation, i.e. the training data set and testing data set follow the same distribution, this may result in performance degradation when faced with various tasks in real-life applications.

2.2 Transfer Learning and Domain Adaptation

In sensor-based HAR, domain shift is a common and must be solved problem due to the variation of devices, locations, personalities, and so on. During the data collection, any change in these factors may result in distribution divergence. Large distribution divergence results in the performance degradation of the trained model. Transfer learning, as a representative method of machine learning, is an effective paradigm to solve the domain shift problems, which makes it possible to reuse existing knowledge. The purpose of transfer learning is to apply the knowledge learned in the existing domains to related but different accessible domains during the training process, to improve the ability to solve new tasks [36, 57].

The commonly used approach of transfer learning is pre-training and fine-tuning, i.e., pre-training on the source dataset to get a pre-trained model, and then fine-tuning this model on the new target dataset. Such a simple and effective approach has been successfully applied to computer vision [46, 60], natural language processing [12], and speech recognition [24, 49]. In order to tackle the domain shift challenge, transfer learning has become an effective method and applied in HAR [7, 10, 27, 39, 43]. Ma M et.al [33] proposed a twin stream network architecture and jointly fine-tuned the two networks to recognize objects, actions, and activities. Wang H B et.al [52] verified that fine-tuning and regular constraints can increase the training efficiency, and fine-tuning is valid in practical HAR applications.

In the scope of transfer learning, domain adaptation (DA) is a major technique and has attracted much attention from researchers in recent years. Domain adaptation aims at improving the performance on the given less annotated or no annotated target domain by exploiting the source domains. Domain adaptation is a popular topic and has been applied in HAR to solve the domain shift problem [7]. Khan et al. [27] proposed a feature-based model HDCNN, which adapts the source and target features after every convolutional and fully connected layer. Chang et al. [7] made a comparison of adaptation techniques to give a guideline on applying unsupervised domain adaptation algorithms to cross-position HAR problems.

However, both transfer learning and domain adaptation assume the target domain can be accessed in the model training process. While in real HAR applications, the target tasks are novel and various and can not be accessed during the training. This requires the trained model has strong generalization capability so that it can perform well in unknown fields, which goes beyond the scope of conventional transfer learning and domain adaptation problem settings. Specifically, the multi-source domain adaptation (MSDA) also has several source domains for training, but its goal is to adapt the trained model to the target domain, which is accessible during training.

2.3 Domain Generalization

Domain generalization (DG) is an emerging topic and is attracting increasing attention in recent years. The goal of DG is to learn a robust and well-generalized prediction function on several given source domains to obtain the minimum prediction error on any possible unknown domain [55]. The most striking difference between DA and DG is that whether the target domain can be accessed during the training, the target data can only be used for the model test in DG. Thus, DG is more suitable for the situation of various unknown target HAR tasks in real life. DG has been widely applied in the computer vision field and many DG methods have been proposed and evaluated on image datasets. The DG methods can be mainly divided into three branches, i.e. data manipulation [38, 45], which works on the input data to assist the general representation learning; representation learning [17, 18, 30], which aims at learning domain-invariant representation or disentangling the features; and learning strategy [5, 23, 29, 42], which exploits the general learning strategy such as meta-learning and ensemble learning for generalization. Yan et al. [58] solved the

DG problem from the perspective of data generation by linear interpolation between instances and their labels. However, unlike image data, it is hard to intuitively assess the semantics and diversity of sensor data. Although some works assess sensor data by training a post-hoc classification or prediction model or using other techniques, these assessments are still less intuitive and explainable to some extent [31, 59]. Data generation methods may be not straightly applicable to sensor-based HAR since there is still a lack of the intuitive quantitative assessment of quality for generated sensor data. Li et al. [29] proposed a model agnostic training procedure by leveraging the meta-learning for DG, i.e. MLDG, while this kind of method may not straightly applicable due to the high dependence of sensor data. Many methods follow distribution alignment in DA to minimize the distribution discrepancy between domains by adversarial training [19], Wasserstein distance [62] etc. to learn domain-invariant representations, and the spirit is followed in DG. In the other aspect, some ensemble-learning-based methods focus on the domain-specific, such as domain-specific neural networks [34, 56], domain-specific batch normalization [44], weight averaging [6] etc., more details can be found in [55]. However, few works both consider domain-invariant and domain-specific representations, especially for HAR applications.

Last, it is worth noting that domain generalization is not Leave-One-Out-Cross-Validation (LOOCV) in traditional machine learning. For domain generalization, we sequentially leave one domain for the final test to construct several domain generalization tasks, and these tasks are independent of each other. LOOCV is mainly used for selecting models where all domains are used for validation in turn and it also has an inaccessible test dataset for the final testing. Therefore, LOOCV can be seen as a typical model selection method that can also be used in DG.

3 OUR METHOD: ADAPTIVE FEATURE FUSION

In this section, we present our Adaptive Feature Fusion method for activity recognition in detail.

3.1 Problem Definition: Domain-generalized Activity Recognition

In a typical human activity recognition (HAR) problem, we are given a training dataset $\mathcal{D}^{tr} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, where $\mathbf{x} \in \mathbb{R}^d$ denotes its d-dimensional features and $y \in \{1, 2, \cdots, C\}$ denotes its corresponding activity categories, such as walking or running. n denotes the total number of samples. The goal is to build a machine learning model $h : \mathbf{x} \mapsto y$ such that it can accurately recognize the activities in the training data, i.e., achieving the minimum training error:

$$h^* = \arg\min_h \frac{1}{n} \sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i), \tag{1}$$

where h^* denotes the optimal model and $\ell(\cdot, \cdot)$ is the loss function such as cross-entropy loss.

However, achieving the minimum error on the training data $\mathcal{D}^{tr}$ does not necessarily guarantee optimal performance when we apply the model to the **unseen** test data $\mathcal{D}^{te}$. For instance, a well-trained HAR model can perform poorly when deployed to recognize different persons' activities with different body shapes or activity styles. Moreover, we can never collect all the possible training data to build a generalized HAR model. While transfer learning and domain adaptation [10, 36, 39, 54] are popular to perform cross-domain learning, they can not be used in our problem since they require the availability of the test domain.

We aim to solve this practical and challenging problem, which we refer to as **Domain-generalized Activity Recognition**, or **DGAR**. Here, "domain" is a general notion of "dataset", i.e., a dataset is a domain, or it can be split into several domains [55]. In DGAR, we assume there are several training (source) domains available, i.e., there are K different but related training domains $\mathcal{D}^{tr} = \left\{\mathcal{D}^1, \mathcal{D}^2, \cdots, \mathcal{D}^K\right\}$ available. $\mathcal{D}^k = \left\{(\mathbf{x}_i^k, y_i^k)\right\}_{i=1}^{n_k}$ denotes the k_{th} training domain with n_k samples.

Our goal is to learn a generalized model h on the K training domains such that it can achieve minimum error on the *unseen* test domain $\mathcal{D}^{te} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{te}}$. We often assume that all domains share the same kinds of sensors and activities, i.e. the feature space and the label space are the same: $\mathcal{X}^{tr} = \mathcal{X}^{te}$ and $\mathcal{Y}^{tr} = \mathcal{Y}^{te}$. In real applications, different domains tend to have *different* probability distributions, i.e. $P^i(\mathbf{x}) \neq P^j(\mathbf{x}) \neq P^{te}(\mathbf{x}), 1 \leq i \leq j \leq K$. Since the probability distribution of the test dataset is different from the training domains, DGAR is a practical setting to evaluate the generalization ability of activity recognition algorithms.

3.2 Motivation and Main Idea

Domain generalization (DG) [4, 55] is the general learning setting of DGAR. Over the past few years, domain generalization has attracted the increasing attention of researchers, and a large number of DG methods have been proposed [55]. Most of these methods are developed for general-purpose learning and they are often evaluated on image classification tasks, with data augmentation [5], domain-invariant representation learning [35], or meta-learning methods [29]. Intuitively, it is natural to ask: *can we directly apply these existing DG methods to our DGAR problem?* The short answer is yes but no. "Yes" means we can always do that, "no" means this ignores the characteristics of activity recognition and we can do it better. The reasons are as follows.

First, to achieve strong generalization capability, some of them focus on manipulation of data including data augmentation and data generation which are specific to the image domain [5]. However, such methods may not be straightly applicable to sensor-based HAR due to the lack of a quantitative assessment of quality for generated sensor data. Second, some focus on representation learning, which aims to learn a better feature representation for better generalization. However, activity recognition is a special area where we should not only care about domain-invariant features but also domain-specific features to capture the individually-specific features from each domain, which can preserve the diversity of different persons. Third, although meta-learning-based DG methods could be used for our problem, our empirical experiments (ref. Table 2) indicate that its performance is even worse than the empirical risk minimization method. The reason may be the second-order gradient optimization can have mode collapse for the special activity data. To sum up, we need to develop special algorithms for this DGAR problem.

In this paper, we proposed a novel **Adaptive Feature Fusion** method for domain-generalized Activity Recognition, abbreviated as **AFFAR**. Our key assumption is that although we cannot get access to the test data in training, we can still learn to represent the test data using the aggregation of existing training data. This is reasonable since different persons may generate different sensor readings while performing the same activities, but they could share some similarities such as body shapes and activity styles. Thus, we can learn to represent each data as the weighted aggregation of existing training domains. Meanwhile, since the activity data from different domains have different probability distributions, we also need to learn domain-invariant features to regularize the model to facilitate knowledge transfer.

The core of AFFAR is to learn both domain-invariant and domain-specific feature representations. Specifically, domain-invariant representation learning is to capture the general and transferable knowledge from the training domains, and domain-specific representation learning is to learn the specific characteristics of each source domain to enhance the generalization ability. We depict the overall learning process of AFFAR in Figure 2.

AFFAR conceptually consists of four modules: feature extraction module (purple blocks), activity classification module (blue blocks), domain-specific representation learning module (yellow blocks), and domain-invariant learning module (green blocks). The feature extraction module is used to extract features from the raw sensor data. In this paper, we leverage two convolutional neural network (CNN) layers along with max-pooling operations to extract features. These layers are

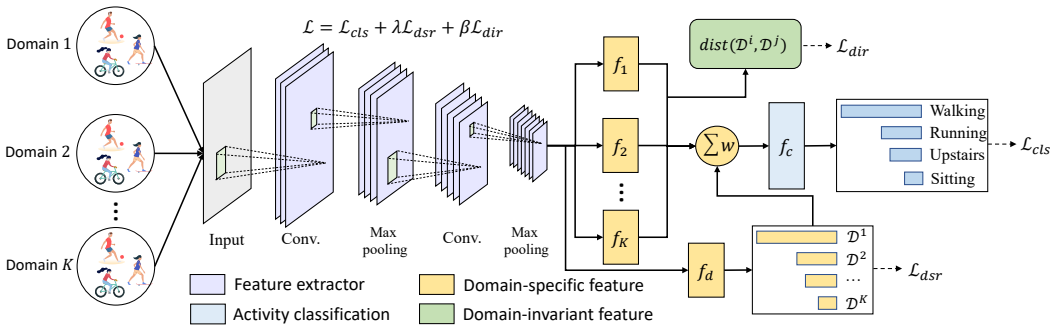


Fig. 2. Illustration of the AFFAR framework.

shared by all training domains to reduce parameter amounts. The domain-specific representation learning module is used to learn domain-specific features, thus this is not shared, but specific for each domain. We implement it by adding K fully connected (FC) layers for each domain after feature extraction. To aggregate the specific information of each domain, we design a weighting function. The domain-invariant representation learning module is used to reduce the distribution discrepancy between each domain $\mathcal{D}^i$ and domain $\mathcal{D}^j$ to learn domain-invariant features. Finally, the activity classification module is an FC layer. Since the classification uses the fusion of domain-specific and domain-invariant features, we call our method adaptive feature fusion.

The learning objective of AFFAR can be formulated as:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{dsr} + \beta\mathcal{L}_{dir}, \tag{2}$$

where $\mathcal{L}_{cls}$ is the classification loss, $\mathcal{L}_{dsr}$ is the loss of domain-specific representation learning, $\mathcal{L}_{dir}$ is the loss of domain-invariant representation learning. λ and β are the tradeoff hyperparameters. For classification, we take cross entropy as the classification loss:

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log P\left(y_i \mid x_i\right), \tag{3}$$

where $N = \sum_{k=1}^{K} n_k$ is the amount of the samples from all the training domains.

In the next sections, we will elaborate on the details of domain-specific and domain-invariant representation learning modules.

3.3 Domain-specific Representation Learning

The domain-specific representation learning module aims to learn domain-specific features and then aggregate them for unified feature representations by fusing the features from multiple sources for the unseen target feature. More formally, given a new test data $\mathbf{x}$, its feature $\mathbf{z}$ is formulated as:

$$\mathbf{z} = \sum_{k=1}^{K} w_k f_k(f_e(\mathbf{x})), \text{ where } w_k > 0 \text{ and } \sum_{k=1}^{K} w_k = 1, \tag{4}$$

where w_k is the weight on domain $\mathcal{D}^k$, indicating the similarity between the data on domain $\mathcal{D}^k$ and the true data. f_k is the feature learning function of domain $\mathcal{D}^k$ and f_e is the shared feature extraction function, i.e., CNNs.

This process can be viewed as a certain type of ensemble learning [13], where each specific base learner is trained to ensemble a stable model which can perform well in all aspects. In our problem, the base learner is composed of two parts: the shared CNN feature extractor f_e and the specific

feature extraction layer f_k for each domain $\mathcal{D}^k$. This split is motivated by existing research on the transfer learning ability of deep networks [60] that the lower layers tend to learn low-level and general features while the higher layers tend to learn domain-specific features.

To learn the weights w_k for each domain $\mathcal{D}^k$, we build a domain classifier $f_d : \mathbf{x} \mapsto \mathbb{R}^K$ that takes as input the features after CNN layers, and then use softmax to satisfy Equation 4. At training time, the domain label $d_k \equiv k$ for each sample is known a priori, thus, the domain-specific loss for each domain k can be computed as:

$$\mathcal{L}_{dsr}^k = \frac{1}{n_k} \sum_{i=1}^{n_k} \ell(f_d(f_e(\mathbf{x}_i)), d_k). \tag{5}$$

Then, we can get the total domain-specific loss by averaging the losses on all domains as $\mathcal{L}_{dsr} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{L}_{dsr}^k$.

When testing on the target data, we cannot access the domain label of the unseen target data. After the network extracts the features, the output of the domain weight branch is used as the weight to fuse the target features extracted by each domain-specific branch. It can be understood that the learned domain branch network adaptively fuses the domain-specific features extracted from each source-specific branch to construct the feature representation of the target domain. In this way, AFFAR can adaptively learn feature representation of any unseen domains.

3.4 Domain-Invariant Representation Learning

While domain-specific feature learning encourages the model to learn specific information for each domain, the feature distribution gap could also be enlarged due to their diverse representations. Thus, to enhance the generalization capability, we further design a domain-invariant representation learning module to seek balance with domain-specific representation learning.

Recall that features in the lower level of deep neural networks focus on learning common and low-level features, while the higher layers focus more on specific tasks [60]. Thus, we make feature adaptation to reduce the distribution discrepancy between source domains in the domain-specific layers. In the domain adaptation field, the strategy is to reduce the distribution discrepancy between the source domain and the target domain so that the model can be robust on the target. Under domain generalization scenarios, we cannot access the target domain data during the training process, so it is of vital importance to learn domain-invariant feature representation so that any unseen target can be represented. To enable feature adaptation in DGAR, we turn to reducing the distribution divergence between each domain pair $\mathcal{D}^i$ and $\mathcal{D}^j$, i.e., to minimize $dist(\mathcal{D}^i, \mathcal{D}^j)$ where $dist(\cdot, \cdot)$ is a distribution distance measurement.

Specifically, we adopt the widely used distance metric Maximum Mean Discrepancy (MMD) [20] to help reduce the distribution divergence. MMD embeds distributions in Reproducing Kernel Hilbert Space (RKHS) and calculates the distance between these embeddings as the test statistic. Thus, it is often adopted to justify whether two distributions are the same or used to measure how similar two distributions are. The MMD loss between domains $\mathcal{D}^i$ and $\mathcal{D}^j$ is formulated as:

$$\mathcal{L}_{dir}^{ij} = \left\| \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}^i} \phi\left(\mathbf{x}\right) - \frac{1}{n_j} \sum_{\mathbf{x} \in \mathcal{D}^j} \phi\left(\mathbf{x}\right) \right\|_{\mathcal{H}}^2, \tag{6}$$

where i and j are indexes of domains, $\phi(\cdot)$ is the feature map which maps the original instances into the RKHS $\mathcal{H}$. Then, the total loss for domain-invariant learning can be computed by taking average on all possible domain pairs $\mathcal{L}_{dir} = \frac{2}{K(K-1)} \sum_{i,j} \mathcal{L}_{dir}^{ij}$.

Although we adopt MMD as a metric for distribution divergence in this paper, AFFAR is a general approach that can embed other metrics instead of MMD.

3.5 Training and Inference

As for training, after loading the training data, the feature extractor extracts the lower-level features and then inputs them to each domain-specific branch. Each branch learns the higher-level domain-specific representations. Meanwhile, the domain classifier takes the lower-level features as input and outputs the softmax weights to fuse the domain-specific features. At the same time, it makes the in-between-source domain adaptation to learn domain-invariant representation. Then, it takes the fused features as input of the final activity classifier to make activity classification.

As for inference, we fix the model parameters and learn domain-specific and domain-invariant representations for the target. Different from the training, without prior knowledge of the domain label of target data, the output of the domain classifier can be regarded as the similarity between the target and each source domain-specific branch.

The complete learning process of AFFAR is summarized in Algorithm 1.

Algorithm 1 AFFAR for domain-generalized activity recognition

Input: K training domains $\mathcal{D}_s^1, \cdots, \mathcal{D}_s^K$, and λ, β.
Output: Classification results on test domain.

```
Randomly initialize the model parameters θ;
while not converge do
    Sample a mini-batch B = {B^1, ··· , B^K} from K domains;
    Extract the lower-level features f_e(x) by the feature extractor;
    Extract the domain-specific features f_k(f_e(x)) by K domain-specific FC layers;
    Calculate the domain-specific loss L_dsr and output the weight for each source branch;
    Calculate the domain-invariant loss L_dir.
    Fuse the domain-specific features with weight according to Eq. (4);
    Calculate the total loss of AFFAR according to Eq. (2);
    Update the model parameter θ using SGD.
end while
Make inference on the target HAR data.
return Classification results on target HAR data.
```

3.6 Discussions

The proposed AFFAR learns domain-specific and domain-invariant representations in a unified framework to seek their balance in feature learning, which can take other distribution discrepancies as the domain-invariant learning loss. We show two possible losses: the domain-adversarial neural networks (DANN) [17] and the CORelation ALignment (CORAL) loss [50]. DANN introduced an adversarial training objective where it used a min-max optimization to maximize the loss of domain discriminator and minimize the loss of both feature extractor and classification. However, DANN only aims at learning domain-invariant representations, which is good for domain adaptation tasks (that explains why DANN is the base model for modern domain adaptation models). DANN ignores the specific features for each domain that is useful for domain generalization tasks, making it less favorable for our DGAR problem. We will empirically show this argument in later experiments (ref. Section 4.4.2). On the other hand, AFFAR is a general framework for DGAR tasks where we can also employ the domain discriminator loss of DANN to replace MMD:

$$\mathcal{L}_{dir}^{dann} = \mathbb{E}_{1\le i\ne j\le K}\mathbb{E}_{\mathbf{x}^i\in\mathcal{D}^i,\mathbf{x}^j\in\mathcal{D}^j}\log[D(f_i(f_e(\mathbf{x}^i)))] + \log[1 - D(f_j(f_e(\mathbf{x}^j)))], \tag{7}$$

where $\mathbb{E}$ denotes expectation operation and D is a domain discriminator (typically a two-layer feed-forward network). Thus, DANN for domain generalization requires to build $\frac{K(K-1)}{2}$ domain discriminators, which is not efficient. We can also replace MMD with CORAL loss as:

$$\mathcal{L}_{dir}^{coral} = \frac{1}{4d^2}||C^i - C^j||_F^2, \tag{8}$$

where d denotes the feature dimension, C^i, C^j denotes the covariance matrices for two domains, and $||\cdot||_F$ denotes the Frobenius norm. In later experiments (Section 4.4.2), we will show that our AFFAR can also achieve competitive performance with these two losses.

3.7 Theoretical Insights

Finally, we show that our algorithm is theoretically-motivated using the theory proposed in [1].

Theorem 1 (Risk upper bound on unseen domain [1]). *Let $\gamma = d_{\mathcal{H}}(\mathcal{D}^{te}, \bar{\mathcal{D}}^{te})$ denote the $\mathcal{H}$-divergence between target domain and its nearest neighbor in source domain convex hull, then, the risk on unseen domain $\mathcal{D}^{te}$ of hypothesis h is upper-bounded by the weighted risk on source set S:*

$$R_{te}[h] \leq \sum_{i=1}^{N_S} \pi_i R_S^i[h] + \gamma + \epsilon + \min\left\{\mathbb{E}_{\bar{\mathcal{D}}^{te}}\left[\left|f_{S_\pi} - f^{te}\right|\right], \mathbb{E}_{\mathcal{D}^{te}}\left[\left|f^{te} - f_{S_\pi}\right|\right]\right\}, \tag{9}$$

where ϵ is the largest distribution divergence between unseen target domain and any source domain and $\min\left\{\mathbb{E}_{\bar{\mathcal{D}}^{te}}\left[\left|f_{S_\pi} - f^{te}\right|\right], \mathbb{E}_{\mathcal{D}^{te}}\left[\left|f^{te} - f_{S_\pi}\right|\right]\right\}$ denotes the difference between labeling functions.

In our problem, the categories between training and testing are the same, the main distribution difference between training and testing data is the activity patterns (i.e., $P(\mathbf{x})$). So it is close to the covariate shift assumption: the labeling function error ($\min\{\cdot,\cdot\}$) and γ are both relatively small [1]. In this way, the risk on unseen domain is bounded by two terms: the weighted source risk $\sum_{i=1}^{N_S} \pi_i R_S^i[h]$ and the source-target distribution divergence ϵ. Obviously, our domain-specific learning module (Eq. (4)) corresponds to minimizing the weighted source risk and the domain-invariant learning module (Eq. (6)) minimizes the risk ϵ. Thus, our algorithm can also be interpreted from the theory. Additionally, we also provide a visualization study to help better analyze the algorithm in Section 4.4.

4 EXPERIMENTAL EVALUATION

In this section, we evaluate the performance of the proposed AFFAR approach via extensive experiments on domain-generalized activity recognition.

4.1 Datasets and Preprocessing

We adopt three large public activity datasets as summarized in Table 1. In the following, we briefly introduce their basic information, and detailed descriptions are in their original papers.

Table 1. Statistical information of three public activity recognition datasets

Dataset	Subject	Activity	Sample	Body Position	Sampling rate	Training : Test size
DSADS	8	19	1.14M	5	25Hz	~15:1
USC-HAD	14	12	2.81M	1	~100Hz	~16:1
PAMAP2	9	18	2.84M	3	100Hz	~15:1

DSADS. UCI Daily and Sports Data Set [3] collects 19 activities through 8 subjects (four males and four females between the ages of 20 and 30) wearing body-worn sensor units including triaxial accelerometer, triaxial gyroscope, and triaxial magnetometer on 5 body parts: torso, right arm, left arm, right leg and left leg. To construct the domain-generalized activity recognition scenario, we divide the 8 subjects into 4 groups where each group consists of 2 different subjects, and data from each group is regarded as a domain, leading to 4 domains in total.

USC-HAD. USC Human Activity Dataset [61] consists of data collected from 14 subjects (7 males and 7 females) performing 12 activities. A motion mode is equipped at the front right hip of subjects to capture triaxial accelerometer and triaxial gyroscope sensor readings. In order to construct a domain-generalized activity recognition scenario, we divide the 14 subjects into 5 groups where each of the first four groups consists of 3 different subjects, and the last group consists of two subjects. That leads to 5 domains in total.

PAMAP2. PAMAP2 [41] consists of data collected from 9 subjects performing 18 activities. Each subject wears 3 inertial measurement units (IMU) and a heart rate monitor. We use the data from IMU in the experiments. Each IMU consists of two 3-axis accelerometers, one 3-axis magnetometer, and one 3-axis gyroscope. To construct a domain-generalized activity scenario, we choose 8 subjects (subjects IDS 1-8) and their common eight activities: lying, sitting, standing, walking, ascending stairs, descending stairs, vacuum cleaning, and ironing. Data are divided into 4 domains.

For each task in one dataset, we select one domain as the test domain while other domains serve as the training domains. We further split a validation set from the training set with a ratio of 0.2 for hyperparameter tuning. Our main focus is to test the performance on cross-person settings in a dataset (i.e., different person, same sensor device).

4.2 Comparison Methods and Implementation Details

We compare AFFAR with several state-of-the-art domain generalization methods:

- Empirical Risk Minimization (ERM [51], i.e., CNN baseline): minimizes the sum of errors over data. We regard it as the naive baseline to learn a single model on all source domains.
- Meta-Learning Domain Generalization (MLDG [29]): learns how to generalize cross domains by leveraging Model-Agnostic Meta Learning (MAML [16]).
- Domain Adversarial Neural Network (DANN [17]): employs an adversarial network that consists of a generator and a discriminator to adapt feature distribution. Under the domain generalization setting, we perform DANN across multiple source domains as no target can be accessed during the training process.
- Group Distributionally Robust Optimization (GroupDRO [42]): couples group DRO models with increased regularization to increase the importance of the worst-group loss.
- Representation Self-Challenging (RSC [23]): discards the dominant features i.e. representations associated with the higher gradients at each epoch, and forces the model to predict with the remaining information.
- AND-mask [37]: learning explanations that are hard to vary, which uses AND-mask to improve the consistency in gradients for better generalization.

In addition, we compare the results of all methods with the results trained on the target domain (split the target data into train and test set with a rate around 8:2), which are *ideal* cases since our problem does not access the target domain data:

- ERM-t: directly trains models on the target domain using ERM.
- Fine-tune: trains a model using ERM on the training set and then fine-tunes it on the target.

Experimental settings are as follows. In order to evaluate the classification performance of AFFAR, we construct non-iid cross-person HAR tasks under the domain generalization scenario. First, we divide the data of subjects into several groups as illustrated in Section 4.1, data from each group is regarded as a domain. Each domain plays the role of the unseen target domain and remains as source domains. We use 2-D convolutions for our implementations with the kernel size of $(1, 6)$ and $(1, 9)$, depending on different datasets.

For the comparison methods, we adopt the implementations from DomainBed [22] while we change their network structures to be the same as ours for the comparison study. We perform hyperparameter tuning for each comparison method to achieve its best performance on each task. Specifically, we tune the following hyperparameters: learning rate is selected in $\{0.0003, .., 0.001\}$, batch size is set as 128. We set the total training epochs as 500 and early stop patience to 30. We run the experiments five times and report the average results. F1 score is the main evaluation metric, and we also analyze the accuracy, precision, recall, and ROC curves in detailed analysis.

4.3 Classification Performance

The test weighted F1 score on three datasets are shown in Table 2, 3, and 4, respectively. From these results, we can make some observations that: 1) The proposed AFFAR can achieve the best classification on almost all tasks. Concretely speaking, AFFAR significantly outperforms the second-best comparison methods by 2.5%, 1.7%, and 3.7% on three datasets, respectively. 2) Other comparison methods such as AND-mask and RSC can achieve good classification on some tasks while behaving less satisfied on others, this may be because they neglect the domain-specific knowledge, which may neglect some latent information between the distributions. 3) ERM is regarded as the naive baseline and the experimental results are worse than other methods on several tasks, it is because it only minimizes the empirical risk on the training source data without reducing the distribution discrepancy and investigating the latent information, the generalization capability is less satisfied due to the large distribution discrepancy. Thus, it is necessary to explore and utilize domain generalization approaches in real-life cross-domain HAR application tasks. 4) The performance of DANN is also not comparable to ours. It aims to learn domain-invariant representations. However, it does not consider the domain-specific information, making it less effective for generalization tasks. 5) The average performance of MLDG in Table 2 is significantly worse than ERM, indicating it is not feasible to directly apply meta-learning-based DG algorithms to activity recognition problems. This may be because the split of meta-train and meta-test datasets depends heavily on the independence of data, while the activity sensor data are highly dependent, thus making the results worse. The same conclusions go for data augmentation methods whose results are also not comparable, thus we did not list them. 6) Finally, the results of the two ideal-case methods ERM-t and Fine-tune which are trained on the target domain are better than all DG methods, indicating the importance of target train data. While AFFAR achieves the best results, there is still room for improvement.

Table 2. Weighted F1 score (%) on DSADS dataset. The bold is the best result except for two ideal conditions.

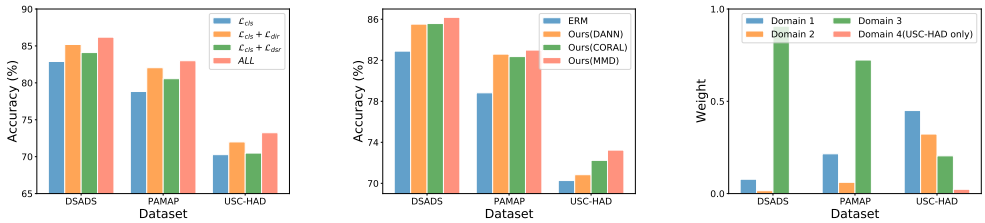
Target	ERM	MLDG	DANN	GroupDRO	RSC	AND-mask	AFFAR (ours)	ERM-t	Fine-tune
T-1	82.58	64.19	**83.94**	81.58	81.16	81.86	82.74	98.90	99.12
T-2	80.04	72.32	79.52	80.75	79.86	80.78	**84.92**	98.90	98.90
T-3	82.87	80.79	83.36	83.24	83.53	83.24	**87.60**	99.34	100
T-4	82.82	52.81	84.76	82.69	84.03	80.83	**86.45**	98.22	98.22
Average	82.10	67.53	82.90	82.07	82.15	81.68	**85.43**	98.84	99.06

Table 3. Weighted F1 score (%) on USC-HAD. The bold is the best result except for two ideal conditions.

Target	ERM	DANN	GroupDRO	RSC	AND-mask	AFFAR (ours)	ERM-t	Fine-tune
T-1	72.79	74.68	73.32	73.75	72.35	**75.11**	90.80	91.22
T-2	76.87	77.64	75.82	77.12	75.17	**78.57**	89.90	90.99
T-3	72.82	72.00	69.27	73.70	70.86	**74.59**	87.73	89.74
T-4	59.21	60.05	57.67	59.66	58.88	**62.20**	82.67	84.72
T-5	65.83	68.18	60.68	70.17	67.38	**72.41**	87.72	90.33
Average	69.50	70.51	67.35	70.88	68.93	**72.58**	87.76	89.40

Table 4. Weighted F1 score (%) on PAMAP2. The bold is the best result except for two ideal conditions.

Target	ERM	DANN	GroupDRO	RSC	AND-mask	AFFAR (ours)	ERM-t	Fine-tune
T-1	56.25	57.92	57.55	57.63	56.74	**65.37**	94.88	95.58
T-2	84.06	86.11	86.91	86.58	86.54	**89.35**	93.83	94.95
T-3	85.21	86.21	85.24	84.85	84.93	**87.43**	94.04	94.70
T-4	82.72	83.32	84.02	84.12	82.28	**86.46**	93.90	95.21
Average	77.06	78.39	78.43	78.29	77.62	**82.15**	94.16	95.11



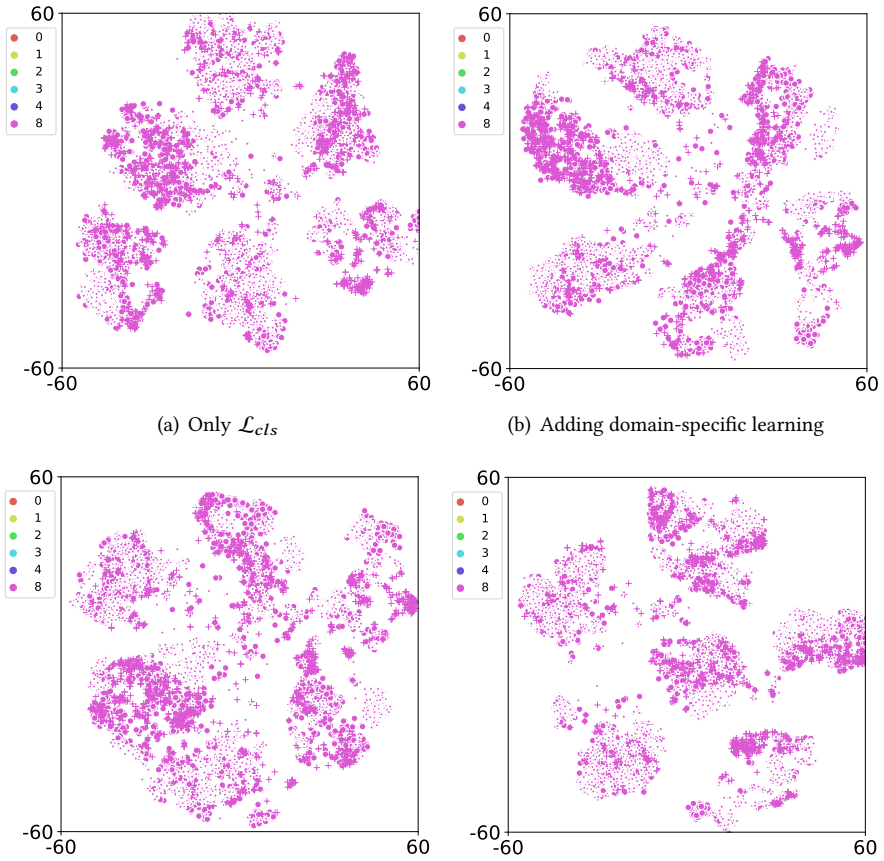
(a) Ablation study (b) AFFAR with DANN and CORAL (c) Learned weights

Fig. 3. Detailed analysis of AFFAR. (a) Ablation study to show the effectiveness of domain-invariant and domain-specific representation learning modules. (b) Replace the MMD loss with DANN and CORAL loss. (c) Weights of a randomly selected target test sample for each dataset.

4.4 Ablation Study

4.4.1 Domain-invariant and domain-specific learning modules. AFFAR consists of two important modules: (1) domain-invariant learning module and (2) domain-specific learning module. In this section, we conduct an ablation study by evaluating the importance of each module. We compare four variants of our method: (1) $\mathcal{L}_{cls}$, (2) $\mathcal{L}_{cls} + \mathcal{L}_{dir}$, (3) $\mathcal{L}_{cls} + \mathcal{L}_{dsr}$, and (4) the full version of AFFAR. Figure 3(a) reports the average classification accuracy of these variants on all tasks. It can be observed that by combining domain-invariant and domain-specific learning modules, the whole AFFAR can achieve the best performance. It evaluates that both the modules are very important and make contributions to the accurate classification of HAR tasks.

We also show the feature embeddings of each module of our method in Figure 4. (1) Compare Figure 4(a) and Figure 4(b), we see that adding domain-specific learning to the model will enlarge the domain margin and the classes are far from each other. This is because domain-specific learning focuses on separating classes in each domain. However, the domains are not aligned well (in each class, different domains denoted by shapes are still far). (2) Compare Figure 4(a) with Figure 4(c), we see that domains (denoted by shapes) are more invariant in each class since domain-invariant



(a) Only $\mathcal{L}_{cls}$ (b) Adding domain-specific learning

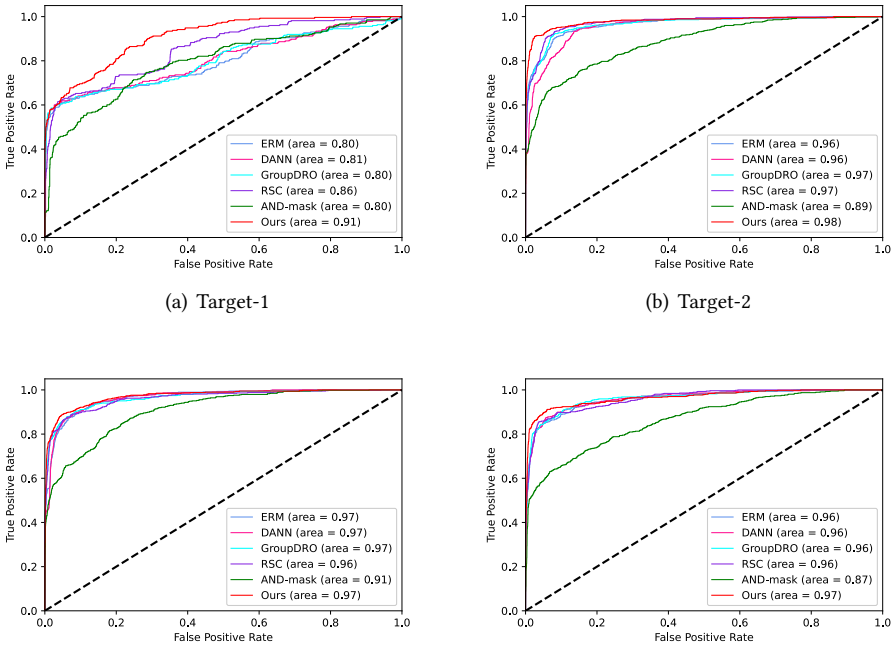
(c) Adding domain-invariant learning (d) Domain-specific + domain-invariant learning

Fig. 4. Visualization of the t-SNE embeddings of USC-HAD dataset. Each class is denoted by color and each domain is denoted by a shape. The classes denoted by numbers are walking forward, walking left, walking right, walking upstairs, walking downstairs, and standing. *Best viewed in color and zoom in.*

learning focuses on learning general features that can transfer across domains. But the classification is worse than the whole version (class margins are small, making it easy to misclassify samples). (3) Finally, we see from Figure 4(d) that adding two modules can not only make the domains more invariant but also enhance the classification results.

4.4.2 Extending domain-invariant learning with other distances. In Section 3.6, we show that our AFFAR can also take other distribution matching techniques such as domain-adversarial learning (DANN, Eq. (7)) and CORAL loss (Eq. (8)). In this section, we replace the original MMD measure with the DANN and CORAL loss to evaluate the performance of AFFAR.

We thoroughly test the performance in three different datasets and record their average accuracy in Figure 3(b). It shows that our method is general and flexible that can use other distribution matching metrics to achieve competitive performance, which is better than the original ERM. We also observe that AFFAR with MMD loss gives the best performance. On the other hand, comparing their computational complexity ($O_{MMD} \approx O_{CORAL} < O_{DANN}$), we use MMD as our main distribution matching loss.



(a) Target-1 (b) Target-2

(c) Target-3 (d) Target-4

Fig. 5. Micro-average ROC curves of AFFAR and other comparison methods for each task on PAMAP2.

4.4.3 Analysis of domain-specific module. In this section, we analyze the domain-specific module by investigating the learned domain weights to the target domain. The weights can act as the similarity between the training domains and the test data, representing how much information can be transferred from these domains. Figure 3(c) shows the (normalized) weights given to a target test sample for each dataset. Note that only the USC-HAD dataset has four training domains and the other two have three training domains. The weights reflect the similarity between the target dataset and each training domain. Thus, it shows that our AFFAR can effectively learn such similarity, which acts as the contribution of each domain to the target dataset for better generalization.

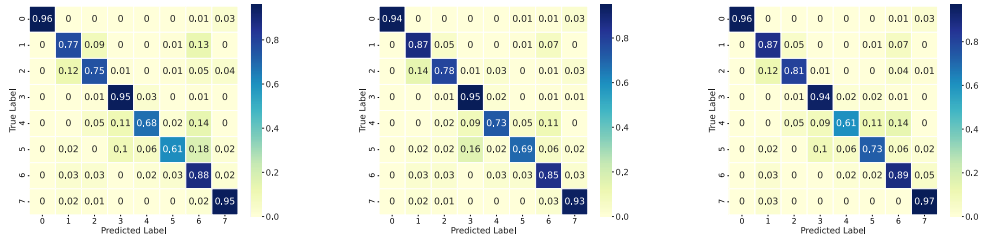
4.5 Detailed Analysis

To further evaluate the performance of AFFAR on each class, we provide the fine-grained analysis by the micro-average Receiver Operating Characteristics (ROC) curves in Figure 5. ROC is a more effective metric for the areas of cost-sensitive learning and unbalanced class issues [14]. We also calculate the AUC (Area Under Curve) on the four tasks of the PAMAP2 dataset. Figure 5 illustrates the following conclusions: 1) Our AFFAR can achieve good performance on four tasks. The micro-average results of four target tasks are all higher than 0.91 which evaluates the classification effectiveness of AFFAR. 2) Results of all methods on Target-1 are less satisfying, which may be because the distribution discrepancy of the other domains is smaller so the generalization capability is reduced. The improvement of AFFAR is more obvious on Target-1 than on the other three target tasks, which shows that our approach is more robust on hard tasks.

Meanwhile, we randomly select a task of PAMAP2 to further make the fine-grained evaluation from the perspective of comparison with some state-of-the-art methods by utilizing the metrics of multi-class precision (P), recall (R), F1 scores (F1) in Table 5 and the visualization of the confusion matrix for each category in Figure 6. From these experimental results, it can be observed that

Table 5. Precision, recall, and F1 score of AFFAR and other comparison methods for each class on PAMAP2. Numbers in each cell denote the percentage of the prediction for each class.

	DANN			GroupDRO			RSC			AND-mask			AFFAR		
Activity	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Lying	**1.00**	0.94	0.97	**1.00**	**0.96**	**0.98**	0.99	**0.96**	0.97	0.99	**0.96**	0.97	**1.00**	**0.96**	**0.98**
Sitting	0.83	**0.87**	0.85	0.83	**0.87**	0.85	0.82	0.83	0.82	0.89	0.77	0.82	**0.92**	0.80	**0.86**
Standing	0.85	0.78	0.81	**0.89**	0.81	0.85	0.82	0.82	0.82	0.82	0.81	0.81	0.83	**0.92**	**0.87**
Walking	0.88	0.95	0.91	0.90	0.94	0.92	**0.95**	0.93	0.94	0.91	0.94	0.92	**0.95**	**0.96**	**0.95**
Ascending stairs	0.82	0.73	0.77	0.82	0.61	0.70	**0.84**	0.70	0.77	0.80	**0.80**	**0.80**	**0.84**	0.73	0.78
Descending stairs	0.85	0.69	0.76	0.80	0.73	0.77	0.88	**0.76**	0.81	0.83	0.71	0.77	**0.97**	**0.76**	**0.85**
Vacuum cleaning	0.73	0.85	0.79	**0.74**	0.89	**0.81**	0.69	0.82	0.75	0.70	0.88	0.78	0.72	**0.92**	**0.81**
Ironing	0.91	0.93	0.92	0.93	**0.97**	0.95	0.93	**0.97**	0.95	**0.94**	0.96	0.95	**0.94**	**0.97**	0.95
Average	0.86	0.84	0.85	0.86	0.85	0.85	0.87	0.85	0.85	0.86	0.85	0.85	**0.90**	**0.88**	**0.88**



(a) ERM (b) DANN (c) GroupDRO

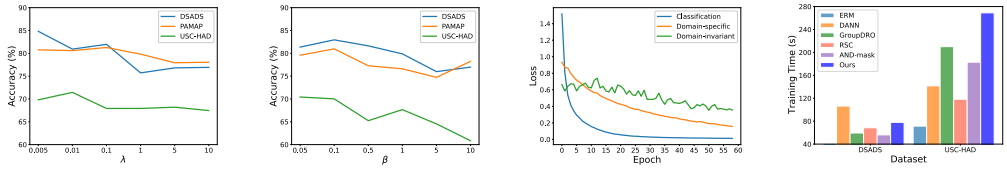
(d) RSC (e) AND-mask (f) Ours

Fig. 6. Confusion matrices of all methods on PAMAP2. Classes 0-7 denote the following activities: lying, sitting, standing, walking, ascending stairs, descending stairs, vacuum cleaning, and ironing.

AFFAR can get the most number of the best precision-recall-F1 results on all categories than other comparison methods, and can get the best average precision-recall-F1 across each class. Besides, from the confusion matrix, we can see that AFFAR can get more balance results on each category. Most methods get less satisfying results on the fourth and fifth class, AFFAR can reduce the performance degradation. Although AND-mask can get the best performance on the fourth class, it gets less satisfying results on other classes than AFFAR except for the first class where all the comparative methods can achieve a satisfying performance.

4.6 Parameter Sensitivity, Convergence, and Time Analysis

We empirically evaluate the sensitivity of two parameters λ and β by setting their values from $\{0.005, 0.01, 0.1, 1, 5, 10\}$ and $\{0.05, 0.1, 0.5, 1, 5, 10\}$, respectively . The results are shown in Figure 7(a) and 7(b). It indicates that our AFFAR stays robust to a wide range of parameter choices.



(a) Sensitivity of λ (b) Sensitivity of β (c) Convergence (d) training Time

Fig. 7. Parameter sensitivity of (a) λ and (b) β. (c) is convergence analysis. (d) is training time analysis.

Table 6. Comparison of inference time and weighted F1-score

Method	ERM	DANN	GroupDRO	RSC	AND-mask	AFFAR (Ours)
Inference time (s)	.000010	.000012	.000011	.000011	.000012	.000057
Weighted F1 (%)	82.10	82.90	82.07	82.15	81.68	85.43

We also empirically analyze the convergence of our method and draw the loss curve on a randomly chosen task of the PAMAP2 dataset in Figure 7(c). Other tasks follow similar observations. The results show that AFFAR can converge in dozens of epochs, indicating that it is easy to train.

Furthermore, we show the training time of each method on one task in Figure 7(d). We see that the training time of our method is almost the same as others, while slightly takes more time in some circumstances. This is reasonable because our method is an ensemble-based learning process with several domain-specific branches to be learned, which is comparable with other methods. We obtain similar observations for the inference time.

As inference time is also very crucial to make quick classification while performing accurate activity recognition. Table 6 shows the comparison of the average inference time of each sample and weighted F1 score on the DSADS dataset. We can observe that the inference time of different methods are similar and the proposed method is slightly longer, while the recognition performance is the best among the comparison methods. This indicates the method can make better activity recognition in applications with a little more time. And we will make efforts to further reduce the inference time for better applications in future work.

5 APPLICATION TO ADHD RECOGNITION

Attention Deficit Hyperactivity Disorder (ADHD) is one of the most common mental disorders in children [26]. ADHD is characterized by inappropriate inattention, hyperactivity, and impulsivity [2]. It is often accompanied by some motor abnormality, thus it is possible and necessary to utilize the HAR method to assist the diagnosis. In this section, we apply the proposed AFFAR to the ADHD application to further evaluate its effectiveness.

We apply our algorithm to a real-world ADHD dataset [26]. This dataset is collected with a designed wearable diagnostic assessment system in the room environment with little tension. Ten diagnostic tasks including six interactions with the screen tasks and four physical objects interaction tasks (Schulte grid, Multi-ball tracking, Catching grasshopper, Drinking birds, Limb reaction, Reading, Finger holes, Shape-color conflicting, Catching worms, and Keeping balance) are designed for children for ADHD symptoms according to DSM-5. Six wearable sensors are attached to children, one on each wrist and each ankle, one on the head, one on the waist during the ten assessment tasks and accelerometer data are collected. 54-dimensional motion features are extracted according to [26]. For more detailed information about this dataset, please refer to [26].

In our experiment, features from 83 normal children and 83 children with ADHD are involved. Meanwhile, the children with ADHD are diagnosed by doctors to confirm they meet the diagnostic criteria for ADHD. Furthermore, we have more fine-grained labels, i.e., children with ADHD are diagnosed with subtypes, i.e. predominantly inattentive (50 children), predominantly hyperactive-impulsive (14 children), and combination (19 children). This is a multi-class classification task.

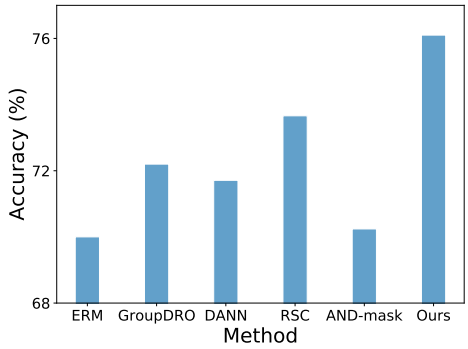


Fig. 8. Multi-class classification in ADHD application.

We divide the data into the source training set and target test set, with a number of participants of 125 and 41, respectively. To harness the different distributions in the training set, we further divide it into 3 domains. The classification results of all methods are shown in Figure 8. We can observe that AFFAR improves the classification accuracy with a rate of around 2.44%. It indicates the effectiveness of AFFAR in classification with ADHD and the effectiveness of fine-grained classification on subtypes. Experimental results also show that AFFAR has the potential to the applied in real-life wearable healthcare.

6 CONCLUSIONS AND FUTURE WORK

Generalization to the unseen test data has always been the key research and application problem in human activity recognition. While transfer learning and domain adaptation approaches rely on the availability of test data during the training stage, in this paper, we propose AFFAR to solve this problem by learning both the domain-invariant and domain-specific features. The key of our algorithm is to preserve the specific representations of the training data while learning transferable representations, which could be informative to the generalization on unseen test data. Experiments on both public datasets and the real application have demonstrated the superiority of our method.

In the future, we plan to extend AFFAR in the following two directions. First, apply it to more healthcare applications such as the diagnosis of Parkinson's disease. Second, experiments show that there is still a gap between our method and fine-tuning, which motivates us to improve the performance of our method by introducing other domain-invariant learning modules.

ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Plan of China (No. 2021YFC2501202), Natural Science Foundation of China (No. 61972383, No. 61902377, No. 62101530, No. 61902379), Science and Technology Service Network Initiative, Chinese Academy of Sciences (No. KFJ-STS-QYZD-2021-11-001).

REFERENCES

[1] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. 2019. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804* (2019).

[2] Russell A Barkley. 1998. Attention-deficit hyperactivity disorder. *Scientific American* 279, 3 (1998), 66–71.

[3] Billur Barshan and Murat Cihan Yüksek. 2014. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *Comput. J.* 57, 11 (2014), 1649–1667.

[4] Gilles Blanchard, Gyemin Lee, and Clayton Scott. 2011. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems* 24 (2011), 2178–2186.

[5] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. 2019. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2229–2238.

[6] Junbum Cha, Hancheol Cho, Kyungjae Lee, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021. Domain generalization needs stochastic weight averaging for robustness on domain shifts. *arXiv 2102.08604* (2021).

[7] 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.

[8] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–40.

[9] Liming Chen, Jesse Hoey, Chris D Nugent, Diane J Cook, and Zhiwen Yu. 2012. Sensor-based activity recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 790–808.

[10] Diane Cook, Kyle D Feuz, and Narayanan C Krishnan. 2013. Transfer learning for activity recognition: A survey. *Knowledge and information systems* 36, 3 (2013), 537–556.

[11] Ditte Demontis, Raymond K Walters, Joanna Martin, Manuel Mattheisen, Thomas D Als, Esben Agerbo, Gísli Baldursson, Rich Belliveau, Jonas Bybjerg-Grauholm, Marie Bækvad-Hansen, et al. 2019. Discovery of the first genome-wide significant risk loci for attention deficit/hyperactivity disorder. *Nature genetics* 51, 1 (2019), 63–75.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Thomas G Dietterich et al. 2002. Ensemble learning. *The handbook of brain theory and neural networks* 2 (2002), 110–125.

[14] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern recognition letters* 27, 8 (2006), 861–874.

[15] Kyle Dillon Feuz and Diane J Cook. 2014. Heterogeneous transfer learning for activity recognition using heuristic search techniques. *International Journal of Pervasive Computing and Communications* (2014).

[16] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*. PMLR, 1126–1135.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[18] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. 2015. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*. 2551–2559.

[19] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. 2019. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2477–2486.

[20] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13, Mar (2012), 723–773.

[21] Yu Guan and Thomas Plötz. 2017. Ensembles of deep lstm learners for activity recognition using wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 2 (2017), 1–28.

[22] Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. In *International Conference on Learning Representations (ICLR)*.

[23] Zeyi Huang, Haohan Wang, Eric P Xing, and Dong Huang. 2020. Self-challenging improves cross-domain generalization. In *European Conference on Computer Vision (ECCV)*, Vol. 2.

[24] Navdeep Jaitly, Patrick Nguyen, Andrew Senior, and Vincent Vanhoucke. 2012. Application of pretrained deep neural networks to large vocabulary speech recognition. (2012).

[25] 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.

[26] Xinlong Jiang, Yiqiang Chen, Wuliang Huang, Teng Zhang, Chenlong Gao, Yunbing Xing, and Yi Zheng. 2020. WeDA: Designing and Evaluating A Scale-driven Wearable Diagnostic Assessment System for Children with ADHD. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[27] 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)*. IEEE, 1–9.

[28] Oscar D Lara and Miguel A Labrador. 2012. A survey on human activity recognition using wearable sensors. *IEEE communications surveys & tutorials* 15, 3 (2012), 1192–1209.

[29] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[30] Wen Li, Zheng Xu, Dong Xu, Dengxin Dai, and Luc Van Gool. 2017. Domain generalization and adaptation using low rank exemplar svms. *IEEE transactions on pattern analysis and machine intelligence* 40, 5 (2017), 1114–1127.

[31] Xi'ang Li, Jinqi Luo, and Rabih Younes. 2020. ActivityGAN: generative adversarial networks for data augmentation in sensor-based human activity recognition. In *ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 249–254.

[32] Na Lu, Yidan Wu, Li Feng, and Jinbo Song. 2018. Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data. *IEEE journal of biomedical and health informatics* 23, 1 (2018), 314–323.

[33] Minghuang Ma, Haoqi Fan, and Kris M Kitani. 2016. Going deeper into first-person activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1894–1903.

[34] Massimiliano Mancini, Samuel Rota Bulo, Barbara Caputo, and Elisa Ricci. 2018. Best sources forward: domain generalization through source-specific nets. In *2018 25th IEEE international conference on image processing (ICIP)*. IEEE, 1353–1357.

[35] Toshihiko Matsuura and Tatsuya Harada. 2020. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11749–11756.

[36] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[37] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. 2020. Learning explanations that are hard to vary. In *International Conference on Learning Representations (ICLR)*.

[38] Fengchun Qiao, Long Zhao, and Xi Peng. 2020. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12556–12565.

[39] 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.

[40] Nishkam Ravi, Nikhil Dandekar, Preetham Mysore, and Michael L Littman. 2005. Activity recognition from accelerometer data. In *Aaai*, Vol. 5. Pittsburgh, PA, 1541–1546.

[41] Attila Reiss and Didier Stricker. 2012. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*. IEEE, 108–109.

[42] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*.

[43] Andrea Rosales Sanabria and Juan Ye. 2020. Unsupervised domain adaptation for activity recognition across heterogeneous datasets. *Pervasive and Mobile Computing* 64 (2020), 101147.

[44] Mattia Segu, Alessio Tonioni, and Federico Tombari. 2020. Batch normalization embeddings for deep domain generalization. *arXiv preprint arXiv:2011.12672* (2020).

[45] Shiv Shankar, Vihari Piratla, Soumen Chakrabarti, Siddhartha Chaudhuri, Preethi Jyothi, and Sunita Sarawagi. 2018. Generalizing across domains via cross-gradient training. *ICLR* (2018).

[46] Ling Shao, Fan Zhu, and Xuelong Li. 2014. Transfer learning for visual categorization: A survey. *IEEE transactions on neural networks and learning systems* 26, 5 (2014), 1019–1034.

[47] Gulbadan Sikander and Shahzad Anwar. 2018. Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems* 20, 6 (2018), 2339–2352.

[48] Elnaz Soleimani and Ehsan Nazerfard. 2021. Cross-subject transfer learning in human activity recognition systems using generative adversarial networks. *Neurocomputing* 426 (2021), 26–34.

[49] Mihaela C Stoian, Sameer Bansal, and Sharon Goldwater. 2020. Analyzing ASR pretraining for low-resource speech-to-text translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7909–7913.

[50] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*. Springer, 443–450.

[51] Vladimir Vapnik. 1992. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*. 831–838.

[52] Hong-Bo Wang, Yanze Xue, Xiaoxiao Zhen, and Xuyan Tu. 2018. Domain Specific Learning for Sentiment Classification and Activity Recognition. *IEEE Access* 6 (2018), 53611–53619.

[53] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119 (2019), 3–11.

[54] Jindong Wang, Yiqiang Chen, Lisha Hu, Xiaohui Peng, and S Yu Philip. 2018. Stratified transfer learning for cross-domain activity recognition. In *2018 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 1–10.

[55] 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).

[56] Shujun Wang, Lequan Yu, Kang Li, Xin Yang, Chi-Wing Fu, and Pheng-Ann Heng. 2020. Dofe: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Transactions on Medical Imaging* 39, 12 (2020), 4237–4248.

[57] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data* 3, 1 (2016), 1–40.

[58] Shen Yan, Huan Song, Nanxiang Li, Lincan Zou, and Liu Ren. 2020. Improve unsupervised domain adaptation with mixup training. *arXiv preprint arXiv:2001.00677* (2020).

[59] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. 2019. Time-series generative adversarial networks. *Advances in Neural Information Processing Systems* 32 (2019).

[60] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[61] Mi Zhang and Alexander A Sawchuk. 2012. USC-HAD: a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *ACM UbiComp*. ACM, 1036–1043.

[62] Fan Zhou, Zhuqing Jiang, Changjian Shui, Boyu Wang, and Brahim Chaib-draa. 2020. Domain generalization with optimal transport and metric learning. *arXiv preprint arXiv:2007.10573* (2020).