Cross-Modal Federated Human Activity Recognition via Modality-Agnostic and Modality-Specific Representation Learning

Xiaoshan Yang[1,3,4], **Baochen Xiong**[2,4], **Yi Huang**[1,3], **Changsheng Xu**[1,3,4*]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]Zhengzhou University
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4]Peng Cheng Laboratory

Abstract

In this paper, we propose a new task of cross-modal federated human activity recognition (CMF-HAR), which is conducive to promote the large-scale use of the HAR model on more local devices. To address the new task, we propose a feature-disentangled activity recognition network (FDARN), which has five important modules of altruistic encoder, egocentric encoder, shared activity classifier, private activity classifier and modality discriminator. The altruistic encoder aims to collaboratively embed local instances on different clients into a modality-agnostic feature subspace. The egocentric encoder aims to produce modality-specific features that cannot be shared across clients with different modalities. The modality discriminator is used to adversarially guide the parameter learning of the altruistic and egocentric encoders. Through decentralized optimization with a spherical modality discriminative loss, our model can not only generalize well across different clients by leveraging the modality-agnostic features but also capture the modality-specific discriminative characteristics of each client. Extensive experiment results on four datasets demonstrate the effectiveness of our method.

Introduction

Human activity recognition (HAR) aims at identifying the types of activities performed by humans based on information received from different devices, e.g., cameras and motion sensors. It has attracted increasing attention due to its promising application in many fields, such as security monitoring, health management and smart home. Nowadays, the popular use of mobile phones and smart bracelets makes it much easier to collect video and sensor data, which has significantly inspired the research on vision-based (Simonyan and Zisserman 2014; Song et al. 2016a), sensor-based (Bulling, Blanke, and Schiele 2014) and multimodal (Nakamura et al. 2017; Possas, Pinto-Caceres, and Ramos 2018) activity recognition approaches.

Most existing HAR methods assume that all data samples collected from local devices should be uniformly stored and processed on a central server, which may lead to the leakage of private user information. To avoid uploading local data to central servers, McMahan et al. (2017) propose the first federated learning (FL) algorithm, FedAvg, which can aggregate the gradients of models independently learned from different local clients to establish a global model. Afterwards, the FedAvg has also been successfully applied to HAR (Sozinov, Vlassov, and Girdzijauskas 2018). As an improvement, Li et al. (2021a) propose to leverage meta learning to learn an embedding network for addressing the heterogeneity in both label and sample distributions.

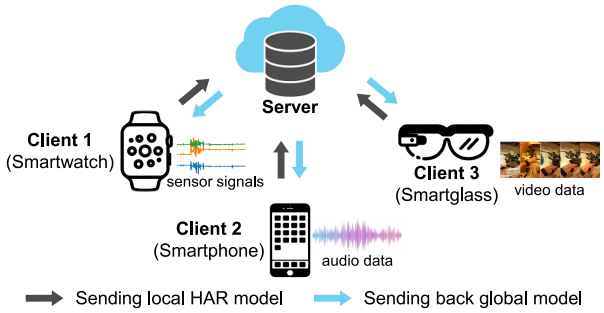


Figure 1: Illustration of the proposed CMF-HAR task.

Although existing FL-based HAR methods (Sozinov, Vlassov, and Girdzijauskas 2018; Ek et al. 2020; Jain et al. 2021; Li et al. 2021a) have achieved important progress, they only consider the case where the local data of different clients are from the same modality. e.g., sensor signals. It is still questionable whether existing FL methods can be directly used in a more common scenario where local data are from different modalities, e.g., some local clients may provide sensor signals while others can only provide visual data. We believe that this scenario deserves more attention because it is conducive to distributively learn or apply the HAR model on more local devices. Therefore, in this paper, **we propose a new task of cross-modal federated human activity recognition (CMF-HAR)**, which focuses on distributively aggregating local models learned on clients with different modalities as shown in Figure 1.

Compared with conventional FL, **the CMF-HAR task has at least two more challenges**. **(1)** *How to collaboratively build a common feature subspace for different clients that have cross-modal heterogeneity.* In conventional FL-based HAR (Sozinov, Vlassov, and Girdzijauskas 2018; Li et al. 2021a), local data are collected from the same type of sensor and the heterogeneity in sample distributions is

*Corresponding author.

mainly resulted from the various motion patterns of different users. Whereas, in the CMF-HAR task, the data of different local clients always have extraordinary different structure and content, e.g., the sensor signal is recorded with the sequence of three-axis values while the video has much more complex spatial and temporal information, which results in much larger distribution heterogeneity. Moreover, the restriction of data privacy brings extra difficulties because we cannot simply organize the distributed local data together to learn a shared feature space as in conventional cross-modal tasks (Wang et al. 2017). **(2)** *Not all knowledge learned from one client is useful for HAR on other clients with different modalities.* Conventional FL-based HAR methods only consider instances from the same type of modality, where the discriminative patterns learned on one local client can always be shared by other local clients. Whereas, in the CMF-HAR task, different clients probably have different discriminative patterns due to the large heterogeneity. For example, for the local models learned on clients with video data, it is useful to identify the visual patterns of *tennis racket* to correctly recognize an activity *play tennis*. However, this kind of discriminative ability on visual data is superfluous on clients with only accelerometer or gyroscope signals.

To address above challenges of the CMF-HAR task, **we propose a feature-disentangled activity recognition network (FDARN)**, which has five important modules of altruistic encoder, egocentric encoder, shared activity classifier, private activity classifier and modality discriminator. The altruistic encoder aims to embed the instances on different clients into a modality-agnostic feature subspace where the shared activity classifier can be effectively learned across different clients. The egocentric encoder aims to produce the modality-specific feature that is only important for one modality and can be used to learn the private activity classifier. To encourage the altruistic and egocentric encoders produce different representations, a separation loss is adopted to ensure the orthogonality of their output features. In addition, the modality discriminator is adopted to recognize the modality label and jointly guide the parameter learning of the altruistic and egocentric encoders in an adversarial manner. To avoid learning a trivial modality discriminator from distributed local data where each client has only one type of modality, we leverage a spherical modality discriminative loss to enhance the intra-class compactness and inter-class discrepancy for the hidden representations and parameters of the modality discriminator in a hyper-sphere. Through decentralized optimization with alternative local update and global aggregation steps, the proposed FDARN can not only generalize well across different clients by leveraging the modality-agnostic features but also capture the modality-specific discriminative characteristics of each client.

The main contributions of this paper are three-fold. **(1)** We propose a new task of cross-modal federated human activity recognition, which is important for the large-scale use of the HAR model on more local devices that have different modalities of data. **(2)** To solve the new task, we propose a feature-disentangled activity recognition network. Through decentralized optimization with a spherical modality discriminative loss, our model can comprehensively build a modality-agnostic feature subspace for collaboratively learning activity classifiers on different clients and capture the modality-specific discriminative characteristics of each client. **(3)** We evaluate the proposed method on four datasets and demonstrate its effectiveness with extensive experimental results.

Related Work

Federated Learning

The first FL algorithm FedAvg was proposed by McMahan et al. (2017), which has promising performance in decentralized model learning and privacy protection. However, the FedAvg suffers from an inevitable performance reduction on non-iid data. To deal with this problem, Zhao et al. (2018) improve the FedAvg by sharing a small set of data between different local clients. FedMA (Wang et al. 2020) constructs a layer-wise federated learning algorithm for deep learning models by matching and averaging hidden elements of the neural layers with similar feature extraction signatures. FedRobust (Reisizadeh et al. 2020) attempts to address the device-dependent data heterogeneity, which assumes that local instances are shifted from a ground distribution by an affine transformation. Li, He, and Song (2021) propose to conduct contrastive learning in model-level to constrain the local model and the global model to produce consistent representations. FedBN (Li et al. 2021b) uses local batch normalization to alleviate the feature shift between different clients. FedDis (Bercea et al. 2021) is a FL method for unsupervised brain pathology segmentation, which can disentangle the parameter space into shape and appearance. FedCMR (Zong et al. 2021) focuses on a federated cross modal retrieval task where each client has both text and image data. In contrast, samples from different modalities in CMF-HAR are stored on different clients. More related work to ours is LG-FedAvg (Liang et al. 2020) and FedRep (Collins et al. 2021). The former learns the representation of local data with a client-specific local model while the latter learns a globally consistent function to compute the shared representations for different clients. Unfortunately, these methods only consider the heterogeneity across clients that have the same modality of data.

FL-based Activity Recognition

Sozinov, Vlassov, and Girdzijauskas (2018) are the first to address the privacy issue of HAR using FL. Ek et al. (2020) show that, when directly applying existing FL methods to HAR, the FedAvg (McMahan et al. 2017) performs even better than more sophisticated FL algorithms, e.g., FedPer (Arivazhagan et al. 2019) and FedMA (Wang et al. 2020), which calls for more dedicated research on FL-based HAR. More recently, Li et al. (2021a) propose a meta learning-based HAR algorithm that can reduce the heterogeneity of different clients by a deep embedding network. Jain et al. (2021) propose a knowledge distillation based asynchronous federated optimization method to handle the heterogeneity in the computing resources of different clients. The above FL-based HAR methods do not consider the case that local clients have different modalities of data, and thus cannot be directly applied to solve the proposed CMF-HAR task.

Disentangled Representation Learning

Early work on disentangled representation learning is mostly based on auto-encoders (Kingma and Welling 2014; Bousmalis et al. 2016; Higgins et al. 2017) and generative adversarial networks (Odena, Olah, and Shlens 2017). Kim and Mnih (2018) propose to constrain the representation to be factorial and independent across the dimensions. More recently, the disentangled representation learning has been successfully applied to cross-modal tasks. For example, Wu et al. (2019) propose a disentangled variational representation method for heterogeneous face recognition based on Wasserstein CNN by aligning the correlation between different modality variations. Guo et al. (2019) propose to disentangle the modality exclusive information from the learned common representations for cross-modal retrieval with deep mutual information estimation. Xu, Zhang, and Duan (2020) propose a feature aggregation network, which can learn domain-private features and domain-agnostic features for modality adaptive face recognition. The above cross-modal disentangled learning methods assume that the data of different modalities are simultaneously accessible in model learning, and thus cannot be directly used in the CMF-HAR task, where each client has only one type of modality and the data are not allowed to be shared across different clients.

Methodology

Problem Definition

We assume that there is a set of data $\mathcal{D} = \{(\mathcal{D}^k, z^k)\}_{k=1}^K$ from K clients, where $\mathcal{D}^k = \{\mathbf{x}_i^k, \mathbf{y}_i^k\}_{i=1}^{n_k}$ is the set of n_k instances on the k-th client. The total number of instances is $n = \sum_{k=1}^K n_k$. $\mathbf{x}_i^k \in \mathcal{X}$ is the i-th instance of the k-th client. The $\mathbf{y}_i^k \in \mathbb{R}^N$ denotes the one-hot encoding of the activity class. N denotes the total number of activity classes. $z^k \in \{1, ..., M\}$ denotes the modality label of the k-th client. M is the total number of modalities and $M \leqslant K$.

The CMF-HAR task aims to distributively learn a HAR model that can work well on each local client by leveraging the cross-modal discriminative knowledge learned on other clients without centrally collecting the data. The objective of the CMF-HAR task is formulated as:

$$\sum_{k=1}^K \frac{n_k}{n} \sum_{i=1}^{n_k} \frac{1}{n_k} \ell\big(f^k(\mathbf{x}_i^k), \mathbf{y}_i^k\big), \tag{1}$$

where f^k: $\mathcal{X} \to \mathbb{R}^N$ is a classification function learned on the local data of the k-th client. ℓ is a loss function (e.g., cross-entropy loss). To achieve the above objective, conventional FL methods (McMahan et al. 2017) focus on learning a single shared model f that performs averagely well on different clients. Whereas, in the CMF-HAR task, due to the large heterogeneity of different clients, it is difficult to learn a single classification function. Therefore, we redefine f^k in a disentangled form:

$$f^k(\mathbf{x}) = \psi_{sc}(\phi_a(\mathbf{x}; \mathbf{\Theta}_a^{z^k}); \mathbf{\Theta}_{sc}) + \psi_{pc}(\phi_e(\mathbf{x}; \mathbf{\Theta}_e^{z^k}); \mathbf{\Theta}_{pc}^{z^k}), \tag{2}$$

where ϕ_a: $\mathcal{X} \to \mathbb{R}^d$ is an altruistic encoder that explicitly maps an input instance $\mathbf{x}$ to a modality-agnostic feature subspace. ϕ_e: $\mathcal{X} \to \mathbb{R}^d$ is an egocentric encoder that computes the modality-specific feature. ψ_{sc}: $\mathbb{R}^d \to \mathbb{R}^N$ and ψ_{pc}: $\mathbb{R}^d \to \mathbb{R}^N$ are shared activity classifier and private activity classifier learned on the modality-agnostic feature and the modality-specific feature, respectively. Here, we assume that all clients share a single activity classifier ψ_{sc} while ϕ_a, ϕ_e and ψ_{pc} are shared across clients with the same modality.

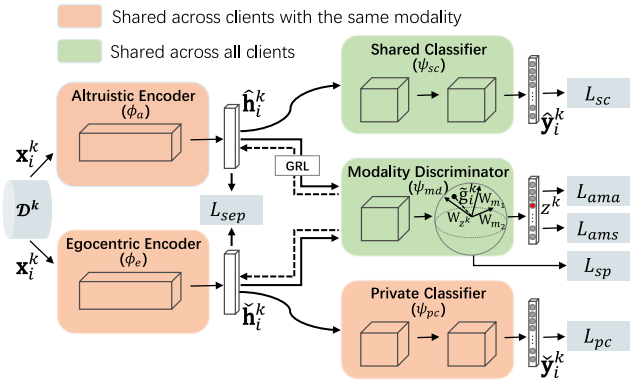


Figure 2: Overview of the FDARN on a local client.

In this work, the classification function f^k defined in Eq. (2) is instantiated as a feature disentanglement activity recognition network.

Feature-Disentangled Activity Recognition Network

As shown in Figure 2, the proposed FDARN has five important modules, i.e., altruistic encoder ϕ_a, egocentric encoder ϕ_e, shared classifier ψ_{sc}, private classifier ψ_{pc} and modality discriminator ψ_{md}. For the altruistic and egocentric encoders, we can use different backbones for different modalities and we leave the details in the experiment section. For the shared activity classifier and the private classifier, we define them as two-layer perceptrons with ReLU activation functions and Softmax outputs. The modality discriminator ψ_{md}: $\mathbb{R}^d \to \mathbb{R}^M$ is adopted to identify whether the altruistic encoder and the egocentric encoder can produce modality-agnostic and modality-specific features, respectively. Next, we will illustrate more details of the modality discriminator and the loss functions used to constrain the FDARN.

Separation Loss. The separation loss aims to encourage the altruistic and egocentric encoders to produce different features that represent different aspects of the input instance. Inspired by domain separation networks (Bousmalis et al. 2016), we adopt soft subspace orthogonality constraint:

$$\mathcal{L}_{sep}(\mathcal{D}^k) = \|\hat{\mathbf{H}}_k^\top \check{\mathbf{H}}_k\|_F^2, \tag{3}$$

where $\hat{\mathbf{H}}_k \in \mathbb{R}^{d \times n_k}$ is a matrix whose columns are the output features $\{\hat{\mathbf{h}}_i^k\}_{i=1}^{n_k}$ of the altruistic encoder, $\hat{\mathbf{h}}_i^k = \phi_a(\mathbf{x}_i^k; \mathbf{\Theta}_a^{z^k})$. Similarly, $\check{\mathbf{H}}_k \in \mathbb{R}^{d \times n_k}$ consists of output features $\{\check{\mathbf{h}}_i^k\}_{i=1}^{n_k}$ of the egocentric encoder, $\check{\mathbf{h}}_i^k = \phi_e(\mathbf{x}_i^k; \mathbf{\Theta}_e^{z^k})$. $\|\cdot\|_F$ is Frobenius norm.

Modality Discriminator. The separation loss can encourage the altruistic and egocentric encoders to produce different representations. However, the separation loss still cannot guarantee that the output feature of the altruistic encoder ϕ_a belongs to a latent space shared across different modalities and the output feature of the egocentric encoder ϕ_e exactly reflects the modality-specific characteristics of the client. To deal with this problem, we use the modality discriminator

ψ_{md} to identify the modality label and guide the parameter learning of the altruistic and egocentric encoders.

To make it easier to explain, we formally define the modality discriminator as $\psi_{md}(\mathbf{g};\mathbf{\Theta}_{md}) = Softmax\big(\mathbf{W}^\top\varphi(\mathbf{g})\big)$, where $\mathbf{g}\in\mathbb{R}^d$ is the input feature of the modality discriminator, which can be either the output $\hat{\mathbf{h}}_i^k$ of the altruistic encoder or the output $\check{\mathbf{h}}_i^k$ of the egocentric encoder. φ: $\mathbb{R}^d\to\mathbb{R}^{\tilde{d}}$ is a single-layer perception with the nonlinear activation function of ReLU. Here, following (Deng et al. 2019), we assume that the last linear transformation layer of the modality discriminator does not have bias term and its weight matrix is denoted as $\mathbf{W}\in\mathbb{R}^{\tilde{d}\times M}$.

As a multi-class classifier, it is straightforward to learn the modality discriminator with cross-entropy loss. However, directly using the cross-entropy will lead to a trivial solution. Because each client has only one type of modality and the requirement of data privacy impedes the direct accessing of instances from other clients. To address this issue, we firstly normalize the hidden representation $\tilde{\mathbf{g}} = \varphi(\mathbf{g})$ and each column of $\mathbf{W}$ with L2-norm. Then, we use additive angular margin loss (Deng et al. 2019) to enhance the intra-class compactness and inter-class discrepancy for the modality discriminator in a hyper-sphere:

$$\mathcal{L}_{am}(\mathbf{g},z) = -\log\frac{e^{\alpha\cos(\theta_z+\tau)}}{e^{\alpha\cos(\theta_z+\tau)}+\sum\limits_{m=1,m\neq z}^{M}e^{\alpha\cos(\theta_m)}}, \quad (4)$$

where $\theta_z = \arccos(\mathbf{W}_z^\top\tilde{\mathbf{g}}), \theta_m = \arccos(\mathbf{W}_m^\top\tilde{\mathbf{g}})$. z denotes the ground-truth modality label. $\mathbf{W}_z\in\mathbb{R}^{\tilde{d}}$ denotes the z-th column of the weight matrix $\mathbf{W}$. Similarly, $\mathbf{W}_m\in\mathbb{R}^{\tilde{d}}$ denotes the m-th column of the weight matrix $\mathbf{W}$. τ is a margin factor and α is a scale factor.

Since we need to use the modality discriminator to measure the output features of both the altruistic encoder and the egocentric encoder, we compute the additive angular margin loss for all instances on the local client as follows:

$$\mathcal{L}_{ama}(\mathcal{D}^k,z^k) = \frac{1}{n_k}\sum\nolimits_{i=1}^{n_k}\mathcal{L}_{am}(\hat{\mathbf{h}}_i^k,z^k), \quad (5)$$

$$\mathcal{L}_{ams}(\mathcal{D}^k,z^k) = \frac{1}{n_k}\sum\nolimits_{i=1}^{n_k}\mathcal{L}_{am}(\check{\mathbf{h}}_i^k,z^k), \quad (6)$$

where $\hat{\mathbf{h}}_i^k = \phi_a(\mathbf{x}_i^k;\mathbf{\Theta}_a^{z_k})$, and $\check{\mathbf{h}}_i^k = \phi_e(\mathbf{x}_i^k;\mathbf{\Theta}_e^{z_k})$.

In addition, to avoid learning identical weight parameters for negative modalities, we adopt a spreadout regularizer (Yu et al. 2020) to ensure that parameters of different modalities are separated from each other:

$$\mathcal{L}_{sp} = \sum_{m=1}^{M}\sum_{\substack{m'=1\\m'\neq m}}^{M}\max\big\{0,\nu-\big(1-\mathbf{W}_m^\top\mathbf{W}_{m'}\big)\big\}, \quad (7)$$

where ν is a margin factor.

Finally, the spherical modality discriminative loss for learning the modality discriminator is defined as $\mathcal{L}_{md} = \mathcal{L}_{sp}-\mathcal{L}_{ama}+\mathcal{L}_{ams}$, where the minus before L_{ama} is used to adversarially learn the altruistic encoder. As shown in Figure 2, with the spherical modality discriminative loss, we can constrain the hidden representation $\tilde{\mathbf{g}}_i^k$ of the input $\mathbf{x}_i^k$ to be close to the parameter $\mathbf{W}_{z^k}$ of its ground-truth modality label z^k in a hyper-sphere while the parameters of the negative modalities (e.g., m_1 and m_2) are constrained to be scattered and far away from $\mathbf{W}_{z^k}$.

Activity Classification Loss. We use cross-entropy loss to learn the shared activity classifier and private classifier:

$$\mathcal{L}_{sc}(\mathcal{D}^k) = -\sum_{i=1}^{n_k}\mathbf{y}_i^k\log\hat{\mathbf{y}}_i^k, \mathcal{L}_{pc}(\mathcal{D}^k) = -\sum_{i=1}^{n_k}\mathbf{y}_i^k\log\check{\mathbf{y}}_i^k, \quad (8)$$

where $\hat{\mathbf{y}}_i^k$ and $\check{\mathbf{y}}_i^k$ are predictions: $\hat{\mathbf{y}}_i^k = \psi_{sc}(\hat{\mathbf{h}}_i^k;\mathbf{\Theta}_{sc})$, $\check{\mathbf{y}}_i = \psi_{pc}(\check{\mathbf{h}}_i^k;\mathbf{\Theta}_{pc}^{z_k})$. $\mathbf{y}_i^k$ is the ground-truth activity label. Finally, the activity classification loss is defined as $\mathcal{L}_c = \mathcal{L}_{sc}+\mathcal{L}_{pc}$.

Federated Optimization

We train the FDARN with two alternate steps of local update and global aggregation for T rounds. In each round, the client receives a global model and updates it for ε epochs on local data while the server receives the updated local models from different clients and updates the global model with weighted averaging. Before introducing the optimization details, we define $\mathbf{\Theta}_*^m = \{\mathbf{\Theta}_a^m,\mathbf{\Theta}_e^m,\mathbf{\Theta}_{pc}^m\}$ for simplicity, where $m\in\{1,...,M\}$ is the modality label.

Adversarial Local Update. For the k-th client, the server firstly sends the initial modal parameters $\{\bar{\mathbf{\Theta}}_{sc},\bar{\mathbf{\Theta}}_{md},\bar{\mathbf{\Theta}}_*^{z^k}\}$ before local updating. Then, we will update these parameters based on the local data. It is worth noting that the modality discriminator plays two contrary roles in learning the FDARN. On the one hand, it is used to learn an altruistic encoder that can map instances from different modalities into a common feature space. On the other hand, it is also used to encourage an egocentric encoder to produce modality-specific features. To achieve the adversarial goal, we update the FDARN on the local data as follows:

$$(\mathbf{\Theta}_{sc}^k,\mathbf{\Theta}_*^{z^k,k}) = (\bar{\mathbf{\Theta}}_{sc},\bar{\mathbf{\Theta}}_*^{z^k}) - \eta\nabla_{(\bar{\mathbf{\Theta}}_{sc},\bar{\mathbf{\Theta}}_*^{z^k})}\mathcal{L}(\mathcal{D}^k,z^k), \quad (9)$$

$$\mathbf{\Theta}_{md}^k = \bar{\mathbf{\Theta}}_{md} - \eta\nabla_{\bar{\mathbf{\Theta}}_{md}}\mathcal{L}'_{md}(\mathcal{D}^k,z^k), \quad (10)$$

where $\mathcal{L} = \mathcal{L}_c+\gamma_1\mathcal{L}_{sep}+\gamma_2\mathcal{L}_{md}$ and $\mathcal{L}'_{md} = \mathcal{L}_{sp}+\mathcal{L}_{ama}+\mathcal{L}_{ams}$. η is learning rate. This adversarial local updating can be simply implemented by adding a gradient reversal layer (GRL) (Ganin and Lempitsky 2015) between the altruistic encoder and the modality discriminator.

Modality-aware Global Aggregation. Now, we introduce how to collect the local modals learned on different clients to build a global model.

The shared activity classifier ψ_{sc} is learned on features that are invariant across different modalities. Therefore, its parameters learned on one client can also be shared by other clients. We aggregate the local parameters of the shared classifier from all clients with weighted average as in FedAvg (McMahan et al. 2017). For the modality discriminator, to ensure that it can consistently identify the modality labels of all instances from different clients, we also take the same parameter aggregation scheme:

$$\bar{\mathbf{\Theta}}_{sc} = \sum\nolimits_{k=1}^{K}\frac{n_k}{n}\mathbf{\Theta}_{sc}^k, \quad \bar{\mathbf{\Theta}}_{md} = \sum\nolimits_{k=1}^{K}\frac{n_k}{n}\mathbf{\Theta}_{md}^k. \quad (11)$$

For the altruistic encoder ϕ_a, the egocentric encoder ϕ_e and the private activity classifier ϕ_{pc}, we only make the clients with the same modality share common parameters:

$$\bar{\mathbf{\Theta}}_*^m = \sum_{k\in\mathcal{K}_m}\frac{n_k}{\sum\limits_{k'\in\mathcal{K}_m}n_{k'}}\mathbf{\Theta}_*^{m,k}, m=1,...,M, \quad (12)$$

where $\mathcal{K}_m$ is a set of clients that have the modality label m.

Inference

In the test time, each $\mathbf{x}$ on the k-th client is classified by: $\bar{f}^k(\mathbf{x}) = \psi_{sc}(\phi_a(\mathbf{x}; \bar{\mathbf{\Theta}}_a^{z_k}); \bar{\mathbf{\Theta}}_{sc}) + \psi_{pc}(\phi_e(\mathbf{x}; \bar{\mathbf{\Theta}}_e^{z_k}); \bar{\mathbf{\Theta}}_{pc}^{z_k})$.

Experiment

Datasets and Evaluation Metrics

Epic-Kitchens. This is the largest public multimodal dataset in egocentric HAR (Damen et al. 2020). The Epic-Kitchens-100 has 89977 video segments of human-object interaction captured by 37 participants. 16 participants also provide audio and sensor data. For the CMF-HAR task, we use the unique 97 verb labels as activity classes and consider four modalities (i.e., video, optical flow, audio and sensor), where the optical flow is extracted from the video. We randomly select 4 participants for each modality and treat each participant as a local client, where each client only retains the data of its corresponding modality. Finally, we obtain a 16-client 97-class 4-modality CMF-HAR task with 34018 instances.

Multimodal-EA. This is an earlier multimodal dataset for egocentric HAR (Song et al. 2016b), which contains 50 minutes of videos with sensor signals of 20 activities. Since the user information is not recorded, we randomly split this dataset into 4 clients, where two clients have 100 instances of the video modality and the other two have 100 instances of the sensor modality. Then, we obtain a 4-client 20-class 2-modality CMF-HAR task with 400 instances.

Stanford-ECM. This dataset contains 31 hours of egocentric videos with sensor signals of 23 activities (Nakamura et al. 2017). The video duration ranges from 2 minutes to 51 minutes, and each video may contain multiple activities. We divide each video into multiple instances as in (Huang et al. 2021) so that each instance has a unique activity label. Since the user information is not recorded, we randomly split this dataset into 10 clients, where 5 clients have 112 instances of the video modality and the other 5 clients have 112 instances of the sensor modality. Then, we obtain a 10-client 23-class 2-modality CMF-HAR task with 1120 instances.

Ego-Exo-AR. This is a new multimodal HAR dataset collected by ourselves. It contains sensor signals captured by smart bracelets and images captured by third perspective phone cameras. It is collected by 14 participants for one month and each participant has 465 instances of 15 daily activities. Similar to the Epic-Kitchens, we treat each participant as a client and randomly select 7 clients for each modality. Finally, we obtain a 14-client 15-class 2-modality CMF-HAR task with 6510 instances, where each client only retains the data of its corresponding modality. It is worth noting that the image data and the sensor data of the Ego-Exo-AR are captured in different perspectives while the vision and sensor data of the above three public datasets are all from the egocentric perspective. Therefore, our Ego-Exo-AR dataset has larger heterogeneity between different modalities and brings more challenges to the CMF-HAR.

Training/Test Splitting. For all the above four datasets, we randomly split local instances on each client into the training and test sets with a ratio of $0.75 : 0.25$.

Evaluation Metrics. Following the existing FL-based HAR methods (Li et al. 2021a), we adopt accuracy as the evaluation metric. More specifically, we firstly compute the accuracy for each client, and then average the results of different clients. We repeat the training and testing process 5 times and report mean accuracy and standard deviation.

Baselines

We compare our model with seven state-of-the-art FL algorithms: **FedAvg** (McMahan et al. 2017), **pFedMe** (Dinh, Tran, and Nguyen 2020), **FedProx** (Li et al. 2020), **PerAvg** (Fallah, Mokhtari, and Ozdaglar 2020), **LG-FedAvg** (Liang et al. 2020), **FedRep** (Collins et al. 2021), **FedBN** (Li et al. 2021b), and one closely related FL-based HAR method: **Meta-HAR** (Li et al. 2021a). We also compare with a baseline **SingleSet**, which trains a local model (i.e., only the egocentric encoder and the private classifier of our method are used) for each client without using FL.

Implementation Details

For the four datasets in our experiment, there are 5 unique modalities (i.e., video, optical flow, audio, sensor, image). To facilitate the fair comparison with existing methods, we firstly extract the raw features with the same dimension (i.e., 1024) for different modalities. Then, the raw features will be input to our model or baselines to conduct the CMF-HAR task. It is worth noting that using the same dimension features of different modalities is not a mandatory requirement of our method in practice.

The overall framework of our method is implemented with Pytorch (Paszke et al. 2019). For all baselines, we use the publicly released code. Our model and baselines are all trained with SGD optimizer, where the weight decay is set to $1e$-5 and the momentum is set to 0.9. On the Epic-Kitchens, the learning rate η of the local client is set to 0.001 and the batch size is set to 64. On the other three datasets, the learning rate and the batch size are set to 0.01 and 32, respectively. On all four datasets, the number of local epochs ε is set to 2, and the number of communication rounds T is 300. For the SingleSet baseline, the number of local epochs is set to 300. Unless explicitly specified, other hyper-parameters of each baseline are tuned within the range provided by the authors and the best results are reported.

In our method, both the altruistic encoder and the egocentric encoder are implemented as two-layer perceptrons with the activation function of ReLU, where the dimension of the hidden layer and the output dimension d are all set to 1024. For the shared activity classifier and the private classifier, the dimension of the hidden layer is set to 1024. For the modality discriminator, the output dimension $\tilde{d}$ of the single-layer perceptron φ is set to 128. The margin factor τ and scale factor α of the additive angular margin loss in Eq. (4) are set to 0.5 and 72. The margin factor ν of the spreadout regularizer in Eq. (7) is set to 1.5. The balance weights γ_1 and γ_2 of the loss function are set to 0.6 and 0.4.

Comparison with State-of-the-art Methods

Table 1 shows the average accuracy of the clients with the same modality. Overall, the proposed FDARN performs better than all baselines, which demonstrates that our model can effectively mitigate the cross-modal heterogeneity. On the Epic-Kitchens, our model has improvements of 2.00%,

Methods	Epic-Kitchens				Multimodal-EA		Stanford-ECM		Ego-Exo-AR	
	Video	Flow	Audio	Sensor	Video	Sensor	Video	Sensor	Image	Sensor
SingleSet	29.91(0.3)	40.07(0.6)	44.80(0.3)	30.12(0.4)	53.60(2.0)	32.80(1.6)	49.24(1.8)	27.79(0.9)	34.69(0.9)	31.34(0.8)
FedAvg	30.57(0.2)	36.76(0.5)	42.73(0.4)	28.90(0.5)	57.60(0.8)	31.20(0.9)	57.14(1.1)	29.74(1.2)	33.93(0.7)	32.96(0.9)
pFedMe	26.47(0.5)	31.73(0.9)	36.03(0.6)	26.97(0.4)	48.80(2.4)	28.40(2.0)	53.61(1.8)	25.49(1.4)	30.90(1.1)	28.71(1.2)
FedProx	32.16(0.5)	34.64(0.7)	41.37(0.5)	27.46(0.5)	57.60(0.8)	26.80(1.7)	54.37(1.3)	31.74(1.1)	34.24(1.0)	31.76(0.8)
PerAvg	30.95(0.4)	40.59(0.4)	42.67(0.3)	29.61(0.3)	62.40(0.8)	36.40(0.8)	60.86(1.0)	30.18(1.6)	37.53(0.8)	34.10(0.8)
LG-FedAvg	30.68(0.5)	41.25(0.5)	40.25(0.5)	30.51(0.6)	56.40(1.0)	31.60(1.7)	55.12(1.4)	29.87(1.4)	35.09(0.9)	32.23(0.8)
FedRep	31.93(0.4)	40.89(0.5)	43.12(0.5)	33.01(0.6)	64.80(1.7)	40.00(1.3)	61.09(1.7)	36.27(1.0)	40.02(0.6)	36.15(0.7)
FedBN	32.18(0.4)	38.98(0.4)	44.95(0.2)	32.87(0.4)	64.00(0.0)	42.00(1.9)	59.24(1.3)	33.37(0.8)	40.34(0.5)	35.37(0.7)
Meta-HAR	31.27(0.3)	39.90(0.6)	45.26(0.4)	**34.62(0.5)**	57.60(1.7)	42.00(1.3)	56.61(1.3)	33.49(1.6)	36.64(0.8)	37.23(0.8)
FDARN	**34.18(0.5)**	**43.57(0.4)**	**47.52(0.4)**	33.90(0.5)	**68.40(0.8)**	**42.80(1.7)**	**65.62(1.8)**	**38.49(1.0)**	**42.38(0.6)**	**38.31(0.7)**

Table 1: Comparison with state-of-the-art methods on four datasets.

Methods	Epic-Kitchens			
	Video	Flow	Audio	Sensor
SingleSet	22.35(0.9)	37.04(1.7)	39.91(0.4)	25.14(0.3)
FedAvg	23.37(0.4)	41.50(1.3)	39.26(0.4)	25.93(0.2)
pFedMe	22.53(0.8)	36.20(1.1)	37.05(0.5)	24.92(0.6)
FedProx	23.73(0.3)	39.27(0.6)	39.86(0.1)	26.18(0.4)
PerAvg	24.39(0.3)	42.88(1.2)	40.03(0.2)	27.26(0.4)
LG-FedAvg	21.76(0.4)	40.12(0.7)	36.54(0.3)	25.91(0.5)
FedRep	24.56(0.2)	42.09(0.6)	37.63(0.3)	29.22(0.3)
FedBN	21.55(0.5)	41.16(0.6)	38.71(0.4)	28.71(0.3)
Meta-HAR	22.47(0.4)	41.96(0.8)	38.27(0.6)	27.74(0.3)
FDARN	**26.86(0.2)**	**43.23(0.8)**	**40.60(0.4)**	**31.01(0.3)**

Table 2: Comparison with state-of-the-art methods under the 4-client 97-class 4-modality CMF-HAR task.

Methods	Epic-Kitchens	Multimodal-EA	Stanford-ECM	Ego-Exo-AR
w/o ψ_{sc}	33.76(0.7)	44.60(2.0)	42.86(1.4)	35.43(1.2)
w/o ψ_{pc}	39.12(0.4)	54.00(0.6)	50.47(0.4)	39.17(0.4)
w/o (ϕ_a, ψ_{sc})	35.96(0.6)	49.00(0.6)	48.13(0.3)	37.27(0.5)
w/o (ϕ_e, ψ_{pc})	37.77(0.5)	53.20(0.4)	49.24(0.5)	37.86(0.7)
w/o ψ_{md}	36.10(0.2)	51.20(0.4)	47.51(0.3)	36.12(0.3)
w/o $\mathcal{L}_{sep}$	37.24(0.2)	50.80(1.1)	50.21(0.3)	38.38(0.6)
w/o $\mathcal{L}_{sp}$	38.57(0.3)	54.80(0.4)	51.05(0.5)	39.47(0.4)
FDARN	**39.80(0.5)**	**55.60(1.2)**	**52.05(1.4)**	**40.35(0.7)**

Table 3: Ablation studies on four datasets.

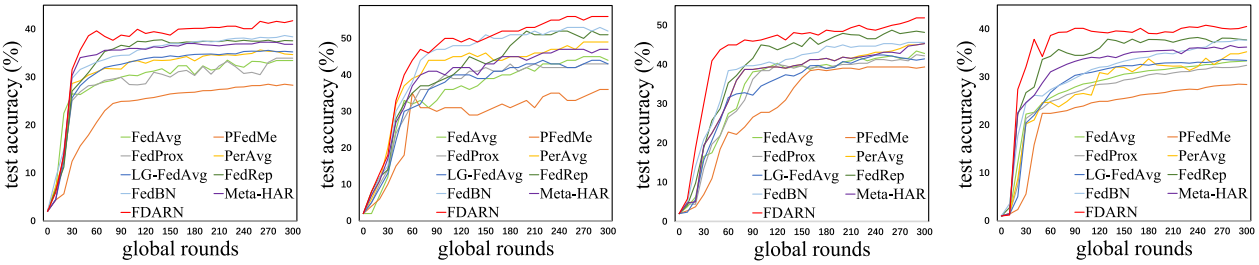
2.32% and 2.26% over the second best method on the video, optical flow and audio modality, respectively. For the sensor modality, although our method cannot outperform the Meta-HAR, we still achieve competitive results. On the Multimodal-EA, Stanford-ECM and Ego-Exo-AR datasets, our method performs consistently better than all baselines on different modalities. It is worth noting that the Meta-HAR has better performances than all other baselines for the sensor data on the Epic-Kitchens and Ego-Exo-AR. This is due to that each user has different distribution of the sensor signal for the same activity class (Li et al. 2021a) on these two datasets while the Meta-HAR can effectively reduce the distribution gap by a meta-learned embedding network. However, the inferior performances of the Meta-HAR on other modalities demonstrate that it cannot be directly applied to diminish the cross-modal heterogeneity.

To further evaluate the effectiveness of our model in mitigating the cross-modal heterogeneity, we conduct an extra experiment on the Epic-Kitchens under a more challenging 4-client 97-class 4-modality CMF-HAR task, where we randomly select one client for each modality and ensure that no two clients have the same modality. The results in Table 2 show that our model consistently outperforms all the baselines on different modalities.

Ablation Studies

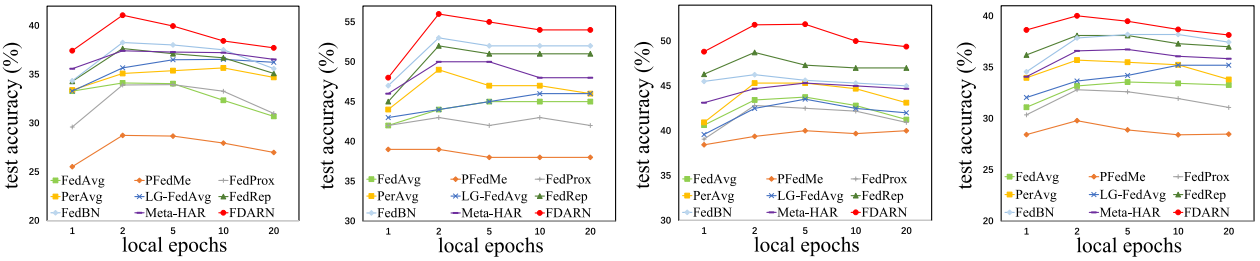
Here, we present the results for several variants of our model to demonstrate the effectiveness of the primary modules in our method. We use w/o ψ_{sc} to denote the variant that does not use the shared classifier, i.e., we learn disentangled features, but only the private activity classifier is adopted for classification. w/o ψ_{pc} denotes the variant that does not use the private classifier, i.e., we learn disentangled features, but only the shared activity classifier is adopted for classification. w/o (ϕ_a, ψ_{sc}) denotes the variant that does not use the altruistic encoder and shared classifier, i.e., we distributively learn the egocentric encoder and private classifier across clients with the same modality without considering the knowledge sharing across different modalities. w/o (ϕ_e, ψ_{pc}) denotes the variant that does not use the egocentric encoder and private classifier, i.e., we only use the altruistic encoder and modality discriminator to learn modality-agnostic features that are further used to train the shared activity classifier. w/o ψ_{md} denotes the variant that does not use the modality discriminator. w/o $\mathcal{L}_{sep}$, w/o $\mathcal{L}_{sp}$ denotes the variants that do not use the separation loss and the spreadout regularizer, respectively.

Table 3 shows the average accuracy of all clients on each dataset. As shown, the w/o ψ_{sc} has the worst performance, which demonstrates that the shared classifier has better discriminative ability than the private classifier. The w/o ψ_{pc} performs slightly worse than the FDARN, which demonstrates that the modality-specific information is still complementary to the shared classifier. The w/o (ϕ_e, ψ_{pc}) performs much worse than the w/o ψ_{pc}, which demonstrates that the proposed feature disentanglement is more effective in learning client-shared features across different modalities than directly embedding all instances from different clients into a common space. The w/o ψ_{sc} performs even worse than the w/o (ϕ_a, ψ_{sc}), which demonstrates that the feature



(a) Epic-Kitchens (b) Multimodal-EA (c) Stanford-ECM (d) Ego-Exo-AR

Figure 3: Effect of the number of global rounds.



(a) Epic-Kitchens (b) Multimodal-EA (c) Stanford-ECM (d) Ego-Exo-AR

Figure 4: Effect of the number of local epochs.

disentanglement may tend to allocate more discriminative information into the modality-agnostic features than into the modality-specific features. The $\texttt{w/o}\ \psi_{md}$ decreases the performance by about 4.00% on each dataset, which demonstrates the effectiveness of the proposed modality discriminator. The decreased performances of the $\texttt{w/o}\ \mathcal{L}_{sep}$ and $\texttt{w/o}\ \mathcal{L}_{sp}$ show the necessity of the separation loss and the spread-out regularizer, respectively.

Further Remarks

Number of Communication Rounds. Figure 3 shows the average test accuracy of all clients with different number of communication rounds. With a small number of rounds (e.g., less than 20 on the Epic-Kitchens), our model has similar performance as the baselines, e.g., PerAvg, FedBN, and Meta-HAR. Thanks to the proposed feature disentanglement scheme, our model achieves consistently better accuracy than all baselines after more rounds of training (e.g., about 40 rounds on the Epic-Kitchens).

Number of Local Epochs. Figure 4 shows the effect of the number of local update epochs. Our model has the best results when the number of local epochs is 2, which is consistent with most of the baselines. With a smaller number of local updates, the training speed will slow down and will impact the final accuracy in a fixed number of communication rounds. In contrast, with a larger number of local updates, it is difficult to find a globally consistent HAR model.

Spherical Modality Discriminative Loss. In the proposed FDARN, we adopt the spherical modality discriminative loss $\mathcal{L}_{md}$ to learn the modality discriminator for disentangling modality-agnostic and modality-specific features. To show its effectiveness, we evaluate the performance of our model when replacing $\mathcal{L}_{md}$ with cross-entropy loss (CE-Loss). As shown in Table 4, the CE-Loss decreases the performance by at least 2.00% on each dataset. If we further remove the separation loss $\mathcal{L}_{sep}$ and simply use the modality discriminator learned with cross-entropy loss (CE-Loss*) to constrain the feature disentanglement, the performances are much worse than FDARN* that simply uses the modality discriminator learned with $\mathcal{L}_{md}$. Because the local client only has instances of one modality and the conventional cross-entropy loss cannot effectively capture the inter-class discrepancy in this case.

Methods	Epic-Kitchens	Multimodal-EA	Stanford-ECM	Ego-Exo-AR
CE-Loss	37.75(0.3)	53.60(0.5)	49.24(0.3)	37.26(0.5)
FDARN	**39.80(0.5)**	**55.60(1.2)**	**52.05(1.4)**	**40.35(0.7)**
CE-Loss*	34.55(0.3)	47.60(0.4)	45.56(0.2)	34.56(0.3)
FDARN*	**37.24(0.2)**	**50.80(1.1)**	**50.21(0.3)**	**38.38(0.6)**

Table 4: Results with different kinds of loss for learning the modality discriminator.

Conclusion

In this paper, we propose a feature-disentangled activity recognition network to solve the new task of cross-modal federated HAR by embedding instances on different local clients into a modality-agnostic feature space and producing the modality-specific feature that cannot be shared across clients with different modalities. Through decentralized optimization with a spherical modality discriminative loss, our model obtains state-of-the-art results by combining the shared activity classifier and the private activity classifier that are learned on the modality-agnostic features and the modality-specific features, respectively.

Acknowledgments

This work was supported by National Key Research and Development Program of China (No. 2018AAA0100604), National Natural Science Foundation of China (No. 61720106006, 62036012, 62072455, 61721004, U1836220, 61872424), Beijing Natural Science Foundation (L201001).

References

Arivazhagan, M. G.; Aggarwal, V.; Singh, A. K.; and Choudhary, S. 2019. Federated Learning with Personalization Layers. *CoRR*, abs/1912.00818.

Bercea, C. I.; Wiestler, B.; Rueckert, D.; and Albarqouni, S. 2021. FedDis: Disentangled Federated Learning for Unsupervised Brain Pathology Segmentation. *CoRR*, abs/2103.03705.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain Separation Networks. In *NIPS*, 343–351.

Bulling, A.; Blanke, U.; and Schiele, B. 2014. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Comput. Surv.*, 46(3): 33:1–33:33.

Collins, L.; Hassani, H.; Mokhtari, A.; and Shakkottai, S. 2021. Exploiting Shared Representations for Personalized Federated Learning. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 2089–2099. PMLR.

Damen, D.; Doughty, H.; Farinella, G. M.; Furnari, A.; Kazakos, E.; Ma, J.; Moltisanti, D.; Munro, J.; Perrett, T.; Price, W.; and Wray, M. 2020. Rescaling Egocentric Vision. *CoRR*, abs/2006.13256.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 4690–4699. Computer Vision Foundation / IEEE.

Dinh, C. T.; Tran, N. H.; and Nguyen, T. D. 2020. Personalized Federated Learning with Moreau Envelopes. In *NeurIPS*.

Ek, S.; Portet, F.; Lalanda, P.; and Vega, G. 2020. Evaluation of federated learning aggregation algorithms: application to human activity recognition. In Tentori, M.; Weibel, N.; Laerhoven, K. V.; Abowd, G. D.; and Salim, F. D., eds., *UbiComp/ISWC '20: 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and 2020 ACM International Symposium on Wearable Computers, Virtual Event, Mexico, September 12-17, 2020*, 638–643. ACM.

Fallah, A.; Mokhtari, A.; and Ozdaglar, A. E. 2020. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. In *NeurIPS*.

Ganin, Y.; and Lempitsky, V. S. 2015. Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 1180–1189.

Guo, W.; Huang, H.; Kong, X.; and He, R. 2019. Learning Disentangled Representation for Cross-Modal Retrieval with Deep Mutual Information Estimation. In *ACM Multimedia*, 1712–1720. ACM.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR (Poster)*. OpenReview.net.

Huang, Y.; Yang, X.; Gao, J.; Sang, J.; and Xu, C. 2021. Knowledge-driven Egocentric Multimodal Activity Recognition. *ACM Trans. Multim. Comput. Commun. Appl.*, 16(4): 133:1–133:133.

Jain, P.; Goenka, S.; Bagchi, S.; Banerjee, B.; and Chaterji, S. 2021. Federated Action Recognition on Heterogeneous Embedded Devices. *CoRR*, abs/2107.12147.

Kim, H.; and Mnih, A. 2018. Disentangling by Factorising. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, 2654–2663. PMLR.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Li, C.; Niu, D.; Jiang, B.; Zuo, X.; and Yang, J. 2021a. Meta-HAR: Federated Representation Learning for Human Activity Recognition. In Leskovec, J.; Grobelnik, M.; Najork, M.; Tang, J.; and Zia, L., eds., *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, 912–922. ACM / IW3C2.

Li, Q.; He, B.; and Song, D. 2021. Model-Contrastive Federated Learning. In *CVPR*, 10713–10722. Computer Vision Foundation / IEEE.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated Optimization in Heterogeneous Networks. In *MLSys*. mlsys.org.

Li, X.; Jiang, M.; Zhang, X.; Kamp, M.; and Dou, Q. 2021b. FedBN: Federated Learning on Non-IID Features via Local Batch Normalization. In *ICLR*. OpenReview.net.

Liang, P. P.; Liu, T.; Liu, Z.; Salakhutdinov, R.; and Morency, L. 2020. Think Locally, Act Globally: Federated Learning with Local and Global Representations. *CoRR*, abs/2001.01523.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, 1273–1282. PMLR.

Nakamura, K.; Yeung, S.; Alahi, A.; and Fei-Fei, L. 2017. Jointly Learning Energy Expenditures and Activities Using Egocentric Multimodal Signals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 6817–6826. IEEE Computer Society.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional Image Synthesis with Auxiliary Classifier GANs. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 2642–2651. PMLR.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 8024–8035.

Possas, R.; Pinto-Caceres, S. M.; and Ramos, F. 2018. Egocentric Activity Recognition on a Budget. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 5967–5976. IEEE Computer Society.

Reisizadeh, A.; Farnia, F.; Pedarsani, R.; and Jadbabaie, A. 2020. Robust Federated Learning: The Case of Affine Distribution Shifts. In *NeurIPS*.

Simonyan, K.; and Zisserman, A. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 568–576.

Song, S.; Chandrasekhar, V.; Mandal, B.; Li, L.; Lim, J.; Babu, G. S.; San, P. P.; and Cheung, N. 2016a. Multimodal Multi-Stream Deep Learning for Egocentric Activity Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2016, Las Vegas, NV, USA, June 26 - July 1, 2016*, 378–385. IEEE Computer Society.

Song, S.; Cheung, N.; Chandrasekhar, V.; Mandal, B.; and Lin, J. 2016b. Egocentric activity recognition with multimodal fisher vector. In *ICASSP*, 2717–2721. IEEE.

Sozinov, K.; Vlassov, V.; and Girdzijauskas, S. 2018. Human Activity Recognition Using Federated Learning. In *IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking, Sustainable Computing & Communications, ISPA/IUCC/BDCloud/SocialCom/SustainCom 2018, Melbourne, Australia, December 11-13, 2018*, 1103–1111. IEEE.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial Cross-Modal Retrieval. In Liu, Q.; Lienhart, R.; Wang, H.; Chen, S. K.; Boll, S.; Chen, Y. P.; Friedland, G.; Li, J.; and Yan, S., eds., *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, 154–162. ACM.

Wang, H.; Yurochkin, M.; Sun, Y.; Papailiopoulos, D. S.; and Khazaeni, Y. 2020. Federated Learning with Matched Averaging. In *ICLR*. OpenReview.net.

Wu, X.; Huang, H.; Patel, V. M.; He, R.; and Sun, Z. 2019. Disentangled Variational Representation for Heterogeneous Face Recognition. In *AAAI*, 9005–9012. AAAI Press.

Xu, Y.; Zhang, L.; and Duan, Q. 2020. Domain Private and Agnostic Feature for Modality Adaptive Face Recognition. In *IJCB*, 1–9. IEEE.

Yu, F. X.; Rawat, A. S.; Menon, A. K.; and Kumar, S. 2020. Federated Learning with Only Positive Labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, 10946–10956. PMLR.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated Learning with Non-IID Data. *CoRR*, abs/1806.00582.

Zong, L.; Xie, Q.; Zhou, J.; Wu, P.; Zhang, X.; and Xu, B. 2021. FedCMR: Federated Cross-Modal Retrieval. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, 1672–1676. ACM.