Spectrum-Guided Adversarial Disparity Learning

Anonymous Author(s)

ABSTRACT

It has been a significant challenge to portray intraclass disparity precisely in the area of activity recognition, as it requires a robust representation of the correlation between subject-specific variation for each activity class. In this work, we propose a novel end-to-end knowledge directed adversarial learning framework, which portrays the class-conditioned intraclass disparity using two competitive encoding distributions and learns the purified latent codes by denoising learned disparity. Furthermore, the domain knowledge is incorporated in an unsupervised manner to guide the optimization and further boost the performance. The experiments on four HAR benchmark datasets demonstrate the robustness and generalization of our proposed methods over a set of state-of-the-art. We further prove the effectiveness of automatic domain knowledge incorporation in performance enhancement.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *Neural networks*; *Learning latent representations.*

KEYWORDS

Intraclass Variability, Adversarial Autoencoder, Activity Recognition, Generative Models

1 INTRODUCTION

Sensor technologies have inspired a wide range of applications that help people's daily lives in many realms, such as visual recognition [5], brain-computer interface [8, 20], and activity recognition [26, 30]. One of the major components of sensing applications is to establish commonly-used and robust systems over diverse scenarios. The intraclass disparity can be caused by the variety of subjects' temporal conditions and unique physical characters: people have varying habits and body shapes; they may make different movements when performing the same activities, and sensors may perceive differently, given the same person performing the same activities. We call such variability within a class *intraclass disparity*. The intraclass disparity will significantly impair the systems' performance in dealing with new subjects or new environments.

The typical work to address this challenge [5, 6] constructs a dictionary or a projection based on the existing subjects to conjecture the possible variations that may occur in less-seen subjects; the models can thus gain robustness through embracing the variation during training. However, such studies primarily rely on the number of known subjects, which we call subject-dependent studies. Therefore, the model may suffer from significant performance degradation when handling new subjects, due to the proprietary characteristics in existing subjects.

In light of generative models' excellent performance on sparse data, recent studies are increasingly applying generative networks to improve models' robustness on new subjects. Given that generative models usually perform better on sparse data, However, most subject-independent studies [1, 22, 27, 32] are still limited in considering the intraclass disparity as meaningless noise, which neglect the point that intraclass disparity is related to the subject and the class type. They are still inaccurate in exhibiting the relationship between the subject variation and the class, e.g., subject variation within a class should be conditionally constrained.

Besides, signal data may be segmented imprecisely, and the segments may include gaps and noises. Further, the segments carry nonequivalent amounts of information, which is difficult to measure in the time domain. Since any signal wave in the time domain also presents as a variable in the frequency domain, *Spectrum* represents the corresponding frequency composition of signals. Thus, it is promising to use frequency-domain features (e.g., amplitude and entropy) [3] as the domain knowledge to analyze the segments. Take amplitude spectrum [29] for example. Valid information frequencies offer peak amplitudes than gaps and noises. We can thereby measure the amounts of information and mitigate the impact of imprecise segmentation. The only issue is the heuristic, hand-engineered, case-specific nature of signal domain analysis, which limits its applicability.

To address the limitations in existing intraclass disparity research, we propose a novel Spectrum-guided Adversarial Autoencoder (SAAE) for Human Activity Recognition (HAR). HAR is challenging due to its low signal-to-noise ratio. SAAE utilizes two competitive encoding distributions to make intraclass disparity class-conditioned and further eliminates the subject variation through a denoising structure. This way, the purified latent codes will be robust to handle new subjects. The advantages of SAAE are two-fold. First, SAAE applies a learnable prior distribution, which enables posterior distribution to learn class-related distribution. The posterior distribution learns to generate the intraclass disparity with the class information in prior distribution, while the discriminator and decoder further supervise the prior distribution to learn the pure information by fixing posterior distribution. Second, SAAE automates spectrum analysis as domain knowledge, leading to enhanced robustness of the data segmentation. The proposed automatic spectrum score function is capable of dynamically weighting signals and then adjusting the optimization of adversarial training.

We make several contributions in this paper:

- We introduce the signal theory and design a principled unsupervised score function to weight signals dynamically. The score function analyzes signals in the frequency domain to measure the signal information amount and to provide domain knowledge in optimization.
- We propose a novel Spectrum-guided Adversarial Autoencoder (SAAE), which fuses automating spectrum analysis and adversarial training in a unified network. SAAE utilizes two competitive encoding distributions supervised by prediction validity to extract and denoise the learned intraclass disparity. We also analyze and prove the effectiveness of intraclass disparity learning through competitive encoding distributions training.

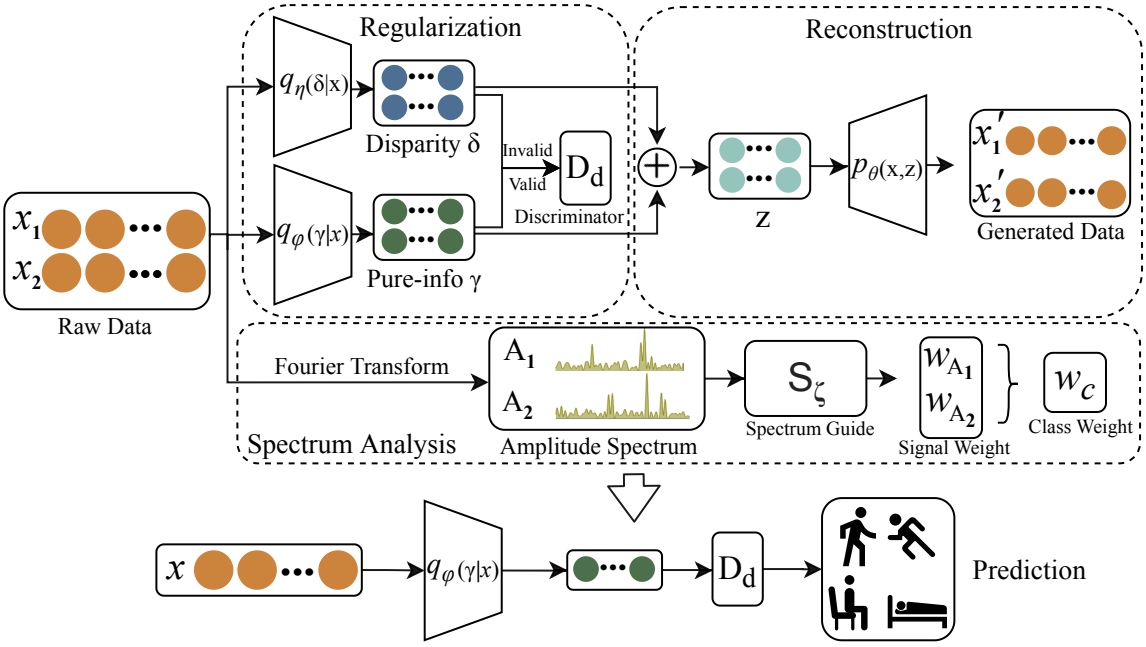


Figure 1: Illustration of SAAE. First, we take two samples of the same class to explain the three phases in optimization: signal weight measurement in spectrum analysis, decomposition in the regularization phase, and denoising in the reconstruction phase. *Spectrum analysis* (in the middle) transforms signals and projects the corresponding amplitude spectra into signal weights W_A for reconstruction by S_ζ. The class weight W_c is summarized from all signals under a class to assist regularization. *Decomposition* feeds the disparity δ (labeled 'invalid') and pure information γ (labeled 'valid') to the discriminator to decompose the original data. *Denoising* takes δ as noises and uses an autoencoder to denoise the disparity components. Then, after optimization, the encoding distribution $a_\varphi(\gamma|x)$ could extract information γ purified from intraclass disparity δ, which enables discriminator D to precisely predict the activities.

- We compare SAAE with both state-of-the-art adversarial autoencoders and human activity-related algorithms in the subject-independent experiments. The superior performance of SAAE on four benchmark datasets demonstrates its effectiveness in solving the intraclass disparity problem on unknown subjects. We further exhibit the effectiveness of intraclass disparity learning through convergence analysis and spectrum analysis.

2 METHODOLOGY

This section introduces the methodology of AAE for class-conditioned intraclass disparity and domain knowledge assistance. AAE consists of two phases: intraclass disparity extraction and denoising. AAE leverages the domain knowledge obtained from frequency-domain analysis and an automated spectrum guide function for optimization.

2.1 Intraclass Disparity Learning

This section first declares the class conditioned intraclass disparity definition, and then propose a two-phase AAE which utilizes two competitive encoding distributions to handle the intraclass disparity learning. We discuss the *Reconstruction phase* and the *Regularization phase* of AAE from the generative perspective and the discriminator's perspective, respectively. We further prove AAE's effectiveness in learning and denoising intraclass disparity.

2.1.1 Notations for Adversarial Autoencoder. Following the notations in Kingma et al.'s work [16], we denote by $p_{data}(x)$ the observed data distribution, $p(x)$ the model distribution of reconstructing data, $q(z|x)$ the encoding distribution, and $p(x|z)$ the decoding distribution. The distribution $q(z|x)$ encodes the original data into latent codes z, which denotes the latent representation for data reconstruction. Suppose the intraclass disparity δ confuses pure information γ in a linear relationship, we define the latent representation z is composed by

$$\gamma \sim q_\varphi(\gamma|x), \quad \delta \sim q_\eta(\delta|x)$$
$$z := \gamma + \delta$$

where $q_\varphi(\gamma|x), q_\eta(\delta|x)$ denote the pure information and disparity encoding distributions, respectively.

2.1.2 Reconstruction Phase. In this phase, we use the denoising autoencoder to denoise disparity and reconstruct the signal waveform. We consider the signal waveform reconstruction with the following requirements: (i) the decoder should be able to reconstruct original data by $z = \gamma + \delta$; (ii) γ should be able to learn the representative and invariant waveform information; (iii) δ should be the confusion factor in waveform only related to class information.

Given the optimal disparity δ, the reconstruction can be demonstrated as a denoising autoencoder from such disparity. The autoencoder fuses the learned disparity δ with raw data as inputs and then train autoencoder to recover the original data. Thus, the model will be less sensitive to the disparity components. In conventional denoising autoencoders [28], the disparity part in original data is vague; thus, they fuse the noise in raw data (e.g., Gaussian noise). Different from the meaningless noise, the proposed disparity encoder $q_\eta(\delta|x)$ can extract the disparity components, which will be proved to be the disparity with class information in the regularization phase by the adversarial training. Therefore, we move the fusion process after encoding.

The requirement (i)-(iii) can be explained with the following definitions. Let γ be pure latent codes and fix the optimal disparity distribution to let δ be the constant noise. Then, the purified latent code $z = \gamma$ and the confused code $z^{'} = \gamma + \delta$. The condition probability distribution of $z^{'}$ on x is $q_\varphi(z^{'}|x)$. Consider an approximate posterior distribution between x and z:

$$q_\varphi(z|x) = \int_{z^{'}} c(z|z^{'}) q_\varphi(z^{'}|x) dz^{'}$$

where $c(z|z^{'})$ is the observed probability distribution between confused codes and pure codes. Then, given the decoding distribution $p_\theta(x, z) = p_\theta(x|z)p(z)$, the lower bound of the autoencoder may be formed in the following way [13] by Jensen's inequality:

$$\begin{aligned} log p_\theta(x) &= log E_{q_\varphi(z^{'}|x)} E_{c(z|z^{'})}[\frac{p_\theta(x,z)}{q_\varphi(z|z^{'})}] \\ &\geq E_{q_\varphi(z^{'}|x)} E_{c(z|z^{'})}[log \frac{p_\theta(x,z)}{q_\varphi(z|z^{'})}] \end{aligned}$$

Therefore, the lower bound of autoencoder can be sorted into minimizing the negative weighted likelihood L_{rec}:

$$\min_{\varphi,\theta} q_\varphi(\gamma|x)[-log \frac{p_\theta(x,\gamma)}{c(\gamma|z^{'})}] \tag{1}$$

2.1.3 Regularization Phase. In this phase, we consider the encoding distribution regularization. The conventional AAE [1] only applies binary discriminator which only predicts 'real' or 'fake' data to regularize generation distribution as a fixed prior distribution by similarity, which fails to portray the class-conditioned intraclass disparity. Therefore, we propose a multi-label classifier and a learnable prior distribution to improve the conventional AAE by concluding the following requirements: (i) discriminator should be able to distinguish γ and δ based on the information validity in prediction; (ii) γ represents the generalized class information without subject variation; (iii) δ represents the subject variation conditioned on activity.

As mentioned above, some intraclass disparity exists due to the motion and body shape variance. The similarity distribution regularization will keep this disparity in the representations. Thus, we derive the original binary classifier to a multi-label classifier to distinguish the validity of learned representations. The representations without disparity should be more robust in predicting different subjects, while the disparity representations will impair the performance. Note that, we only consider the case that the dataset contains the class information; otherwise, the prediction validity cannot be measured. Then, we can explain requirement (ii)-(iii) by denoting the low-validity distribution in regularization as the disparity distribution and high-validity distribution as the pure information in class prediction. The requirements (i) can be modified to let discriminator predict whether the latent codes are 'valid' or 'invalid' according to the likelihood of correct predictions.

First, we discuss the Categorical Cross Entropy (CCE) for multi-label classification:

$$\begin{aligned} CCE(y, \hat{y}) &= -\sum_{i}^{C} y_i log(\hat{y}_i) \\ &= \log(\hat{y}_c) \end{aligned} \tag{2}$$

where y denotes a one-hot probability distribution of ground truth; $\hat{y}_c$ represents the discriminator probability of true label c; $|X_c|$ denotes the number of the samples in class c. Due to the activity type and the subject variance, the information amounts carried by signals are varying from each other and thus we apply the mean class spectrum weights of the ground truth to leverage the discriminator training.

We can find that CCE loss is equivalent to the negative log-likelihood of the true prediction probability. Thus, minimizing the CCE loss equals maximizing the validity (i.e., the likelihood of correct prediction). We utilize the adversarial regularization to optimize requirements (ii)-(iii) and then conclude three requirements in an adversarial format.

Given data (X, Y), the competitive distribution (i.e., the encoder of γ) enables the generator distribution (i.e., the encoder of δ) to learn the class conditioned intraclass disparity by optimizing the loss function L_{reg}:

$$\min_{\eta} \max_{\varphi,d} E_{\gamma\sim q_\varphi(\gamma|x)}[\log D(\gamma)_c] + E_{\delta\sim q_\eta(\delta|x)}[\log(1 - D(\delta)_c)] \tag{3}$$

where d denotes the parameters of the multi-label discriminator D; D_c denotes the probability of correct predictions, respectively.

Proof. We set up parameters, φ, d, η, before the optimization, and the network has enough capacity to acquire optimal solutions $\varphi^{'}, d^{'}, \eta^{'}$. When φ is fixed, we can take the pure information encoding distribution as the 'real' data distribution. According to the Proposition 2 in Goodfellow et. al's work [9], at each iteration, the generator (i.e., the disparity encoding distribution $q_{\eta'}(\delta|x)$) will converge to the 'real' data distribution (i.e., the pure information distribution $q_\varphi(\gamma|x)$). Hence, the network can converge when $q_{\eta'}(\delta|x) = q_\varphi(\gamma|x)$. In other words, the disparity encoder will learn the extracted class information distribution which regularizes the disparity to be conditioned on class. Then, $q_\eta(\delta|x)$ can represent a local optimal solution from last iteration, and the optimal $\varphi^{'}$ will be optimized to let discriminator differentiate the validity of disparity and information. Therefore, $q_{\varphi'}(\gamma|x)$ will generate the pure information apart from the previous disparity validated by D_d. Thus, when the network converges, we can get a local optimal distribution pair $q_\eta(\delta|x)$ and $q_{\varphi'}(\gamma|x)$ to decompose latent codes into disparity and pure information, respectively. □

As mentioned in proof, the optimization $D_{d'}$ converges to precisely predicting $q_\varphi(\gamma|x)$ and $q_{\eta'}(\delta|x)$, and $q_{\varphi'}(\gamma|x)$ is optimized to be distinct from $q_{\eta'}(\delta|x)$. In other words, $q_{\varphi'}(\gamma|x)$ is invisible to

discriminator $D_{d'}$. Thus, only if the $q_{\varphi'}(\gamma|x)$ learn the purified data will $D_{d'}$ predict the accurate classification. To further analyze the loss functions of proposed disparity learning, we specify the details of two competitive encoding training and name the two steps of the min-max optimization of L_{reg} as pure loss and disparity loss according to the corresponding encoding distributions:

$$\begin{aligned} L_{pur} &:= \max_{\varphi,d} E_{\gamma\sim q_\varphi(\gamma|x)}[\log D(\gamma)_c] + E_{\delta\sim q_\eta(\delta|x)}[\log(1-D(\delta)_c)] \\ L_{dis} &:= \min_{\eta} E_{\delta\sim q_\eta(\delta|x)}[\log(1-D(\delta)_c)] \end{aligned} \tag{4}$$

We exhibit curves of L_{pur}, L_{dis} to display the convergence of two competitive encoding losses and the training classification performance of using $D_{d'}$ predicting the purified representations $q_{\varphi'}(\delta|x)$ in Section 3.4 **Convergence Analysis**, which proves the intraclass disparity learning and denoising.

2.2 Principled Spectrum Analysis

Considering the gaps and noises inside data segmentation, invalid segments will obstruct the models from learning factual patterns. Most deep learning algorithms only analyze data in the time domain, where it is difficult to analyze information components. Therefore, we introduce the domain knowledge of signal theory to leverage the model optimization according to the information amount from the frequency domain. This section will introduce the principles of spectrum analysis and the fused Spectrum-guided Adversarial Autoencoder, respectively.

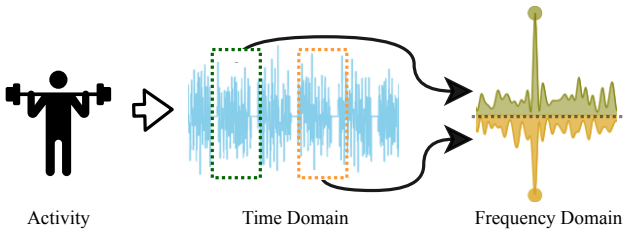


Figure 2: An illustrative example of demonstrating spectrum analysis in the frequency domain. We present two arbitrary data segments while a subject is doing the 'weightlifting' activity. It is represented as a typical time-series sequence in the time domain. By introducing the domain-specific knowledge in the theory of signal processing. It can be transformed into a more robust representation by automatic spectrum analysis proposed in this work, i.e., peak amplitudes are indicated with dotted lines.

2.2.1 Notations for Spectrum Analysis. The spectrum analysis focuses on the patterns of signals in the frequency domain, which is transformed from the time domain. Since sensors are only able to record signals in a certain sampling rate, let $x[n]$ be the observed signals in such discrete-time domain. We can obtain the corresponding frequency domain data X and amplitude spectrum A that records the signal strength of each frequency via Fourier transform[25]:

$$\begin{aligned} X[n] &= \mathcal{F}(x[n]) \\ &= \sum_{n=0}^{N-1} x[n]\cdot e^{-\frac{i2\pi}{N}kn} \quad k\in[0,N-1] \\ A[n] &= |X[n]| \end{aligned} \tag{5}$$

where $A[n]$ denotes the amplitude of frequency n; N denotes the discrete time point number of the sampled signal; i, e represent the imaginary number and the mathematical constant. The amplitude spectrum is perfectly symmetric, so we only analyze the positive frequency part, $k \in [0, N-1]$.

Our goal is to design an amplitude spectrum guide function $S_\zeta(A[i]) \to \mathbb{R}$, which can measure the importance of each frequency according to the signal strength. We define the mean score of the amplitude spectrum to represent the signal information amount (i.e., the signal importance) in the numerical format. To ease our illustration, we express the mean score of an arbitrary frequency set T evaluated by spectrum guide function S_ζ as

$$S_\zeta(T) = \frac{\sum_{i\in T} S_\zeta(i)}{|T|}$$

where ζ denotes the function parameters; $|T|$ represents the element number of T. Therefore, the signal importance can be measured by its amplitude spectrum $S_\zeta(A)$.

To automate the spectrum analysis, we analyze amplitude spectrum in terms of the frequency set. According to [29], the signal information is mainly composed of some frequencies with the highest signal strengths. We let U be the frequency set of the highest amplitudes (i.e., information) and I be the frequency set of the lowest amplitudes (i.e., gap and noise):

$$\begin{aligned} &\forall i\in U, \forall j\in I, A[i] \gg A[j] \\ &s.t. \quad U\cap I = \varnothing, \quad U, I \subseteq A \end{aligned} \tag{6}$$

where U, I are two subsets of A.

2.2.2 Spectrum Analysis Principles. The spectrum analysis principles are based on intra-relationship and inter-relationship in signals, which describe the relationship of frequencies within a signal and between signals, respectively. We illustrate the proposed principles by recalling the phenomenon in Fig. 2: Differing from time-segments, given two arbitrary data segments clipped from a series of noisy continuous signals with gaps, we can easily tell the main composition (i.e., orange and green peak points) and the class of two segments in the frequency domain. Also, we can easily see the green segment contains more information because its curve has more strengths than green segment's, especially around peak points. The conventional band-pass filter analysis [29] manually selects the frequencies with the highest strengths, i.e., the peak points. To automate the spectrum analysis, we transform the frequency selection to weight frequencies based on frequency set analysis and propose two principles:

(1) Given an amplitude spectrum A, $\forall i \in U, \forall j \in I$, function S_ζ should satisfy

$$S_\zeta(A[i]) - S_\zeta(A[j]) \propto A[i] - A[j]$$

(2) Given the certain frequency's amplitudes from two arbitrary different spectra $A_a, A_b, \forall i \in A$, function S_ζ should satisfy

$$S_\zeta(A_a[i]) - S_\zeta(A_b[i]) \propto A_a[i] - A_b[i]$$

The principle (i) describes the traditional band-pass filter theory [29]: if the frequency $A[i]$ is the most significant frequency (i.e., the information frequency), $A[i]$ should have much higher amplitudes than others'. Since principle (i) only considers the intra-amplitude difference relationship, we further derive the principle (ii) to complement the inter-relationship.

Beginning with principle (i), refer to Eq. (6), we consider the **information set** U and the **noise set** I (including noise and gaps) meet the significant amplitude difference requirement. To replace the traditional case-specific band selection, we utilize the information set U to represent the frequencies where information may exist. Then, we can maximize the below loss function L_N to let S_ζ satisfy principle (i):

$$\max_\zeta S_\zeta(U) - S_\zeta(I) \tag{7}$$

Proof. For any two frequencies i, j, the score function S_ζ should let the weight difference between i, j as great as possible, so we should maximize the loss function that $S_\zeta(A[i]) - S_\zeta(A[j])$. Then, for any two subsets U, I,

$$\begin{aligned}\sum_{i\in U}\sum_{j\in I} S_\zeta(A[i]) - S_\zeta(A[j]) &= |U| \cdot |I| \cdot (S_\zeta(U) - S_\zeta(I)) \\ &\propto S_\zeta(U) - S_\zeta(I)\end{aligned}$$

where $|U|, |I|$ denote the element number subsets. Thus, optimizing $S_\zeta(U) - S_\zeta(I)$ is equivalent to the raw formula. □

Further, we utilize principle (ii) to distinguish different signal weights. Similar to principle (i), we can easily conjecture the amplitude difference relationship between the particular frequencies of two spectra in the frequency-set format:

$$\begin{aligned}S_\zeta(A_a) - S_\zeta(A_b) &\propto \frac{1}{N-1}\sum_{i}^{N-1} A_a[i] - A_b[i] \\ &= \overline{A_a} - \overline{A_b}\end{aligned}$$

Then, for arbitrary spectra $\overline{A_a} > \overline{A_b}$, we can minimize the following loss function L_O to meet principle (ii):

$$\min_\zeta |(S_\zeta(A_a) - S_\zeta(A_b)) - \alpha \cdot (\overline{A_a^O} - \overline{A_b^O})| \tag{8}$$

where α denotes the proportionality constant.

2.2.3 Spectrum Guide Function. To put the proposed principles into practice, we further pose a value range in the function $S_\zeta(A) \rightarrow (0, 1)$ (0 means extremely insignificant and 1 means extremely significant), and thus the optimal maximum of $S_\zeta(U) - S_\zeta(I)$ should be 1. Maximizing L_N is equivalent to minimising $1 - L_N$. The empirical proportionality constant holds $\alpha = 1$.

With the above empirical setting, the expected range of A should be $(0, 1)$. To match the different analysis perspectives, we propose to rescale the amplitudes by min-max normalization in two folds. Assuming the spectrum set $\mathcal{A}^{n\times m} := \{A_1, A_2, \ldots, A_n\}$ contains n spectra with m frequency channels. We normalize each spectrum A_i in frequency-level and sample-level to be consistent with two principles (Section 2.2). Denote A_i^N and A_i^O to represent the normalized data of intra- and inter-relationships for an arbitrary spectrum A_i, respectively:

$$\begin{aligned}A_i^N &:= \{\frac{\mathcal{A}_{ij} - \mathcal{A}_i^{min}}{\mathcal{A}_i^{max} - \mathcal{A}_i^{min}} : \mathcal{A}_i \in \mathbb{R}^{1\times m}, j \in [1, m]\} \\ A_i^O &:= \{\frac{\mathcal{A}_{ij} - \mathcal{A}_j^{min}}{\mathcal{A}_j^{max} - \mathcal{A}_j^{min}} : \mathcal{A}_j \in \mathbb{R}^{n\times 1}, j \in [1, m]\}\end{aligned} \tag{9}$$

where $\mathcal{A}_i, \mathcal{A}_j$ denote a row and column vector of $\mathcal{A}$, respectively; $\mathcal{A}_i^{min}, \mathcal{A}_i^{max}, \mathcal{A}_j^{min}, \mathcal{A}_j^{max}$ denote the minimum and maximum of each row or column. Then, the score function evaluates the ith frequency using both intra and inter normalized features by

$$S_\zeta(A[i]) = S_\zeta(\{A^N[i], A^O[i]\})$$

To ensure U and I satisfy our definitions, we empirically take the 20% of frequencies with the highest amplitudes as U and the lowest 50% as I in A^N. With the min-max normalization, the mean amplitude difference between U and I will be large enough to distinguish the noise and information.

The optimization of the score function S_ζ can be demonstrated by two steps: (i) for any amplitude spectrum A, update ζ by minimizing $1 - L_N$; (ii) for any amplitude spectrum pair (A_a, A_b), update ζ by minimizing L_O. We can then unify the steps to optimize ζ by minimizing a pairwise loss function $L_S(A_a, A_b)$:

$$\begin{aligned}L_S(A_a, A_b) = S_\zeta(I_a) - S_\zeta(U_a) + S_\zeta(I_b) - S_\zeta(U_b) + 2 \\ +|S_\zeta(A_a) - S_\zeta(A_b) - (\overline{A_a^O} - \overline{A_b^O})| \\ s.t. \quad A_a := \{A_a^N, A_a^O\}, A_b := \{A_b^N, A_b^O\}\end{aligned} \tag{10}$$

where the first line denotes $1 - L_N$ for A_a^N, A_b^N and the second line is L_O for spectrum pair. Note, we adopt a Fully Connected (FC) layer followed by Sigmoid as S_ζ to match the value range in the definition, and then the optimization will be a traditional regression problem, which can be converged [15].

2.2.4 Spectrum-guided Adversarial Autoencoder. Given a set of signals with labels (X, Y), we calculate the spectrum weighted loss function for an arbitrary signal (x_i, y_i) as follows:

$$\begin{aligned}&\min_{\varphi,\theta} S_\zeta(A_{x_i}) L_{rec}; \min_\eta \max_{\varphi,d} w_c L_{reg} \\ &w_c = \frac{1}{|X_c|}\sum_{x_i\in X_c} S_\zeta(|\mathcal{F}(x_i)|) \\ &s.t. \quad X_c := \{x_i : y_i = c\}\end{aligned} \tag{11}$$

where A_{X_i} denotes the amplitude spectrum of x_i; w_c denotes the weight of x_i's class. We apply two weighting methods in SAAE. Since the reconstruction phase focuses on a single signal waveform, we leverage different signals by its own spectrum. The regularization phase focuses on the class conditioned disparity, so we leverage signals by overall class information amount distribution. The optimization algorithm is shown in Algorithm 1.

Algorithm 1 Adversarial Autoencoder Training

Input Labeled observations (X, Y)
Output $q_\eta, q_{\varphi'}, D_{d'}, S_{\zeta'}$
Initialize network parameters $\zeta, \theta, \varphi, \eta, d$
while epoch < max epoch **do**
  Sample a minibatch $\{(x_1, y_1), (x_2, y_2), ..., (x_k, y_k)\}$
  for $i \leftarrow 1$ to $N-1$ **do**
    $A_{x_i} \leftarrow |\mathcal{F}(x_i)|$
    Calculate $A^N_{x_i}, A^O_{x_i}$
  Sample random spectrum pairs $\{P_1, P_2, ..., P_k\}$
  $L^{'}_S \leftarrow \frac{1}{k}\sum_1^k L_S(P_i)$
  $\zeta^{'} \leftarrow Adam(L^{'}_S)$
  for $i \leftarrow 1$ to k **do**
    Freeze $\zeta^{'}$, $w_{A_{x_i}} \leftarrow S_\zeta(A_{x_i})$
  for $c \leftarrow 1$ to C **do**
    $W_c \leftarrow \{w_{A_{x_i}} : y_i = c\}$
    $w_c \leftarrow \overline{W_c}$
  $w \leftarrow \{w_c : c \in [1, C]\}$
  $L^{'}_{rec} \leftarrow \frac{1}{k}\sum_i^k L_{rec}(x_i, w_{A_{x_i}})$
  $\varphi^{'}, \theta^{'} \leftarrow Adam(L^{'}_{rec})$
  $L^{'}_{reg} \leftarrow \frac{1}{k}\sum_i^k L_{reg}(x_i, w)$
  $\varphi^{'}, d^{'}, \eta^{'} \leftarrow Adam(L^{'}_{reg})$

3 EXPERIMENTS

In this section, we report our comparative experiments and ablation studies in subject-independent settings on four real-world benchmark datasets to evaluate SAAE's robustness and the impact of the spectrum guide function.

3.1 Datasets

We evaluate our approach using four public real-world datasets: 1) MHEALTH [2], a sports activity dataset that records signals of motion and inertial measurement unit (IMU) sensors for 12 different sports activities; 2) PAMAP2 [21], an original daily activity dataset related to 18 daily activities collected from IMUs deployed at different areas of body; 3) UCIDSADS [4], concerning 19 daily and sports activities performed with speed and amplitude variations recorded by motion sensors and IMUs; and 4) OPPORTUNITY [23], concerning 17 hand activities recorded by various body-worn, object-based, and ambient sensors. Considering not all subjects performed all activities in PAMAP2, we excluded six activities (watching TV, computer work, car driving, folding laundry, house cleaning, and playing soccer) and one subject, who conducted very few activities from our experiments.

3.2 Experiment Setting

We execute Leave-One-Subject-Out experiments (with the test subject being excluded from training) to evaluate algorithms' robustness on varying subjects, where we set the time windows as 20 with 50% overlapping to pre-process the time-sequence data. We compare our algorithm, SAAE with five state-of-the-art HAR models and three AAE models, and show a further study on iAAE, a version of the proposed intraclass disparity learning adversarial autoencoder without the knowledge (spectrum) guide function.

- *MC-CNN* [31]: a state-of-the-art Convolutional Neural Network (CNN) that captures temporal correlations along the time axis
- *Bi-LSTM-S* [11]: a bi-directional Long Short-term Memory (LSTM) that captures both forward and backward time information
- *ConvLSTM* [19]: a hybrid model that combines both CNN and LSTM to capture both the spatial and temporal correlations.
- *En-LSTM* [10]: an LSTM-based method that combines multiple individual LSTM learners with epoch-wise bagging
- *AttConvLSTM* [18]: a hybrid model that applies attention layer to learn weighted information
- *AAE* [17]: a benchmark of adversarial autoencoder with FC layers which utilizes adversarial training to regularize autoencoder
- *ConvAAE* [1]: a state-of-the-art adversarial autoencoder algorithm with convolutional layers to learn the long-time multichannel sensor information for gesture recognition
- *DAAE* [7]: a convolutional adversarial autoencoder further enhanced with denoising operation to extract the pure information from low signal-to-noise ratio data

We take MC-CNN as the base model for the encoder, three deconvolutional blocks, namely the deconvolutional layer, the rectified linear unit (ReLU), and the batch normalization layer for the decoder, an FC layer, and a softmax function for the discriminator. The detail of Two FC layers and a sigmoid function for spectrum score function S_ζ are shown in the supplementary Section A. Our model applies the same encoder-decoder structure as in DAAE but differs in using the adversarial structure and the spectrum guide function. To show the robustness of methods, we use the same hyperparameter settings over all the subjects in the same datasets. For SAAE, we set the learning rate of the spectrum score function of to 1e-4 for the four datasets, the learning rates of the adversarial network to 5e-5 for PAMAP2 and 2e-4 for the other datasets.

3.3 Overall Comparison

The results (Table 1) show our models outperforming all other algorithms on the four datasets and having the smallest standard deviations, demonstrating the effectiveness of SAAE and its variant, iAAE (without Spectrum knowledge) on four benchmark datasets. We can easily observe that the proposed models, iAAE and SAAE, consistently outperform all other baselines and hold the smallest standard deviation. In terms of F1-score, SAAE improves about 2%, 3%, 4%, and 6% in four datasets, It also indicates the robustness in handling intraclass variance of new subjects.

3.4 Further Analysis

Effectiveness of Adversarial Training SAAE achieves a 4.3% improvement in accuracy, 4.8% in precision, and 5.2% in F1 Score comparing with a set of encoder-decoder based methods. It demonstrates the advantage of our proposed competitive encoding distribution learning against conventional adversarial training approaches.

We further show comparisons on MHEALTH at the individual level (Fig. 3) (a). For instance, MC-CNN can achieve 97% accuracy on subject 9 while only 87% in subject 3, revealing that MC-CNN may

Table 1: Overall Comparison of SAAE over Four Benchmark Datasets

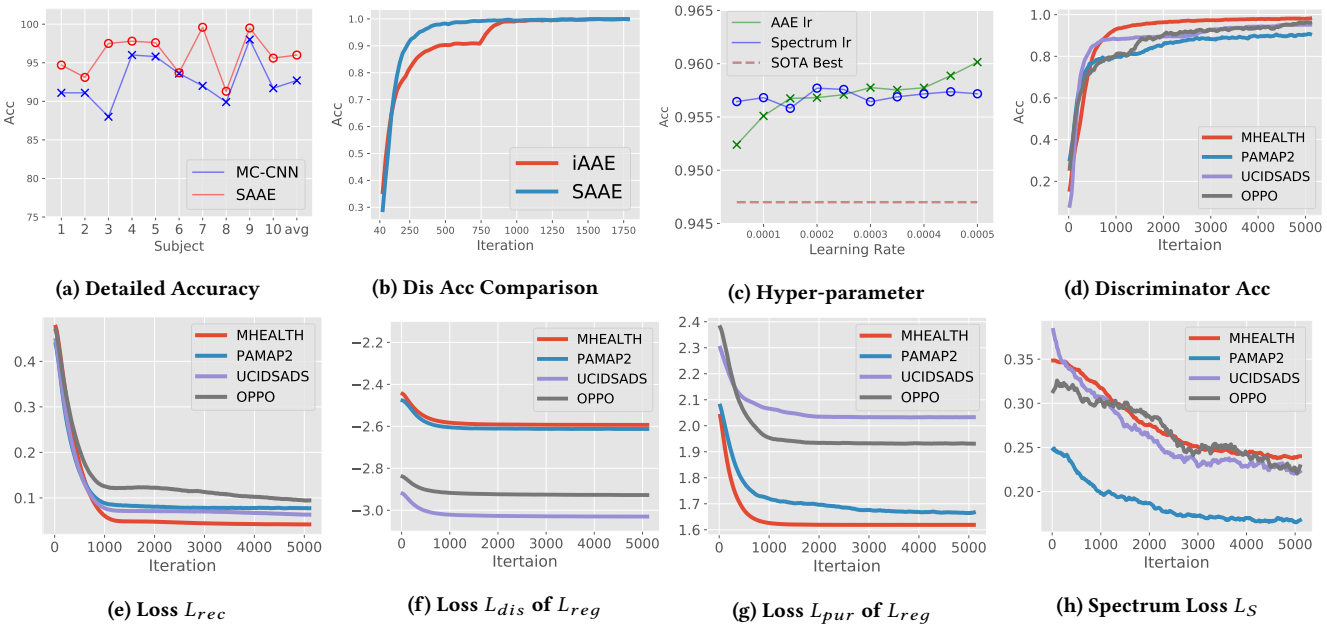
MHEALTH	Metrics	MC-CNN	Bi-LSTM	ConvLSTM	En-LSTM	AttConvLSTM
	Acc	0.927(0.031)	0.899(0.044)	0.913(0.038)	0.860(0.089)	0.921(0.042)
	Pre	0.935(0.039)	0.871(0.062)	0.893(0.051)	0.848(0.085)	0.899(0.052)
	F1	0.922(0.037)	0.879(0.054)	0.899(0.047)	0.846(0.088)	0.908(0.050)
	Metrics	AAE	ConvAAE	DAAE	iAAE	SAAE
	Acc	0.907(0.054)	0.826(0.061)	0.947(0.049)	0.955(0.028)	**0.958(0.028)**
	Pre	0.905(0.067)	0.804(0.078)	0.941(0.063)	0.960(0.026)	**0.963(0.026)**
	F1	0.897(0.063)	0.809(0.072)	0.941(0.060)	0.957(0.028)	**0.960(0.030)**
PAMAP2	Metrics	MC-CNN	Bi-LSTM	ConvLSTM	En-LSTM	AttConvLSTM
	Acc	0.803(0.133)	0.715(0.200)	0.757(0.158)	0.734(0.157)	0.741(0.146)
	Pre	0.806(0.134)	0.711(0.241)	0.730(0.202)	0.739(0.196)	0.739(0.120)
	F1	0.781(0.154)	0.687(0.230)	0.724(0.192)	0.720(0.180)	0.718(0.150)
	Metrics	AAE	ConvAAE	DAAE	iAAE	SAAE
	Acc	0.727(0.194)	0.713(0.147)	0.774(0.180)	0.837(0.112)	**0.840(0.109**)
	Pre	0.771(0.195)	0.730(0.155)	0.805(0.153)	**0.855(0.096)**	**0.855(0.101)**
	F1	0.746(0.215)	0.693(0.168)	0.764(0.197)	0.831(0.124)	**0.836(0.122)**
UCIDSADS	Metrics	MC-CNN	Bi-LSTM	ConvLSTM	En-LSTM	AttConvLSTM
	Acc	0.879(0.067)	0.889(0.049)	0.897(0.046)	0.831(0.043)	0.887(0.048)
	Pre	0.872(0.095)	0.897(0.066)	0.896(0.063)	0.841(0.065)	0.882(0.072)
	F1	0.855(0.087)	0.877(0.061)	0.884(0.059)	0.811(0.055)	0.868(0.064)
	Metrics	AAE	ConvAAE	DAAE	iAAE	SAAE
	Acc	0.846(0.045)	0.815(0.033)	0.889(0.047)	0.918(0.044)	**0.929(0.040)**
	Pre	0.857(0.055)	0.790(0.047)	0.907(0.041)	0.926(0.045)	**0.935(0.048)**
	F1	0.824(0.056)	0.785(0.040)	0.875(0.056)	0.906(0.052)	**0.919(0.047)**
OPPORTUNITY	Metrics	MC-CNN	Bi-LSTM	ConvLSTM	En-LSTM	AttConvLSTM
	Acc	0.635(0.050)	0.575(0.096)	0.537(0.088)	0.531(0.107)	0.566(0.092)
	Pre	0.637(0.024)	0.599(0.101)	0.482(0.124)	0.534(0.094)	0.587(0.093)
	F1	0.613(0.042)	0.549(0.098)	0.464(0.116)	0.485(0.113)	0.530(0.107)
	Metrics	AAE	ConvAAE	DAAE	iAAE	SAAE
	Acc	0.624(0.061)	0.609(0.078)	0.625(0.069)	0.664(0.056)	**0.680(0.049)**
	Pre	0.663(0.025)	0.661(0.062)	0.618(0.076)	0.698(0.053)	**0.713(0.048)**
	F1	0.603(0.064)	0.597(0.077)	0.598(0.075)	0.655(0.055)	**0.674(0.046)**

perform badly while dealing with the new subject whose samples are deviating a bit more from the common distribution of training subjects. In contrast, SAAE steadily improves the performance on most subjects, especially subject 3 and subject 7, which proves its robustness, reliability, and potential to be deployed in real-world applications.

Effectiveness of Spectrum Analysis Compared with SAAE and iAAE, the spectrum guide function can further enhance the performance by around 1% in each subject, validating the effectiveness of frequency domain analysis in model learning. We also take the subject 1 in MHEALTH as an instance to illustrate the spectrum analysis's effectiveness in optimization. Our convergence comparison of discriminator D (Fig. 3 (b)) shows spectrum information (i.e., domain knowledge) can better exclude the disparity during optimization.

Hyper-parameter Analysis. We change the learning rates of different components to explore the optimal hyper-parameters. We set learning rate as 1e-4 for spectrum score function and 2e-4 for AAE as default, and plots the mean accuracy of one-our subject-independent experiments when learning rates range from [5e-5,5e-4] in Fig. 3 (c). We can observe that SAAE is stable over different parameters and constantly outperforms the best state-of-the-art. The spectrum learning rate merely influences the model performance while AAE's learning rate will slightly affects the results. With larger learning rate of AAE, the results gradually become better, which means SAAE could act better than our provided result.

Convergence Analysis. Fig. 3 (d) - (h) plot the averaged loss over four datasets. We can observe that $L_{rec}, L_{dis}, L_{pur}$ could quickly converge around 1000 iterations and then remain stable, where L_{dis} and L_{pur} represent the min-max competition loss functions of intraclass disparity and purified information, respectively. The convergence of L_{pur} and L_{dis} reveals the SAAE's capability of optimizing the corresponding encoding distributions as expected.

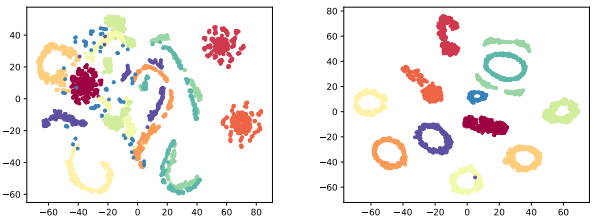


(a) Detailed Accuracy (b) Dis Acc Comparison (c) Hyper-parameter (d) Discriminator Acc

(e) Loss L_{rec} (f) Loss L_{dis} of L_{reg} (g) Loss L_{pur} of L_{reg} (h) Spectrum Loss L_S

Figure 3: Experiment Analysis.

Besides, the discriminator can precisely predict the pure information components, which proves that SAAE is capable of learning to exclude the disparity components in latent codes through the adversarial training. The spectrum loss L_S drops rapidly at first and then smoothly decreases until convergence over iterations.



(a) MHEALTH Raw (b) MHEALTH SAAE

Figure 4: Visualization of raw data and SAAE's learned latent codes.

Visualization of Learned Representation. Fig. 4 visualizes the raw data and the learned features. The original dimension is reduced by t-Distributed Stochastic Neighbor Embedding (t-SNE). The learned features are more similar within classes yet more dissimilar between classes over the sample space than the raw features, indicating the effectiveness of the network's feature learning.

4 RELATED WORK

The previous research on intraclass disparity falls into two categories: subject-dependent and subject-independent models. Some of the subject-dependent algorithms attempt to discover patterns by case-specific analysis: Ren et al. [27] captured the community in natural activities through hand trajectory; Tapia et al. [22] designed a more comprehensive system to record the invariant features of angle movement through wireless sensors.

Another thread of subject-dependent research uses deep neural networks to learn representations for automated feature design and then activity recognition. Such methods focus on capturing the discriminative representations in the signal streams. Yang et al. [31] applied convolution and pooling filters along the temporal dimensions to catch the difference in the long-term time sequence signals; Fransisco et al. [19], who further improved CNN with LSTM but still neglect the varying information amounts in signals. To validate the significance of signals, Vishvak et al. [18] integrated a temporal attention module that aligns the output vector of the last time step with other steps' to learn a relative importance score to modify the learning process. However, subject-dependent methods fail to consider the unique patterns that may occur in new subjects, as the common patterns in existing subjects may not include all the potential conditions.

Subject-independent studies aim to enhance the generalization ability of models into precise recognition of new persons. Sani et al. [24] proposed to construct a support set and match the most similar instances to ease the unique patterns'. Yu et al. [10] ensembled the models from different iterations and promise a more generalized model. The above non-generative models still lack generalization due to the model limitation. Balabka et al. [1] applied AAE to approximate a generalized distribution to simulate human activity distributions. Thus, the encoder in AAE could be more generalized and robust when handling new subjects. Zhang et al. [32] fused the advantages of variational autoencoders and generative adversarial networks and designed a regularized latent representations for

generation. However, these works exclusively analyzed the time-domain information and all considered the intraclass disparity as meaningless noise and failed to extract the disparity distribution by adversarial training.

Besides analyzing raw data patterns, some hand-engineered domain knowledge was introduced to further extract high-level discriminative information to assist predictions. The commonly used features include time-domain features (e.g., mean, variance, skewness) and frequency-domain features (e.g., power spectral density) [14]. Some studies design new features containing temporal and structural information. For example, Hammerla et al. [12] proposed the Empirical Cumulative Density Function (ECDF) feature to extract the spatial information of the signal frames. Such methods are generally heuristic and lack generalization in different scenarios.

Our work differs from the studies above on two aspects: utilization of spectrum analysis and specific intraclass disparity distribution learning. We implement and embrace spectrum analysis as a tool in AAE to leverage the optimization based on signal information amount. Further, we specify and precisely portray the class-conditioned intraclass disparity in a learnable competitive encoding distribution, which enables AAE effectively to extract and to denoise such disparity.

5 CONCLUSION

We propose novel spectrum-guided disparity learning, or SAAE, to address intraclass variability. We design two competitive encoding distributions under a unified adversarial training framework, rather than a fixed prior distribution, to learn robust embeddings that can be generalized to new subjects. We further incorporate the domain-specific knowledge in an unsupervised manner. We experimentally validate our model on four representative benchmark dataset with state-of-the-art methods. The results demonstrate the superior performance and robustness of the proposed model in predictions on unknown subjects. Given SAAE's promising performance in handling intraclass disparity, we will extend it to handle more complex scenarios in the future.

REFERENCES

[1] Dmitrijs Balabka. 2019. Semi-supervised learning for human activity recognition using adversarial autoencoders. In *Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers*. ACM, 685–688.

[2] Oresti Banos, Rafael Garcia, Juan A Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. 2014. mHealthDroid: a novel framework for agile development of mobile health applications. In *International workshop on ambient assisted living*. Springer, 91–98.

[3] Ling Bao and Stephen S Intille. 2004. Activity recognition from user-annotated acceleration data. In *International conference on pervasive computing*. Springer, 1–17.

[4] 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.

[5] Jun Cai, Jing Chen, and Xing Liang. 2015. Single-sample face recognition based on intra-class differences in a variation model. *Sensors* 15, 1 (2015), 1071–1087.

[6] Jun Cai, Jing Chen, and Xing Liang. 2015. Single-sample face recognition based on intra-class differences in a variation model. *Sensors* 15, 1 (2015), 1071–1087.

[7] Antonia Creswell and Anil Anthony Bharath. 2018. Denoising adversarial autoencoders. *IEEE transactions on neural networks and learning systems* 30, 4 (2018), 968–984.

[8] WeiWang Dong-DongChen and Zhi-HuaZhou WeiGao. 2018. Tri-net for semi-supervised deep learning. In *Proceedings of Twenty-Seventh International Joint Conference on Artificial Intelligence*. 2014–2020.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[10] 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), 11.

[11] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 1533–1540.

[12] Nils Y Hammerla, Reuben Kirkham, Peter Andras, and Thomas Ploetz. 2013. On preserving statistical characteristics of accelerometry data using their empirical cumulative distribution. In *Proceedings of the 2013 International Symposium on Wearable Computers*. ACM, 65–68.

[13] Daniel Im Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. 2017. Denoising criterion for variational auto-encoding framework. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[14] Majid Janidarmian, Atena Roshan Fekr, Katarzyna Radecka, and Zeljko Zilic. 2017. A comprehensive analysis on wearable acceleration sensors in human activity recognition. *Sensors* 17, 3 (2017), 529.

[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[16] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[17] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644* (2015).

[18] Vishvak S Murahari and Thomas Plötz. 2018. On attention models for human activity recognition. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers*. ACM, 100–103.

[19] Francisco Ordóñez and Daniel Roggen. 2016. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors* 16, 1 (2016), 115.

[20] Sasank Reddy, Min Mun, Jeff Burke, Deborah Estrin, Mark Hansen, and Mani Srivastava. 2010. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)* 6, 2 (2010), 13.

[21] 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.

[22] Haibing Ren, Guangyou Xu, and SeokCheol Kee. 2004. Subject-independent natural action recognition. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* IEEE, 523–528.

[23] Daniel Roggen, Alberto Calatroni, Mirco Rossi, Thomas Holleczek, Kilian Förster, Gerhard Tröster, Paul Lukowicz, David Bannach, Gerald Pirkl, Alois Ferscha, et al. 2010. Collecting complex activity datasets in highly rich networked sensor environments. In *2010 Seventh international conference on networked sensing systems (INSS)*. IEEE, 233–240.

[24] Sadiq Sani, Nirmalie Wiratunga, Stewart Massie, and Kay Cooper. 2018. Matching networks for personalised human activity recognition. CEUR Workshop Proceedings.

[25] Steven W Smith et al. 1997. The scientist and engineer's guide to digital signal processing. (1997).

[26] Nagender K Suryadevara and Subhas C Mukhopadhyay. 2014. Determining wellness through an ambient assisted living environment. *IEEE Intelligent Systems* 29, 3 (2014), 30–37.

[27] Emmanuel Munguia Tapia, Stephen S Intille, William Haskell, Kent Larson, Julie Wright, Abby King, and Robert Friedman. 2007. Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. In *2007 11th IEEE international symposium on wearable computers*. IEEE, 37–40.

[28] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*. ACM, 1096–1103.

[29] Andrew J Viterbi and Jim K Omura. 2013. *Principles of digital communication and coding*. Courier Corporation.

[30] 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.

[31] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[32] Xiang Zhang, Lina Yao, and Feng Yuan. 2019. Adversarial Variational Embedding for Robust Semi-supervised Learning. *arXiv preprint arXiv:1905.02361* (2019).

A NETWORK ARCHITECTURE

We plot the detailed network structure in Fig. 5. The decoder and encoders consist of three independent blocks and two encoders share the same structure. The architectures of Block are shown in Fig. 6 and the parameters are shown in Table 2. Specially, Conv Block 3 & 6 and Deconv Block 3 do not have Maxpool layer. The stride of layers are 1 and padding way is 0 as default. Discriminator D is composed of one FC layer, so the input dim equals the output element number of encoders and output dim equals the target class number. Given a amplitude spectrum $A \in \mathbb{R}^{1\times m}$, then the Spectrum Score Function S_ζ consists of two FC layers: first layer takes $2 \cdot m$ dimension input (i.e., A^N and A^O and keeps the same output dimension; second layer generates m-dimension outputs representing the corresponding weights for the frequencies in spectrum.

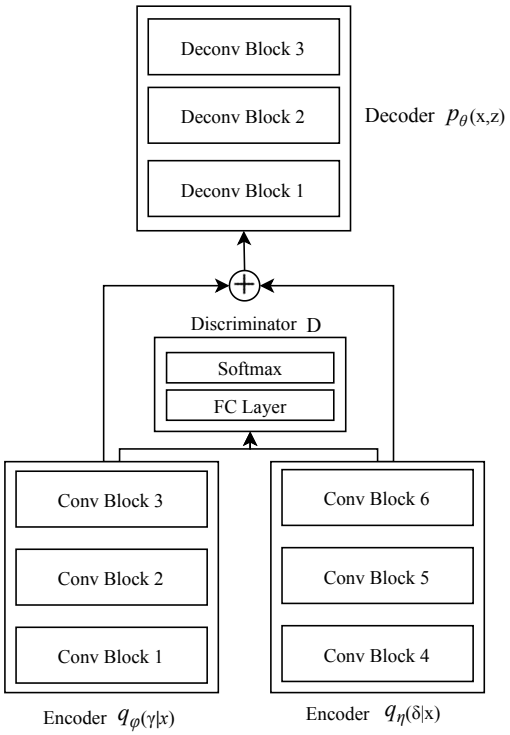


Figure 5: Network Detail

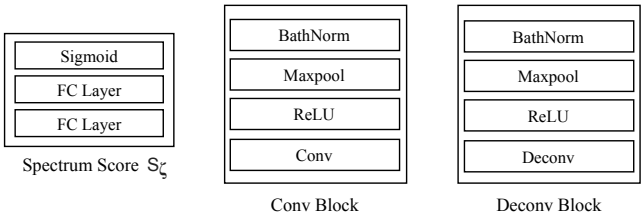


Figure 6: Module Detail

Table 2: Block Parameters.

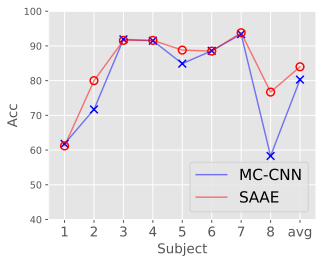
Name	#Kernel	Kernel Shape	Maxpool
Conv Block 1&4	50	(5,1)	(2,1)
Conv Block 2&5	40	(5,1)	(2,1)
Conv Block 3&6	20	(2,1)	none
Deconv Block 1	40	(5,1)	(2,1)
Deconv Block 2	50	(5,1)	(2,1)
Deconv Block 3	2	(2,1)	none

A.1 Supplementary Experiment

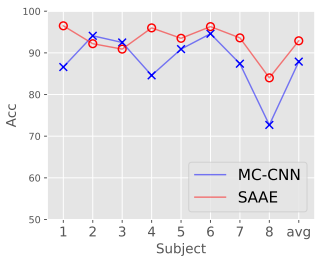
Due to the space limitation, we pose the other three datasets' analysis, i.e., specific subject accuracy reports, spectrum guided fitting curves, and embedding comparison in Fig. 7.

From (a)-(c), we can observe the subject variation in other three datasets, especially the subject 8 in PAMAP2 and subject 8 in UCIDSADS, while SAAE achieves stable performance over these unstable subjects. Also, SAAE improves all three subjects in OPPORTUNITY dataset. (d)-(f) further provide the discriminator fitting curves of subject 1 on other three datasets, which indicates the effectiveness of domain knowledge in diverse scenarios. (g)-(l) exhibit the data distributions before and after SAAE's purification. We can easily see that points of same classes become more gathered and there are fewer scattered points over all three datasets.

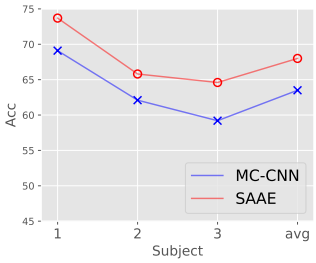
We also exhibit the Confusion matrices over four datasets to assist proving our algorithm's outperformance and robustness. Fig. 8 plots a subject's confusion matrix of the corresponding datasets. We can observe that the classification on MHEALTH is solid, and only a few samples are misclassified. The prediction in PAMAP2 shows that it is difficult to learn the generalized representations in some classes, where the misclassified samples spread over other multiple diverse activities. However, there exist the interclass similarity in UCIDSADS and OPPORTUNITY. Most of the classes can be precisely predicted in these two datasets, but the misclassified samples mainly gather in one class. For instance, the misclassified samples of class 9 (running) centers in class 10 (exercise in steps) in UCIDSADS, which indicates the interclass similarity between class 9 and class 10.



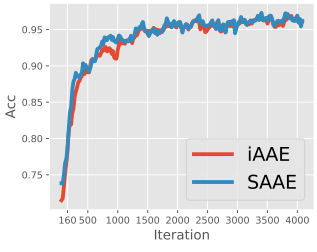
(a) PAMAP2 Acc Report



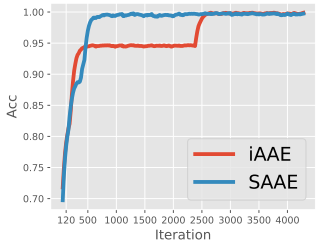
(b) UCIDSADS Acc Report



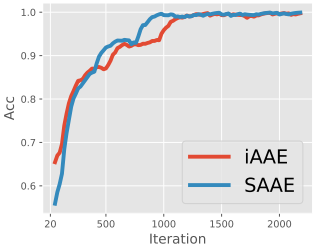
(c) OPPORTUNITY Acc Report



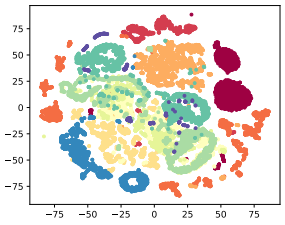
(d) PAMAP2 Dis Acc



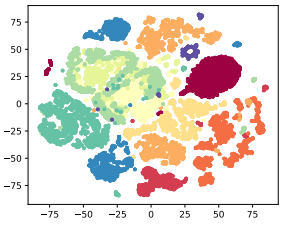
(e) UCIDSADS Dis Acc



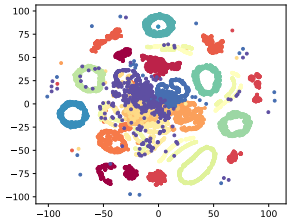
(f) OPPORTUNITY Dis Acc



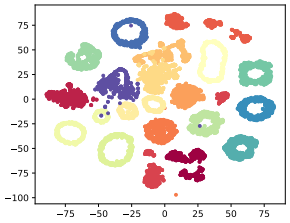
(g) PAMAP2 Raw



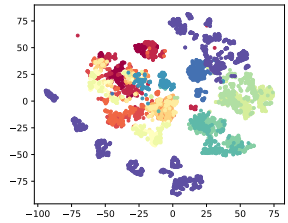
(h) PAMAP2 SAAE



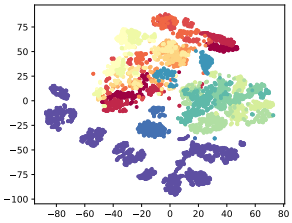
(i) UCIDSADS Raw



(j) UCIDSADS SAAE

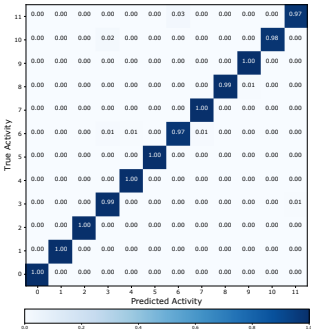


(k) OPPORTUNITY Raw

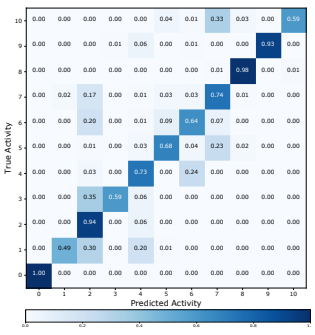


(l) OPPORTUNITY SAAE

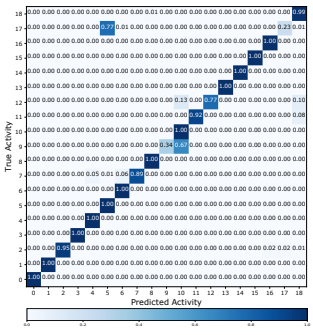
Figure 7: Performance on differnet subjects.



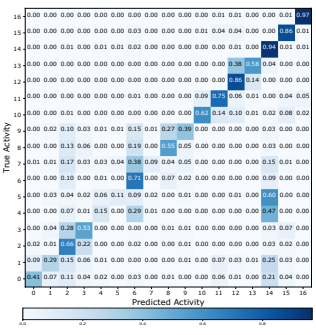
(a) MHEALTH



(b) PAMAP2



(c) UCIDSADS



(d) OPPORTUNITY

Figure 8: Confusion matrices of four datasets.