DWLR: Domain Adaptation under Label Shift for Wearable Sensor (Appendix)

Juren Li[1] , **Yang Yang**[1*] , **Youmin Chen**[1] , **Jianfeng Zhang**[2] , **Zeyu Lai**[1] and **Lujia Pan**[2]
[1]College of Computer Science and Technology, Zhejiang University
[2]Huawei Noah's Ark Lab
{jrlee, yangya, 22251334, jerrylai}@zju.edu.cn, {zhangjianfeng3, panlujia}@huawei.com

A Information Gain

Bayesian neural networks (BNNs) [Gal and Ghahramani, 2015], replaces the weight parameters of a deterministic model by a assumed prior parameter distribution over its weight. Given a set of data $Q = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, a posterior distribution over the parameters is inferred, $p(\theta|Q)$, where θ denotes the parameters. The goal of BNNs is to reduce the number of possible parameters θ, *i.e.*, minimize the uncertainty of parameters which can be computed with Shannon's entropy [Cover and Thomas, 1991]. When sample a subset Q' from Q and achieve the goal of BNNs, Q' needs to satisfy :

$$\text{argmin}_{Q'} H[\theta|Q'] = -\int_{\theta} p(\theta|Q') \log p(\theta|Q') \mathrm{d}\theta \tag{15}$$

Eq 15 is to select the subset that reduce the model's parameters uncertainty as much as possible. In other words, it aims to select the subset that bring more information to the model,*i.e.*, subset with high **information gain** which can be used to measure the reduction of uncertainty of the model when given a random variable [Cover, 1999].

However, solving the problem of Eq 15 is NP-hard. [Houlsby *et al.*, 2011] proposes a myopic policy performing near-optimally [Heckerman *et al.*, 1994; Golovin and Krause, 2010; Dasgupta, 2004]. The object is to seek the data point $\mathbf{x}$ that maximizes the decrease in expected posterior entropy (x with highest information gain, referred to as $\mathbb{IG}(\mathbf{x})$):

$$\underset{\mathbf{x}}{\text{argmax}} H[\theta|Q] - \mathbb{E}_{\hat{y}\sim p(\hat{y}|\mathbf{x},Q)}[H[\theta|\hat{y},\mathbf{x},Q]] \tag{16}$$

However, it is hard to compute the entropy of θ. From the information theory [Cover and Thomas, 1991], Eq 16 is equivalent to the conditional mutual information between the parameters θ and output $\hat{y}$, *i.e.*, $I[\hat{y},\theta|\mathbf{x},Q]$.

$$I[\hat{y},\theta|\mathbf{x},Q] = H[\theta|Q] - \mathbb{E}_{\hat{y}\sim p(\hat{y}|\mathbf{x},Q)}[H[\theta|\hat{y},\mathbf{x},Q]] \tag{17}$$

And we can get:

$$I[\hat{y},\theta|\mathbf{x},Q] = H[\hat{y}|\mathbf{x},Q] - \mathbb{E}_{\theta\sim p(\theta|Q)}[H[\hat{y}|\mathbf{x},\theta]] \tag{18}$$

Consequently, the objective can be rearranged to compute the entropies in $\hat{y}$ space:

$$\underset{\mathbf{x}}{\text{argmax}} H[\hat{y}|\mathbf{x},Q] - \mathbb{E}_{\theta\sim p(\theta|Q)}[H[\hat{y}|\mathbf{x},\theta]] \tag{19}$$

*Corresponding author.

Thus, we only need to compute the entropy of $\hat{y}$ and we can get that $\mathbb{IG}(\mathbf{x}) = H[\hat{y}|\mathbf{x},Q] - \mathbb{E}_{\theta\sim p(\theta|Q)}[H[\hat{y}|\mathbf{x},\theta]]$.

Given a input data $\mathbf{x}$, the probability distribution over the class i is given by:

$$p(\hat{y}=i|\mathbf{x},Q) = \int_{\theta} p(\hat{y}=i|\theta,\mathbf{x}) p(\theta|Q) \mathrm{d}\theta \tag{20}$$

But Eq 20 requires averaging over all possible θ, which is intractable. [Gal *et al.*, 2017] proposes that dropout [Srivastava *et al.*, 2014] can be used as the stochastic regularisation techniques to perform approximate inference in the BNNs.

$$\begin{aligned} p(\hat{y}=i|\mathbf{x},Q) &= \int_{\theta} p(\hat{y}=i|\theta,\mathbf{x}) p(\theta|Q) \mathrm{d}\theta \\ &\approx \int_{\theta} p(\hat{y}=i|\theta,\mathbf{x}) q^{\star}(\theta) \mathrm{d}\theta \\ &\approx \frac{1}{T}\sum_{\tau=1}^{T} p(\hat{y}=i|\theta_{\tau},\mathbf{x}) \end{aligned} \tag{21}$$

where $q^{\star}(\theta)$ is dropout distribution and $\theta_{\tau} \sim q^{\star}(\theta)$. Based on the Shannon's entropy, the $H[\hat{y}|\mathbf{x},Q]$ and $\mathbb{E}_{\theta\sim p(\theta|Q)}[H[\hat{y}|\mathbf{x},\theta]]$ can be compute by

$$H[\hat{y}|\mathbf{x},Q] = -\sum_{i=1}^{k} p(\hat{y}=i|\mathbf{x},Q) \text{log} p(\hat{y}=i|\mathbf{x},Q) \tag{22}$$

and

$$\begin{aligned} &\mathbb{E}_{\theta\sim p(\theta|Q)}[H[\hat{y}|\mathbf{x},\theta]] = \\ &-\mathbb{E}_{\theta\sim q^{\star}(\theta)}[\sum_{i=1}^{k} p(\hat{y}=i|\theta,\mathbf{x}) \text{log} p(\hat{y}=i|\theta,\mathbf{x})] \end{aligned} \tag{23}$$

we can get:

$$\begin{aligned}\mathbb{IG}(\mathbf{x}) =& H[\hat{y}|\mathbf{x}, Q] - \mathbb{E}_{\theta\sim p(\theta|Q)}[H[\hat{y}|\mathbf{x}, \theta]] \\ =& -\sum_{i=1}^{k} p(\hat{y}=i|\mathbf{x}, Q)\log p(\hat{y}=i|\mathbf{x}, Q) \\ &+ \mathbb{E}_{\theta\sim q^{\star}(\theta)}[\sum_{i=1}^{k} p(\hat{y}=i|\theta,\mathbf{x})\log p(\hat{y}=i|\theta,\mathbf{x})] \\ \approx& -\sum_{i=1}^{k}(\frac{1}{T}\sum_{\tau=1}^{T} p(\hat{y}=i|\mathbf{x}, \theta_\tau))\log(\frac{1}{T}\sum_{\tau=1}^{T} p(\hat{y}=i|\mathbf{x}, \theta_\tau)) \\ &+ \frac{1}{T}\sum_{\tau=1}^{T}\sum_{i=1}^{k} p(\hat{y}=i|\mathbf{x}, \theta_\tau)\log p(\hat{y}=i|\mathbf{x}, \theta_\tau)\end{aligned} \tag{24}$$

Then, we get the Eq 8 in the section of Method to compute the information gain approximately.

B DWLR Overview

Algorithm 1 Overview of DWLR

Input: source dataset P, target dataset Q; pre-train epochs E_1, UDA epochs E_2; Encoder $F(\cdot)$, $C(\cdot)$, $WNet(\cdot)$ and discriminator $D(\cdot)$. **Parameter**: α, β

for E_1 epochs **do**
  for all $\mathbf{x}_i^s, y_i^s \in P$ **do**
    $\mathbf{f}_i^s \leftarrow F(\mathbf{x}_i^s)$ # *extrct feature*
    $\mathbf{h}_i^s \leftarrow C(\mathbf{f}_i^s)$ # *classification*
    $\mathcal{L}_{task} \leftarrow$ Eq 5 # *compute loss*
    Update $F(\cdot)$ and $C(\cdot)$ with $\mathcal{L}_{task}$
  end for
end for
for E_2 epochs **do**
  for all $\mathbf{x}_i^s, y_i^s \in P$ **do**
    $\mathbf{f}_i^s \leftarrow F(\mathbf{x}_i^s)$
    $\mathbf{h}_i^s \leftarrow C(\mathbf{h}_i^s)$
    $d_i^s \leftarrow D(\mathbf{f}_i^s)$ # *domain discrimination*
  end for
  for all $\mathbf{x}_i^t \in Q$ **do**
    $\mathbf{f}_i^t \leftarrow F(\mathbf{x}_i^t)$
    $\mathbf{d}_i^t \leftarrow D(\mathbf{f}_i^t)$
    $\mathbf{h}_i^t \leftarrow C(\mathbf{f}_t).detach()$
    $\hat{y}_i^t \leftarrow \underset{j\in\{1,\cdots,k\}}{\text{argmax}}\ \mathbf{h}_i^t[j]$ # *pseudo label*
    $\mathbb{IG}(\mathbf{f}_i^t) \leftarrow$ MC-Dropout with $\mathbf{f}_i^t$ by Eq 8
    $w_i \leftarrow WNet(\mathbf{f}_i^t)$ # weight
  end for
  # *Compute loss*
  $\mathcal{L}_{weight} \leftarrow \mathcal{L}_{label} + \mathcal{L}_{IG} + \mathcal{L}_{conf}$
  $\mathcal{L}_{task} \leftarrow$ Eq 5; $\mathcal{L}_{adv} \leftarrow$ Eq 13;
  $\mathcal{L}_{net} = \mathcal{L}_{task} + \alpha\mathcal{L}_{adv} + \beta\mathcal{L}_{weight}$
  $\mathcal{L}_{dis} \leftarrow$ Eq 12
  Update $F(\cdot)$, $C(\cdot)$ and $WNet(\cdot)$ with $\mathcal{L}_{net}$
  Update $D(\cdot)$ with $\mathcal{L}_{dis}$
end for

In the Method section, we presented a comprehensive description of each component of our proposed DWLR. Additionally, we provided a concise overview of the training process for DWLR. In this section, we present a detailed description of the training process of DWLR using Algorithm 1. The training process is the same in the time domain and the frequency domain, but the specific implementation of the encoder is different, where the frequency domain encoder requires an additional discrete Fourier change (DFT).

Specifically, we first pre-train the encoder $F(\cdot)$ and the classifier $C(\cdot)$ by minimizing the task loss $\mathcal{L}_{task}$ using the labeled source domain data. Then, we perform domain adaptation with label shift. For each source domain data $\mathbf{x}_i^s$, we obtain its feature representation $\mathbf{f}_i^s$ using $F(\cdot)$, predict the scores $\mathbf{h}_i^s$ for the classification task using $C(\cdot)$, and obtain the domain discriminative scores $\mathbf{d}_i^s$ using $D(\cdot)$. For each target domain sample $\mathbf{x}_i^t$, we obtain its feature representation $\mathbf{f}_i^t$ using $F(\cdot)$ and domain discriminative scores $\mathbf{d}_i^t$ using $D(\cdot)$. Meanwhile, we obtain the pseudo labels $\hat{y}_i^t$ with confidence h_i^t. Based on the target domain features $\mathbf{f}_i^t$, we use the MC-Dropout to calculate the approximate information gain $\mathbb{IG}(\mathbf{f}_i^t)$ for the corresponding sample. Meanwhile, $WNet(\cdot)$ output the weight w_i for the sample based on $\mathbf{f}_i^t$. The total weight of data remains unchanged before and after reweighting, i.e., $\sum_{i=1}^{N^t} w_i = N^t$. Then, using Eq 7, Eq 9, Eq 10, and Eq 10, we obtain the objective value $\mathcal{L}_{weight}$ for $WNet(\cdot)$. Next, using Eq 14, we obtain the objective value $\mathcal{L}_{net}$. We minimize $\mathcal{L}_{net}$ to optimize $F(\cdot)$, $C(\cdot)$ and $WNet(\cdot)$. During learnable reweighting, $C(\cdot)$ is used for obtaining pseudo-labels and calculating information gain, but the gradient of $\mathcal{L}_{weight}$ will not be propagated to $C(\cdot)$. Similarly, $\mathcal{L}_{task}$ does not affect $WNet(\cdot)$ and $D(\cdot)$. We use Eq 12 to obtain $\mathcal{L}_{dis}$ and minimize it to optimize $D(\cdot)$.

C Dataset

In our experiments, we use three real-world wearable sensor datasets and a human sensor dataset.

- **WISDM** [Kwapisz *et al.*, 2011]. This dataset comprises labeled accelerometer data from 29 users recording their daily activities, which include walking, jogging, sitting, standing, walking upstairs, and walking downstairs. The dataset consists of 1,098,207 data records in total with sampling frequency of 20Hz. In our experiment, we used a time window of size 128 to construct samples from the dataset. Moreover, for conducting experiments, we selected several user pairs with significant label shift, where one was considered as a source domain and the other as a target domain.
- **UCIHAR** [Anguita *et al.*, 2013]. This dataset collects data of accelerometer, gyroscope and estimated body acceleration from 30 users with sampling frequency of 20Hz. The dataset include six different daily activities recorded, including walking, sitting, lying, standing, walking upstairs, and walking downstairs. Since label distributions among users in this dataset is similar, we applied various data augmentation techniques [Um *et al.*, 2017], such as jittering, scaling, magnitude warping, time warping, and random sampling to create an augmented dataset with label shift between users. And we

randomly selected several user pairs for conducting the experiment.

- **HHAR** [Stisen *et al.*, 2015]. This dataset comprises 3-axis accelerometer data from 9 users, capturing six different daily activities, including walking, standing, walking upstairs, walking downstairs, sitting, and biking. The data is collected using 4 smartwatches and 8 smartphones. But for our experiment, we use only the data collected from smartphones with mixture sampling frequencies of 50Hz, 100Hz, 150Hz and 200Hz. It's worth noting that this dataset is noisier than others due to its non-continuous timestamps and unstable sampling rate. To construct samples from the dataset, we used a time window of size 128, resulting in a total of 697,742 samples. Then, we randomly sampled on each user's data to create a dataset with label shift between users. Also, we randomly selected several user pairs for conducting the experiment.

- **SleepEDF** [Goldberger *et al.*, 2000]. This dataset contains electroencephalography (EEG) collected from 20 healthy users. The data of SleepEDf are to be classified into five sleep stages: wake (W), non-rapid eye movement stages (N1, N2, N3), and rapid eye movement (REM). It is a univariate time series dataset of length 3000. In order to construct the training and test dataset with label shift, we adopted a sliding window of length 128 and conducted random sampling.

To illustrate the label distribution of each domain explicitly, we divided the number of samples in each category by the number of samples in the last category. Consequently, the value of the last category of each user in the result is 1. We presented the calculated results of class proportions for each user in our experiment in Table 7, Table 8, Table 9 and Table 10. Please note that in the presented results, a larger number indicates a higher proportion of the corresponding class. The WISDM dataset uses raw label distribution, while the UCIHAR, HHAR and SleepEDF datasets are sampled and the label distribution is changed.

In our experiment, we randomly select 26 pairs of source and target domains across the four datasets. For each pair, we conduct ten separate experiments for every baseline and the proposed DWLR, each time with a unique random seed. We take the average of these results as the final result, providing robust verification for our proposed method. For the balance of splitting result between different activity, we split each dataset randomly. Given sufficient data quantity, the proportion of data classifications remains similar pre- and post-split. The datasets we use in our experiment are significantly large enough to ensure this consistency.

D Detailed Experimental Result

We present the overall results of DWLR and other baselines on the WISDM, UCIHAR, and HHAR datasets in the Experiment section. On average, our proposed DWLR performs the best. For each dataset, we conducted multiple experiments by randomly selecting source and target domains. Each experiment was repeated ten times. The average and standard deviation of the experimental results are shown in Table 11. It can be observed that in the majority of experiments, DWLR outperforms other baseline methods in terms of performance.

The average std of all experimental results of our proposed DWLR is 1.99, which is the third smallest among the 14 methods. The baselines SLARDA and CLUDA have the smallest average std with 0.89 and 1.92, respectively. However, the average AUC-ROC of DWLR, SLARDA and CLUDA are 90.30, 70.41 and 78.18, respectively. It can be seen that although std of SLARDA and CLUDA is smaller than DWLR, their AUC-ROC is quite a bit smaller than DWLR. Therefore, the performance of our proposed DWLR method is more stable on these datasets.

In addition, we conducted a significance analysis for our proposed DWLR and the optimal baseline COAL. Among the 26 pairs of source domain and target domains, we find that DWLR significantly outperforms COAL in 24 of the experiments with a p_value < 0.05. COAL significantly outperforms DWLR in only one instance. Moreover, among these 26 experiments, DWLR achieves the best overall performance in 20 cases, and 15 of those cases are significantly better than any other best-performing baselines for that specific instance with a p_value < 0.05. Please note, the highest-performing baseline is not always COAL.

E Model Efficiency

We use MC-Dropout to compute the information gain approximately in our learnable reweighting module, which may increase the time complexity of DWLR. In order to explore the efficiency of the proposed model, the following is our analysis and experiment. The MC-Dropout dropout module uses the classifier, which consists of a dropout layer and a fully connected layer. The extra time complexity is $O(TFC)$, where T is the number of prediction, F and C are feature size and number of class respectively. For the whole model, except for the learnable reweighting module, the time cost consists of time and frequency domain encoder. The both time-domain and frequency-domain encoder are implement by CNN. The complexity of the encoder is $O(LMKF + nLKF)$, where L denotes the input length, M represents the number of channels, K signifies the convolutional kernel size, F denotes the channel count or feature dimension of the hidden layer, and n indicates the number of the hidden layers. Notably, this analysis excludes the consideration of pooling layers and the final fully connected layer. In our experiment setting, where $T = 100$, $C = 6$, $L = 128$, $F = 64$, $K = (5, 3)$ and $n = 4$, the extra time is theoretically smaller than the model itself, which indicates that the learnable reweighting module does not increase the time complexity of our model a lot.

We conduct experiments on three datasets to evaluate this, where setting of experiment is the same as that in Section 4. As the result in Table 12, Table 13 and Table 14 show, without learnable reweighting module (w/o $WNet(\cdot)$), each epoch costs an average of 6.66s. With learnable reweighting module, each epoch costs an average of 8.24s. The learnable reweighting module just uses 23.8% extra time, which is acceptable. The result of experiment also shows that learnable

Table 7: Category proportions of WISDM

User	5	6	8	12	13	18	20	21	24
Walk	2.1	9.3	2.1	2.5	1.6	3.1	1.4	1.4	2.2
Jog	2.2	21.4	3.4	5.1	3.1	5.4	3.1	2.7	5.0
Sit	0.5	2.8	0.9	0.9	0.3	0.6	3.7	0.4	0.2
Stand	0.5	1.1	1.1	0.7	0.4	0.8	1.3	0.8	0.2
Up	1.0	1.2	1.3	0.9	1.1	1.0	1.0	1.2	1.1
Down	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Table 8: Category proportions of UCIHAR

User	3	10	28	12	4	20	19	8	27	7	25	16
Walk	0.11	1.02	0.11	1.16	0.20	0.75	6.05	0.53	0.15	8.64	3.95	1.04
Sit	0.02	0.23	0.62	1.87	0.03	0.77	3.75	0.34	1.38	2.25	8.89	0.29
Lie	0.34	0.58	0.69	0.36	0.92	0.14	0.47	3.52	0.36	8.12	1.98	0.88
Stand	1.01	1.03	0.17	4.12	0.21	0.12	9.77	4.26	0.38	8.01	0.80	1.11
Up	0.27	0.25	0.83	7.82	0.80	0.26	0.39	4.90	0.04	5.08	4.86	1.59
Down	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

reweighting part does not significantly increase the training time of DWLR.

Table 9: Category proportions of HHAR

User	a	b	c	d	e	f	g	h	i
Bike	3.96	0.17	0.11	0.11	1.99	13.26	1.31	0.05	1.88
Sit	5.71	1.21	0.46	0.94	0.54	2.44	0.57	0.14	2.46
Down	0.21	0.45	1.21	0.39	1.62	10.81	4.45	0.35	1.61
Up	0.42	0.05	1.30	0.09	2.17	5.77	3.43	1.06	0.43
Stand	4.24	4.24	0.07	0.91	0.26	12.05	3.88	1.04	0.23
Walk	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 10: Category proportions of SleepEDF

User	0	1	3	4	7	8	10	12	13	15	17	19
W	29.71	0.60	2.50	0.24	16.65	1.20	0.16	1.02	0.90	1.96	2.03	0.91
N1	9.36	1.64	3.01	0.65	5.44	0.27	0.45	0.37	0.33	0.17	0.15	0.31
N2	50.01	10.00	25.01	2.44	5.00	1.52	3.15	0.08	2.90	0.08	0.05	0.10
N3	41.45	1.64	5.26	0.12	12.04	1.62	0.07	0.41	0.86	0.71	0.95	0.28
REM	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 11: AUC-ROC on WISDM, UCIHAR, HHAR and SleepEDF. Best in bold, second underlined.

s→t	DDNN	RDANN	VRADA	CoDATs	SLARDA	SASA	AdvSKM	SWL-Adapt	CLUDA	COAL	PAT	ImMMD	RAINCOAT	DWLR
WISDM														
20→6	83.02$_{4.41}$	87.37$_{2.32}$	95.05$_{1.35}$	94.29$_{2.61}$	84.70$_{0.19}$	84.81$_{15.6}$	82.31$_{5.67}$	90.10$_{4.54}$	86.37$_{1.69}$	91.11$_{3.20}$	86.89$_{4.28}$	83.56$_{0.79}$	90.55$_{1.12}$	95.57$_{2.72}$
20→12	73.76$_{3.81}$	68.52$_{5.83}$	76.43$_{8.10}$	86.71$_{6.90}$	79.42$_{0.13}$	71.95$_{6.96}$	90.95$_{3.46}$	87.57$_{5.93}$	76.13$_{3.86}$	85.45$_{6.9}$	59.15$_{11.7}$	92.17$_{3.21}$	95.61$_{0.82}$	97.09$_{1.09}$
24→21	53.06$_{5.86}$	83.63$_{2.18}$	85.55$_{2.83}$	91.40$_{4.28}$	86.79$_{1.08}$	63.05$_{4.88}$	69.80$_{9.22}$	90.70$_{2.24}$	83.16$_{3.75}$	88.99$_{4.17}$	68.08$_{8.52}$	88.75$_{4.45}$	72.47$_{4.01}$	94.36$_{2.54}$
5→18	68.00$_{3.82}$	87.30$_{3.38}$	86.70$_{1.97}$	83.03$_{5.54}$	72.20$_{0.04}$	85.29$_{1.87}$	78.11$_{4.47}$	82.95$_{10.93}$	79.75$_{2.13}$	88.69$_{3.71}$	63.70$_{7.21}$	79.75$_{5.44}$	89.36$_{4.33}$	96.07$_{1.44}$
8→24	74.43$_{3.27}$	77.87$_{3.22}$	77.82$_{4.39}$	78.12$_{6.76}$	68.32$_{0.36}$	72.04$_{2.57}$	65.54$_{7.25}$	83.94$_{5.76}$	75.31$_{3.10}$	82.89$_{3.42}$	63.64$_{8.56}$	69.74$_{10.1}$	84.27$_{4.09}$	93.83$_{4.72}$
13→6	71.34$_{10.9}$	87.80$_{4.01}$	92.51$_{5.19}$	87.78$_{5.54}$	82.20$_{0.32}$	60.73$_{25.7}$	72.54$_{3.52}$	81.11$_{1.98}$	81.18$_{2.04}$	91.79$_{3.16}$	79.28$_{3.53}$	79.17$_{5.03}$	87.80$_{3.83}$	95.52$_{2.68}$
UCIHAR														
3→10	57.17$_{5.56}$	60.94$_{2.59}$	68.82$_{5.18}$	72.06$_{2.64}$	52.67$_{0.42}$	57.58$_{1.97}$	67.07$_{2.70}$	52.15$_{4.62}$	61.58$_{2.90}$	72.09$_{3.41}$	59.60$_{2.43}$	65.46$_{6.21}$	71.81$_{3.97}$	79.06$_{3.41}$
28→12	63.02$_{4.23}$	76.04$_{2.29}$	75.40$_{5.83}$	83.10$_{5.08}$	71.33$_{1.28}$	65.86$_{1.24}$	83.96$_{2.67}$	89.26$_{5.57}$	81.91$_{3.62}$	79.11$_{8.63}$	67.41$_{4.67}$	66.93$_{4.16}$	76.72$_{2.62}$	92.16$_{3.15}$
8→20	66.86$_{4.28}$	70.96$_{5.46}$	81.73$_{2.32}$	80.41$_{3.95}$	71.00$_{0.46}$	70.71$_{2.97}$	79.99$_{1.97}$	87.73$_{3.44}$	83.78$_{1.78}$	83.94$_{5.44}$	75.82$_{5.97}$	81.96$_{6.71}$	82.91$_{4.37}$	94.28$_{2.64}$
4→19	68.03$_{2.29}$	85.08$_{2.21}$	87.79$_{1.84}$	89.96$_{2.40}$	89.45$_{1.37}$	69.77$_{1.93}$	89.32$_{1.99}$	68.85$_{6.78}$	88.39$_{2.19}$	90.67$_{6.10}$	76.69$_{4.94}$	86.09$_{2.61}$	88.25$_{1.14}$	82.68$_{2.84}$
27→7	63.95$_{5.58}$	98.04$_{1.53}$	98.49$_{3.09}$	93.69$_{2.17}$	96.30$_{0.19}$	89.32$_{1.17}$	82.43$_{6.52}$	97.54$_{0.85}$	97.75$_{0.70}$	99.19$_{6.73}$	80.71$_{5.67}$	99.49$_{0.18}$	83.55$_{3.05}$	99.92$_{0.03}$
25→16	61.84$_{1.97}$	79.69$_{5.45}$	74.48$_{2.78}$	83.12$_{1.84}$	65.70$_{0.64}$	69.86$_{1.75}$	83.26$_{3.26}$	80.27$_{0.77}$	84.92$_{0.99}$	86.34$_{6.73}$	61.12$_{5.64}$	82.01$_{2.49}$	75.58$_{1.39}$	93.20$_{1.55}$
16→10	61.52$_{3.05}$	68.90$_{4.13}$	77.21$_{3.09}$	77.17$_{3.46}$	75.35$_{0.46}$	59.75$_{2.88}$	86.72$_{3.39}$	64.64$_{4.76}$	76.49$_{3.23}$	66.19$_{4.44}$	63.60$_{3.44}$	71.41$_{6.17}$	92.50$_{0.92}$	87.15$_{4.09}$
HHAR														
a→c	58.85$_{1.34}$	69.09$_{11.3}$	70.60$_{6.58}$	83.26$_{3.00}$	61.78$_{0.90}$	70.39$_{3.51}$	68.36$_{2.86}$	68.37$_{3.73}$	86.06$_{1.00}$	77.95$_{5.91}$	66.20$_{2.73}$	66.48$_{1.17}$	70.00$_{4.17}$	89.29$_{2.25}$
e→b	55.89$_{1.49}$	82.21$_{3.61}$	71.38$_{10.4}$	68.06$_{2.28}$	66.04$_{0.86}$	54.27$_{9.63}$	73.72$_{6.61}$	80.44$_{3.12}$	91.33$_{1.62}$	86.81$_{3.32}$	74.19$_{1.90}$	75.88$_{7.85}$	74.02$_{3.14}$	90.02$_{3.21}$
d→g	76.96$_{3.02}$	74.25$_{10.9}$	74.24$_{10.8}$	75.57$_{3.88}$	88.76$_{0.30}$	69.61$_{5.56}$	89.03$_{2.35}$	93.76$_{0.73}$	90.97$_{0.82}$	87.94$_{3.85}$	79.83$_{3.52}$	88.91$_{1.29}$	80.90$_{1.82}$	94.33$_{0.56}$
e→d	70.47$_{0.92}$	88.93$_{4.38}$	85.44$_{6.12}$	93.01$_{1.68}$	91.58$_{0.75}$	57.77$_{17.7}$	79.19$_{8.26}$	93.73$_{2.01}$	95.81$_{0.46}$	96.16$_{0.65}$	78.66$_{2.56}$	89.58$_{0.28}$	77.76$_{8.51}$	97.86$_{0.53}$
h→f	83.12$_{16.6}$	87.30$_{4.71}$	85.05$_{3.41}$	94.58$_{0.91}$	65.57$_{0.48}$	65.73$_{7.85}$	75.46$_{9.11}$	84.10$_{1.18}$	85.56$_{1.11}$	86.48$_{5.10}$	68.97$_{2.79}$	88.49$_{1.49}$	83.16$_{1.09}$	87.15$_{0.79}$
i→g	59.09$_{6.92}$	71.29$_{8.28}$	66.80$_{12.5}$	73.95$_{5.41}$	87.28$_{0.13}$	66.64$_{13.1}$	89.24$_{2.34}$	88.66$_{5.14}$	89.69$_{1.73}$	87.38$_{2.36}$	78.48$_{3.23}$	73.87$_{6.13}$	73.69$_{4.91}$	94.19$_{0.47}$
h→i	63.99$_{12.3}$	71.25$_{7.79}$	64.56$_{8.49}$	60.86$_{1.38}$	66.15$_{1.04}$	71.98$_{4.83}$	71.44$_{2.83}$	78.60$_{1.70}$	80.91$_{2.98}$	82.67$_{1.65}$	65.48$_{2.51}$	70.56$_{5.85}$	76.63$_{7.45}$	89.60$_{2.15}$
SleepEDF														
4→8	70.86$_{0.78}$	65.73$_{9.64}$	52.74$_{2.57}$	78.94$_{0.87}$	50.85$_{3.49}$	-	68.02$_{0.92}$	79.50$_{1.53}$	58.84$_{1.66}$	79.50$_{1.01}$	55.58$_{3.97}$	73.44$_{3.14}$	62.28$_{2.81}$	81.19$_{1.83}$
10→17	67.72$_{0.70}$	70.30$_{11.2}$	50.75$_{0.48}$	77.07$_{0.70}$	48.73$_{0.60}$	-	60.19$_{3.51}$	70.87$_{0.27}$	53.29$_{0.60}$	74.17$_{0.82}$	57.50$_{2.25}$	66.85$_{0.88}$	62.43$_{2.16}$	82.75$_{1.58}$
7→13	75.89$_{1.88}$	87.26$_{0.90}$	77.09$_{8.41}$	90.50$_{0.41}$	52.82$_{2.32}$	-	83.97$_{1.07}$	81.48$_{0.50}$	64.63$_{1.30}$	86.40$_{0.95}$	61.36$_{0.73}$	82.72$_{0.55}$	68.88$_{2.94}$	89.64$_{3.54}$
3→12	69.06$_{0.48}$	53.96$_{4.27}$	50.83$_{0.32}$	77.67$_{1.06}$	51.48$_{0.67}$	-	65.65$_{4.57}$	74.03$_{0.39}$	57.05$_{0.45}$	78.02$_{1.05}$	56.54$_{0.56}$	73.83$_{0.61}$	60.79$_{1.63}$	79.36$_{0.46}$
15→1	75.21$_{0.41}$	82.42$_{0.47}$	67.77$_{2.56}$	86.38$_{0.11}$	47.29$_{2.06}$	-	75.97$_{1.81}$	77.32$_{0.45}$	60.64$_{3.67}$	81.58$_{1.23}$	54.44$_{2.25}$	77.18$_{0.78}$	62.86$_{1.50}$	82.73$_{0.70}$
19→0	71.92$_{0.61}$	85.14$_{0.51}$	62.92$_{2.29}$	87.30$_{0.31}$	56.86$_{2.62}$	-	75.14$_{2.37}$	79.62$_{0.38}$	61.24$_{0.43}$	84.23$_{0.98}$	60.46$_{1.14}$	81.71$_{0.80}$	61.42$_{3.04}$	88.84$_{0.84}$
Average	67.88	77.74	75.31	82.60	70.41	68.85	77.21	81.05	78.18	<u>84.45</u>	67.86	79.00	77.16	**90.30**

Table 12: Time cost (second) pre epoch on the WISDM dataset.

Source Target	20 6	20 12	24 21	5 18	8 24	13 6	Avg
w/o $WNet(\cdot)$	6.56	6.39	4.22	3.89	4.05	4.27	4.90
DWLR	8.05	8.28	4.66	4.76	5.51	5.56	6.14

Table 13: Time cost (second) pre epoch on the UCIHAR dataset.

Source Target	3 10	28 12	8 20	4 19	27 7	25 16	16 10	Avg
w/o $WNet(\cdot)$	4.27	4.24	4.93	4.13	4.36	4.88	4.74	4.51
DWLR	4.91	5.00	6.35	5.03	5.02	5.62	5.64	5.37

Table 14: Time cost (second) pre epoch on the HHAR dataset.

Source Target	a c	e b	d g	e d	h f	i g	h i	Avg
w/o $WNet(\cdot)$	9.28	9.75	11.71	11.83	10.03	10.75	10.63	10.57
DWLR	12.76	11.06	13.59	13.47	13.71	13.93	14.02	13.22

References

[Anguita *et al.*, 2013] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra Perez, and Jorge Luis Reyes Ortiz. A public domain dataset for human activity recognition using smartphones. In *Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning*, pages 437–442, 2013.

[Cover and Thomas, 1991] Thomas M. Cover and Joy A. Thomas. Elements of information theory. 1991.

[Cover, 1999] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[Dasgupta, 2004] Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In *NIPS*, 2004.

[Gal and Ghahramani, 2015] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference. *ArXiv*, abs/1506.02158, 2015.

[Gal *et al.*, 2017] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR, 06–11 Aug 2017.

[Goldberger *et al.*, 2000] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[Golovin and Krause, 2010] Daniel Golovin and Andreas Krause. Adaptive submodularity: A new approach to active learning and stochastic optimization. In *Annual Conference Computational Learning Theory*, 2010.

[Heckerman *et al.*, 1994] David E. Heckerman, John S. Breese, and Koos Rommelse. Troubleshooting under uncertainty. 1994.

[Houlsby *et al.*, 2011] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *ArXiv*, abs/1112.5745, 2011.

[Kwapisz *et al.*, 2011] Jennifer R Kwapisz, Gary M Weiss, and Samuel A Moore. Activity recognition using cell phone accelerometers. *ACM SigKDD Explorations Newsletter*, 12(2):74–82, 2011.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15:1929–1958, 2014.

[Stisen *et al.*, 2015] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM conference on embedded networked sensor systems*, pages 127–140, 2015.

[Um *et al.*, 2017] Terry T Um, Franz MJ Pfister, Daniel Pichler, Satoshi Endo, Muriel Lang, Sandra Hirche, Urban Fietzek, and Dana Kulić. Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 216–220, 2017.