Distilling Universal and Joint Knowledge for Cross-Domain Model Compression on Time Series Data

Qing Xu[1,3], **Min Wu**[1], **Xiaoli Li**[1,2] and **Kezhi Mao**[3], **Zhenghua Chen**[1,2*]
[1]Institute for Infocomm Research, A*STAR, Singapore
[2]Centre for Frontier AI Research, A*STAR, Singapore
[3]Nanyang Technological University, Singapore
{Xu_Qing, wumin, xlli}@i2r.a-star.edu.sg, EKZMao@ntu.edu.sg, chen0832@e.ntu.edu.sg

Abstract

For many real-world time series tasks, the computational complexity of prevalent deep leaning models often hinders the deployment on resource-limited environments (*e.g.,* smartphones). Moreover, due to the inevitable domain shift between model training (source) and deploying (target) stages, compressing those deep models under cross-domain scenarios becomes more challenging. Although some of existing works have already explored cross-domain knowledge distillation for model compression, they are either biased to source data or heavily tangled between source and target data. To this end, we design a novel end-to-end framework called **UN**iversal and jo**I**nt **K**nowledge **D**istillation (**UNI-KD**) for cross-domain model compression. In particular, we propose to transfer both the universal feature-level knowledge across source and target domains and the joint logit-level knowledge shared by both domains from the teacher to the student model via an adversarial learning scheme. More specifically, a feature-domain discriminator is employed to align teacher's and student's representations for universal knowledge transfer. A data-domain discriminator is utilized to prioritize the domain-shared samples for joint knowledge transfer. Extensive experimental results on four time series datasets demonstrate the superiority of our proposed method over state-of-the-art (SOTA) benchmarks. The source code is available at https://github.com/ijcai2023/UNI_KD.

1 Introduction

Deep learning (DL) models, particularly convolutional neural networks (CNNs), have achieved remarkable successes in various time series tasks, such as human activity recognition (HAR) [Wang *et al.*, 2019], sleep stages classification [Eldele *et al.*, 2021] and fault diagnosis [Zhao *et al.*, 2019]. These advanced DL models are often over-parameterized for better generalization on unseen data [Chen *et al.*, 2017]. However, deploying those models on a resource-limited environment (*e.g.,* smartphones and robots) is a common requirement for many real-world applications. The contradiction between model performance and complexity leads to the exploration of various model compression techniques, such as network pruning and quantization [Liang *et al.*, 2021], network architecture search (NAS) [Elsken *et al.*, 2019] and knowledge distillation (KD) [Hinton *et al.*, 2015]. Among them, KD has demonstrated its superior effectiveness and flexibility on enhancing the performance of a compact model (*i.e.*, Student) via transferring the knowledge from a cumbersome model (*i.e.*, Teacher). Another well-known problem in many time series tasks is the considerable domain shift between model development and deployment stages. For instance, due to the difference between subject's genders, ages or data collection sensors, a model trained on one subject (*i.e.*, source domain) might perform poorly on another subject (*i.e.*, target domain). Such domain disparity makes cross-domain model compression even more challenging.

Some recent works have already attempted to explore the benefits of applying unsupervised domain adaption (UDA) techniques during compressing cumbersome DL models by knowledge distillation. However, there are some drawbacks in these approaches. For instance, joint training of a teacher with UDA and student with KD would result in unstable loss convergence [Granger *et al.*, 2020], while the knowledge from teachers trained on source domain only [Yang *et al.*, 2020; Ryu *et al.*, 2022] is biased and limited. For cross-domain knowledge distillation, a proper teacher should possess the knowledge of both domains. In particular, the generalized knowledge (namely *Universal Knowledge*) across both domains is more critical in improving student's generalization capability on target domain. However, the aforementioned methods coarsely align teacher's and student's predictions, but neglect to disentangle the domain-shared knowledge (namely *Joint Knowledge*). Due to the existence of domain shift, introducing source-specific knowledge would result in poor adaptation performance.

Fig. 1 presents an example of our proposed universal and joint knowledge under cross-domain scenario. On the one hand, the universal knowledge across source and target domains as shown in Fig. 1(a) is important to improve the generalization capability for the student. On the other hand, the inevitable domain shift makes the distributions of source and target domains overlapped. Suppose that there exists a

*Corresponding author

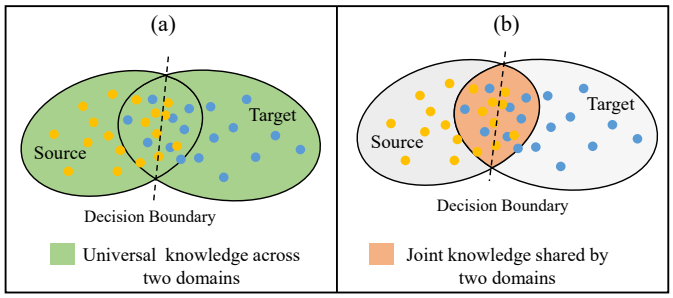

Figure 1: Illustration of universal and joint knowledge.

data-domain discriminator to correctly classify the samples into source or target domain. As depicted in Fig. 1(b), if some samples lie around its decision boundary, then these samples most likely possess some domain-shared information (*i.e.*, joint knowledge) which makes the discriminator incapable of correctly identifying their data domain (*i.e.*, source or target). Meanwhile, those samples which can be very confidently identified by the data-domain discriminator tend to possess domain-specific knowledge. Equally treating all samples like conventional KD approaches would be adverse to diminishing domain disparity, leading to poor generalization on target data. It is thus highly motivated to pay more attentions on samples with joint knowledge than samples with domain-specific knowledge for cross-domain knowledge distillation.

In this paper, we propose an innovative end-to-end model compression framework to improve student's generalization capability under the cross-domain scenarios. Specifically, we design a feature-domain discriminator to align teacher's and student's feature representations for effectively distilling the universal knowledge. Meanwhile, a data-domain discriminator is developed to prioritize the samples with joint knowledge across two domains. It assists to disentangle teacher's logits by paying more attentions on the samples with joint knowledge. Via an adversarial learning scheme, teacher's universal and joint knowledge can be effectively transferred to the compact student. Our main contributions are summarized as follows.

- A novel approach named universal and joint knowledge distillation (UNI-KD) approach is proposed to transfer teacher's universal and joint knowledge, which is an end-to-end framework for cross-domain model compression. Two discriminators (*i.e.*, feature-domain and data-domain discriminators) with an adversarial learning paradigm are designed to distill above two knowledge on feature-level and logit-level, respectively.
- We propose to disentangle teacher's logits with a data-domain discriminator by prioritizing the samples with joint knowledge across source and target domains. The joint knowledge could further boost the generalization ability of the compact student on target domain.
- Extensive experiments are conducted on four real-world datasets across three different time series classification tasks and the results demonstrate the superiority of our approach over other SOTA benchmarks.

2 Related Work

Knowledge distillation, as one of the most popular model compression techniques, has been widely explored in many applications. Originally, the knowledge from a complex teacher model is formulated as the logits soften by a temperature factor in [Hinton *et al.*, 2015]. Then, researchers extend the knowledge to the feature maps as they contain more low-level information than logits. Several works try to minimize the discrepancy between teacher's and student's feature representations via explicitly defining distance metrics, such as L2 [Romero *et al.*, 2014], attention maps [Zagoruyko and Komodakis, 2016], probability distributions [Passalis and Tefas, 2018] and inter-channel correlation matrices [Liu *et al.*, 2021]. On the contrary, other researchers exploit the adversarial learning scheme which implicitly forces the student to generate similar feature maps as the teacher [Gao *et al.*, 2019; Xu *et al.*, 2021; Xu *et al.*, 2022]. However, these approaches cannot be directly applied to cross-domain scenarios as they do not consider the domain shift during the compression.

To tackle the domain shift issue, various UDA approaches have been proposed. Generally, these techniques can be categorized into two types, namely discrepancy-based and adversarial learning-based. The former ones intend to minimize some statistical distribution measurements between source and target domains, *e.g.,* maximum mean discrepancy (MMD) [Tzeng *et al.*, 2014], the second-order statistics [Rahman *et al.*, 2020; Sun and Saenko, 2016] or higher-order moment matching (HoMM) [Chen *et al.*, 2020]. Whereas the adversarial learning-based ones attempt to learn domain-invariant representations via a domain discriminator [Ganin *et al.*, 2016; Long *et al.*, 2018; Shu *et al.*, 2018; Wilson *et al.*, 2020]. Although above mentioned UDA approaches have been successfully applied to many research areas, they seldom consider model complexity issue during domain adaptation, which is more practical for many time series tasks.

Recently, there are some attempts to jointly address model complexity and domain shift problems by integrating UDA techniques with KD for cross-domain model compression. In [Granger *et al.*, 2020], a framework was proposed to employ the MMD to learn domain-invariant representations for teacher and progressively distill the knowledge to the student on both source and target data. However, their approach would lead to difficulty on student's convergence. MobileDA [Yang *et al.*, 2020] performed the distillation on target domain with the knowledge from a source-only teacher. It leveraged the correlation alignment (CORAL) loss to learn domain-invariant representations for student. Similarly, a framework was proposed to perform adversarial learning and distillation on target domain with the knowledge from a source-only teacher in [Ryu *et al.*, 2022]. The teachers in [Yang *et al.*, 2020] and [Ryu *et al.*, 2022] are trained on source data only and the knowledge from such teachers is very biased and limited. Unlike them, our method employs a teacher trained on labeled source domain and unlabeled target domain, and we distill not only the universal feature-level knowledge across both domains but also the joint logit-level knowledge shared by both domains via an adversarial learning scheme.

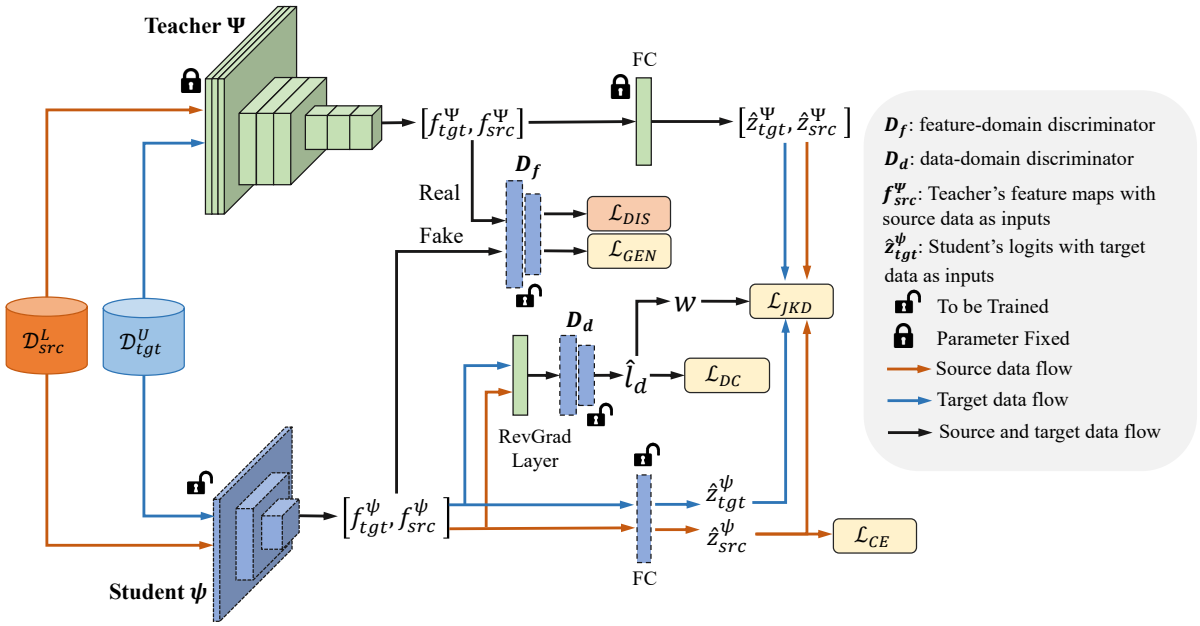

Figure 2: Illustration of proposed UNI-KD. A feature-domain discriminator D_f is employed to identify whether the input feature maps are from teacher Ψ or student ψ. It contributes to transferring the universal knowledge on feature level. A data-domain discriminator D_d is leveraged to identify whether the input feature maps are from source or target domain. The output of D_d is utilized to prioritize the samples for joint knowledge distillation on logits level. D_f is adversarially trained against D_d and student ψ.

3 Methods

3.1 Problem Definition

For the cross-domain model compression scenario, we first assume that a proper teacher Ψ is pre-trained on source and target domain data with SOTA UDA methods (*e.g.,* DANN [Ganin *et al.*, 2016]). Our objective is to improve the generalization capability of the compact student ψ on target domain data. Same as other UDA works, we assume that data come from two domains: *source* and *target*. Data from source domain are labeled, $\mathcal{D}^L_{src} = \{x^i_{src}, y^i_{src}\}^{N_{src}}_{i=1}$, and data from target domain are collected from a new environment without labels, $\mathcal{D}^U_{tgt} = \{x^i_{tgt}\}^{N_{tgt}}_{i=1}$. Here, N_{src} and N_{tgt} refer to the number of samples in source and target domains, respectively. Let $\mathcal{P}(X_{src})$ and $\mathcal{Q}(X_{tgt})$ be the marginal distributions of two domains. UDA problems assume that $\mathcal{P}(X_{src}) \neq \mathcal{Q}(X_{tgt})$ but $\mathcal{P}(Y_{src}|X_{src}) = \mathcal{Q}(Y_{tgt}|X_{tgt})$, indicating that source and target domains have different data distributions but share a same label space.

Our proposed UNI-KD is depicted as Fig. 2. In order to effectively compress the model for the cross-domain scenario, we formulate teacher's knowledge into two categories: feature-level universal knowledge across two domains and logit-level joint knowledge shared by two domains. These two types of knowledge are complementary to each other. We introduce two discriminators with the adversarial learning scheme to efficiently transfer these two types of knowledge.

3.2 Universal Knowledge Distillation

For cross-domain KD, we define the generalized knowledge across both domains as the universal knowledge. Such universal knowledge contains the fundamental characteristics existing in both source and target domains. In order to align teacher's and student's feature representations, adversarial learning scheme is utilized as it is capable of enhancing student's robustness [Maroto *et al.*, 2022]. Previous works also demonstrate that it could improve student's generalization capability on unseen data [Xu *et al.*, 2022; Chung *et al.*, 2020]. Motivated by this, we first design a feature-domain discriminator D_f for transferring the universal feature-level knowledge. D_f is a binary classification network to identify the source of input feature, *i.e.,* whether the input features $[f_{src}, f_{tgt}]$ come from Ψ or ψ.

We train D_f and student ψ in an adversarial manner. To be specific, in the first step, we fix student ψ and train D_f via loss $\mathcal{L}_{DIS}$ as Eq. (1) shows. A batch of 'real' samples (feature maps from teacher with source and target domain samples as input) is forwarded through D_f to calculate the loss $\log(D_f(f^{\Psi}_x))$. The gradients are calculated with back propagation. Then a batch of 'fake' samples (feature maps from student with same inputs) is forwarded through D_f to calculate loss $\log(1 - D_f(f^{\psi}_x))$. The gradients are accumulated to previous gradients from 'real' samples. At last, minimizing $\mathcal{L}_{DIS}$ maximizes the probability of correctly classifying the input features as 'real' (from teacher) or 'fake' (from student). The second step is to fix the D_f and train the student to generate similar feature maps as teacher. By minimizing $\mathcal{L}_{GEN}$ in Eq. (2), the discriminator D_f is expected to be incapable of telling whether the features are from Ψ or ψ. Alternately applying above two steps over all the training samples forces the student to learn similar feature maps as the teacher.

$$\mathcal{L}_{DIS} = -\mathbb{E}_{x\sim(\mathcal{D}_{src},\mathcal{D}_{tgt})}[\log(D_f(f_x^{\Psi}))+\log(1-D_f(f_x^{\psi}))], \quad (1)$$

$$\mathcal{L}_{GEN} = \mathbb{E}_{x\sim(\mathcal{D}_{src},\mathcal{D}_{tgt})}[\log(1 - D_f(f_x^{\psi}))]. \quad (2)$$

However, there are some challenges for the above adversarial learning scheme. Firstly, it can only transfer the universal knowledge but neglect the domain disparity between source and target domains, resulting in poor generalization on target domain. Secondly, the optimization of student ψ heavily relies on the accuracy of D_f and the student would be difficult to converge especially in the early training stage. Thus, we introduce three additional losses in the second step of adversarial learning scheme to cope with the above issues in the following section.

3.3 Joint Knowledge Distillation

Logits from teacher contain more information compared to one-hot labels and thus could be utilized as 'dark' knowledge for distillation [Hinton *et al.*, 2015]. However, we empirically found that simply combining conventional logits knowledge with feature distillation might lead to performance degradation. Due to the existence of domain shift, knowledge from teacher can be divided into domain-joint knowledge shared by two domains and domain-specific knowledge only existing in a particular domain. Since we feed both source and target domain samples to teacher and student, roughly minimizing their logits distributions would transfer both knowledge, leading to poor transferring performance.

Therefore, we intend to transfer the domain-joint knowledge but not domain-specific knowledge to the student. To achieve this, we utilize a data-domain discriminator D_d whose output is a binary probability vector $\hat{l}_d = [p_{c=0}, p_{c=1}]$. The element in above vector represents the probability of the inputs belonging to source domain ($c = 0$) or target domain ($c = 1$). We argue that the samples lying around the distribution boundary of source and target domains in the feature space are more generic than those samples which can be classified with high confidence. In other words, if the data-domain discriminator D_d cannot distinguish certain sample in the feature space, this sample most likely belongs to $\mathcal{P}(X_{src}) \bigcap \mathcal{Q}(X_{tgt})$ and possesses more domain-joint knowledge than others. Mathematically, $p_{c=0}$ and $p_{c=1}$ should be close to each other for these samples. Thus, we can utilize $\hat{l}_d$ to disentangle teacher's logits and let the student pay more attentions on those low-confidence samples during logits distillation. Specifically, for each sample i, we assign a different weight w_i to adjust its contribution for logits-level knowledge distillation in Eq. (3).

$$w_i = 1 - |p^i_{c=0} - p^i_{c=1}|. \quad (3)$$

Then, the loss $\mathcal{L}_{JKD}$ for joint knowledge distillation can be formulated as Eq. (4) where KL represents the Kullback-Lerbler divergence and $\boldsymbol{q}^s$, $\boldsymbol{q}^t \in \mathbb{R}^C$ are the predictions of student and teacher, respectively. C is the number of classes. Each element q_j in $\boldsymbol{q}^s$ or $\boldsymbol{q}^t$ is the probability of input sample belonging to the j^{th} class and $j \in \{1, ..., C\}$. q_j is a function of temperature factor τ used for smoothing the distribution and can be calculated via Eq. (5), and z_j represents model outputs (*i.e.,* logits) before the softmax layer.

$$\mathcal{L}_{JKD} = \tau^2 * \frac{1}{N}\sum_{i=0}^{N} w_i * KL(\boldsymbol{q}^s || \boldsymbol{q}^t), \quad (4)$$

$$q_j = \frac{exp(z_j/\tau)}{\sum_k exp(z_k/\tau)}. \quad (5)$$

It is foreseeable that the efficacy of disentangling domain-joint and domain-specific knowledge to a large extent depends on the accuracy of D_d. Therefore, we introduce a domain confusion loss $\mathcal{L}_{DC}$ to assist the training of D_d in Eq. (6), where l_d and $\hat{l}_d$ are the data domain labels and predictions of data-domain discriminator D_d, respectively.

$$\mathcal{L}_{DC} = -\mathbb{E}_{x\sim(\mathcal{D}_{src},\mathcal{D}_{tgt})}\Big[l_d * \log \hat{l}_d + (1 - l_d) * \log(1 - \hat{l}_d)\Big]. \quad (6)$$

Furthermore, overfitting might occur if we only utilize target domain samples to train the student. To avoid this, we also optimize the student on the source domain via a cross-entropy loss as Eq. (7) shows.

$$\mathcal{L}_{CE} = -\mathbb{E}_{(x_{src},y_{src})\sim\mathcal{D}_{src}} \sum_c \left[\mathbb{1}_{[}y_{src} = c] \log q_c^s\right]. \quad (7)$$

Then, the final loss for training the student ψ in the second step of adversarial learning paradigm is formulated as below:

$$\mathcal{L} = \mathcal{L}_{GEN} + (1 - \alpha) * \mathcal{L}_{DC} + \alpha * \mathcal{L}_{JKD} + \beta * \mathcal{L}_{CE}. \quad (8)$$

Here, we introduce two hyperparameters: α to balance the importance between domain confusion loss and logits KD loss, and β to adjust the contribution of $\mathcal{L}_{CE}$. For α, intuitively we intend to place more weights on UDA to achieve a good data-domain discriminator first and then gradually increase the importance of JKD as the training process goes on. This strategy can assist to stabilize student's training progress at the early stage. At each training epoch m, the corresponding value of α can be calculated by Eq. (9), where M is the total number of epochs, a and b are the starting and end values of α. In our experimental setting, we set $\alpha \in [0.1, 0.9]$. The value of α is exponentially increased with training epochs. Meanwhile, we utilize the grid search approach to identify the optimal value of parameter β.

$$\alpha = a * e^{\frac{m}{M}\log\frac{b}{a}}. \quad (9)$$

Algorithm 1 illustrates the details of our proposed UNI-KD for cross-domain model compression.

4 Experiments

4.1 Datasets

We evaluate our method with four commonly-used time series classification datasets across three different real-world applications: human activity recognition, sleep stage classification and fault diagnosis.

Algorithm 1 UNI-KD for cross-domain model compression

Input: A pre-trained teacher Ψ, a student model ψ, a feature-domain discriminator D_f, a data-domain discriminator D_d, source dataset $\mathcal{D}_{src}^{L}$ and target dataset $\mathcal{D}_{tgt}^{U}$.

for epoch $m \in [1, M]$ **do**
  Randomly shuffle $\mathcal{D}_{src}^{L}$ and $\mathcal{D}_{tgt}^{U}$;
  Update α for current epoch via Eq. (9);
  for each mini-batch in $X_b \in (\mathcal{D}_{src}^{L}, \mathcal{D}_{tgt}^{U})$ **do**
    Fix the parameters of ψ and D_d;
    Forward X_b through Ψ as 'real' samples;
    Forward X_b through ψ as 'fake' samples;
    Calculate $\mathcal{L}_{DIS}$ as Eq. (1);
    Optimize D_f via minimizing $\mathcal{L}_{DIS}$;
    Fix the parameters of D_f;
    Calculate $\mathcal{L}_{GEN}$ as Eq. (2) and $\mathcal{L}_{CE}$ as Eq. (7);
    Calculate $\mathcal{L}_{DC}$ as Eq. (6) and $\mathcal{L}_{JKD}$ as Eq. (4);
    Update ψ and D_d by minimizing $\mathcal{L}$ via Eq. (8);
  end for
end for

UCI HAR [Anguita *et al.*, 2013]: A smartphone is fixed on the waist of 30 experimental subjects and each subject is requested to perform six activities, *i.e.,* walking, walking upstairs, walking downstairs, standing, laying and sitting. The measurements from accelerometer and gyroscope are recorded for identifying each activity. Due to the variability between different subjects, we consider each subject as an independent domain and randomly select five cross-domain scenarios same as [Ragab *et al.*, 2023] for evaluation.

HHAR [Stisen *et al.*, 2015]: In this dataset, each subject conducts six activities, *i.e.,* biking, sitting, standing, walking, walking upstairs and downstairs. Since different brands of smartphones and smart watches are leveraged for data collection, this dataset is considered more challenging than **UCI HAR** in terms of domain shift. We follow [Liu and Xue, 2021] and select five cross-domain scenarios for evaluation.

FD [Lessmeier *et al.*, 2016]: Total 32 bearings are tested under four different operation conditions for rolling bearing fault diagnosis. The motor current signals are recorded for classifying bearing health status, *i.e.*, healthy, artificial damages (D1) and damages from accelerated lifetime tests (D2). We consider each operation condition as an independent domain and select five cross-domain scenarios for evaluation.

SSC [Goldberger *et al.*, 2000]: Sleep stage classification (SSC) dataset aims to utilize electroencephalography (EEG) signals to identify subject's sleep stage, *i.e.,* wake (W), non-rapid eye movement stage (N1, N2 and N3) and rapid eye movement (REM) stage. Each subject is considered as an independent domain and we select five scenarios for evaluation as previous studies [Eldele *et al.*, 2021].

4.2 Experiments Setup

For our method, a well-trained teacher is a pre-requisite to perform cross-domain knowledge distillation. We adopt **1D-CNN** as the backbone of our teacher and student models since it consistently outperforms other advanced backbones such as 1D residual network (1D-Resnet) and temporal convolutional neural network (TCN) as indicated in [Ragab *et al.*, 2023]. We leverage domain-adversarial training of neural networks (DANN) [Ganin *et al.*, 2016] approach to train the teacher. The student is a shallow version of teacher which has less filters. See **Supplementary** for network details of teacher and student. Table 1 summarizes the model complexity of teacher and student in terms of the number of trainable parameters and the number of floating-point operations (FLOPs). We can see that our compact student is about 15× smaller than its teacher in the aspect of parameters and requires less operations during inference. Furthermore, regarding the evaluation metric, considering the fact that accuracy metric might not be representative for imbalanced dataset, we adopt macro F1-score for all experiments as suggested in [Ragab *et al.*, 2023]. For all experiments, we repeat 3 times with different random seeds and report the averaged values.

	Teacher	Student	Compression Rate
No. of Parameters (million)	0.2009	0.0134	14.99 ×
No. of FLOPs (million)	9.328	0.661	14.11 ×

Table 1: Comparison of model complexity.

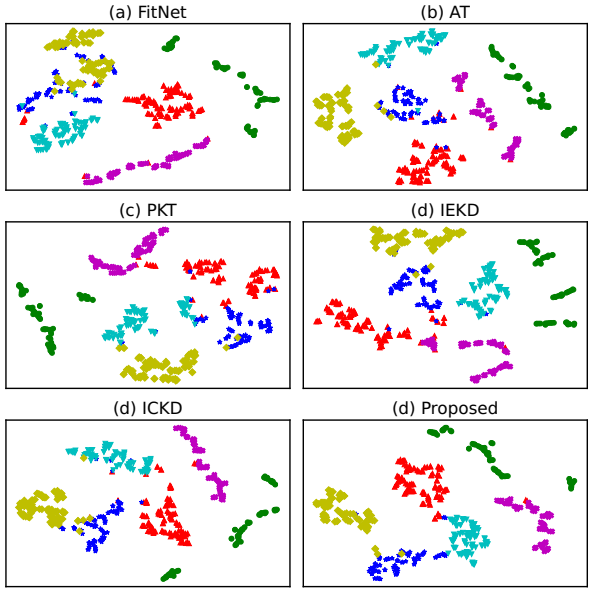

Figure 3: t-SNE of different feature distillation methods on HHAR.

4.3 Effectiveness of Adversarial Distillation

Feature-level knowledge from teacher's intermediate layers has already been known as a good extension of logit-based knowledge as DL models are able to learn multiple levels of feature representations [Gou *et al.*, 2021]. Various feature-based knowledge distillation approaches have been proposed in existing works. However, for cross-domain KD scenarios, we argue that adversarial learning could more effectively

Methods	UCI HAR Transfer Scenario						HHAR Transfer Scenario					
	2 → 11	6→23	7→13	9→18	12→16	Avg	0→6	1→6	2→7	3→8	4→5	Avg
Teacher	100.00	96.33	93.20	86.45	68.36	88.87	56.16	94.10	54.28	98.75	98.67	80.39
Student (src-only)	60.95	53.48	84.61	35.39	56.81	58.25	44.45	66.95	42.00	68.84	65.22	57.49
DDC	99.39	84.44	84.95	50.50	63.29	76.51	52.04	80.94	36.78	74.01	72.86	63.33
MDDA	99.44	93.72	**93.20**	61.20	59.67	81.45	**63.51**	92.44	51.56	86.60	93.01	77.42
HoMM	**100.00**	93.81	85.23	68.38	63.39	82.16	52.48	90.04	50.82	82.18	91.52	73.41
CoDATS	88.29	76.00	92.23	70.97	57.83	77.06	48.96	92.27	46.00	77.91	75.86	68.20
CDAN	**100.00**	91.43	91.71	65.21	60.44	81.76	45.36	92.75	50.79	91.60	86.27	73.35
DIRT-T	**100.00**	96.11	91.32	67.12	60.74	83.06	51.48	93.78	56.05	92.51	97.36	78.24
JKU	97.31	81.54	91.84	51.45	66.44	77.72	49.99	85.76	47.65	84.30	88.65	71.27
AAD	92.69	94.30	91.52	72.21	64.28	83.00	46.01	93.11	53.75	91.03	92.50	75.28
MobileDA	88.66	94.68	92.83	75.47	**66.67**	83.66	45.17	93.84	51.39	98.39	78.64	73.49
Proposed	**100.00**	**96.33**	**93.20**	**79.77**	64.91	**86.84**	46.66	**94.89**	**59.20**	**98.45**	**97.42**	**79.32**

Table 2: Marco F1-score on UCI HAR and HHAR across three independent runs.

Methods	FD Transfer Scenario						SSC Transfer Scenario					
	0→1	0→3	2→1	1→2	2→3	Avg	0→11	12→5	16→1	7→18	9→14	Avg
Teacher	84.86	82.39	99.59	90.34	99.34	91.30	54.20	66.45	64.78	71.48	72.85	65.95
Student (src-only)	35.49	40.46	87.88	75.28	91.13	66.05	33.02	50.78	52.25	57.75	62.05	51.17
DDC	47.31	59.03	89.23	73.57	89.31	71.69	**53.09**	52.29	57.39	63.75	68.53	59.01
MDDA	71.22	58.90	96.06	84.02	98.80	81.80	32.08	63.66	56.98	65.19	72.04	57.99
HoMM	55.44	48.18	95.66	76.39	96.96	74.53	44.88	55.47	56.89	63.66	68.87	57.95
CoDATS	55.72	64.10	89.27	87.58	91.14	77.54	36.69	61.18	61.65	64.47	62.06	57.21
CDAN	71.62	69.53	97.52	89.85	94.48	84.60	33.46	63.72	62.04	65.62	63.53	57.67
DIRT-T	76.98	75.92	**99.26**	**92.74**	98.95	88.77	31.71	**65.53**	62.80	69.87	69.47	59.88
JKU	40.32	51.44	88.20	79.51	85.80	69.05	41.25	51.32	55.34	66.01	65.31	55.85
AAD	64.06	67.50	86.94	79.27	91.38	77.83	51.40	58.60	55.51	68.77	59.64	58.78
MobileDA	40.32	43.30	96.82	76.99	85.10	68.51	53.10	51.86	55.60	65.06	67.63	58.65
Proposed	**78.85**	**82.68**	97.29	92.14	**99.34**	**90.06**	44.48	60.13	**62.99**	**71.03**	**72.21**	**62.17**

Table 3: Marco F1-score on Bearing FD and SSC across three independent runs.

transfer the universal knowledge from teacher to student. To prove it, we first compare the adversarial feature KD with some commonly-used feature distillation approaches: Fitnet [Romero *et al.*, 2014], PKT [Passalis and Tefas, 2018], AT [Zagoruyko and Komodakis, 2016], IEKD [Huang *et al.*, 2021] and ICKD [Liu *et al.*, 2021]. Please refer to **Supplementary** for details of each approach.

It is worth noting that we adapt above methods to our joint logit-level knowledge distillation. The only difference between these methods and ours is the feature distillation part. We utilize the t-SNE to visualize the learnt feature maps of above feature distillation approaches on **HHAR** dataset. As depicted in Fig. 3, the features learned from our proposed UNI-KD are more concentrated and all classes are well separated without overlapping. These observations demonstrate that the adversarial feature KD scheme could efficiently transfer the universal knowledge to the student for the cross-domain scenario. More t-SNE visualization results on other three datasets can be found in **Supplementary**.

4.4 Benchmark Results and Discussions

We compare our method with SOTA UDA algorithms, including some discrepancy-based and adversarial learning-based approaches as follows: deep domain confusion (DDC) [Tzeng *et al.*, 2014], minimum discrepancy for domain adaptation (MDDA) [Rahman *et al.*, 2020], higher-order moment matching (HoMM) [Chen *et al.*, 2020], convolutional deep domain adaptation for time series data (CoDATS) [Wilson *et al.*, 2020], conditional adversarial domain adaptation (CDAN) [Long *et al.*, 2018] and decision-boundary iterative refinement training with a teacher (DIRT-T) [Shu *et al.*, 2018]. Note that these UDA methods are directly applied to compact student. Meanwhile, we also include the results of some advanced works which integrate UDA with KD as follows: joint knowledge distillation and unsupervised domain adaptation (JKU) [Granger *et al.*, 2020], adversarial adaptation with distillation (AAD) [Ryu *et al.*, 2022] and MobileDA [Yang *et al.*, 2020]. See **Supplementary** for the details of benchmark approaches. Besides, the performance of teacher and student trained on source domain ("Student src-only") are also reported as they can be considered as the upper and lower limits of the compact student. Tables 2 and 3 summarize the evaluation results over different domain adaptation scenarios on four datasets. More experimental results on additional transfer scenarios can be found in **Supplementary**.

From Tables 2 and 3, some observations can be found. Firstly, directly applying UDA on a compact student, either the discrepancy-based or adversarial-based UDA approaches, could somehow boost the student's performance on target domain in most of cross-domain scenarios as expected. However, in certain transfer scenarios, negative transfer might also occur. For instance, students trained with the DDC method on 2→7 for **HHAR** and DIRT-T on 0→11 for **SSC** even achieve lower performance than the student trained on source only,

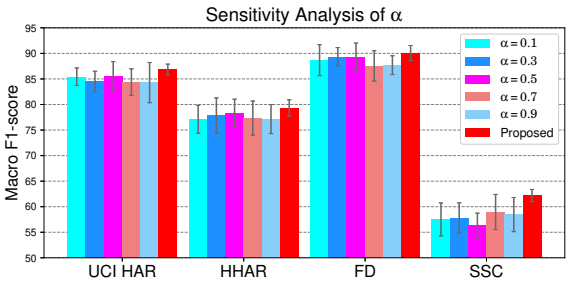

Figure 4: Performance with different α values.

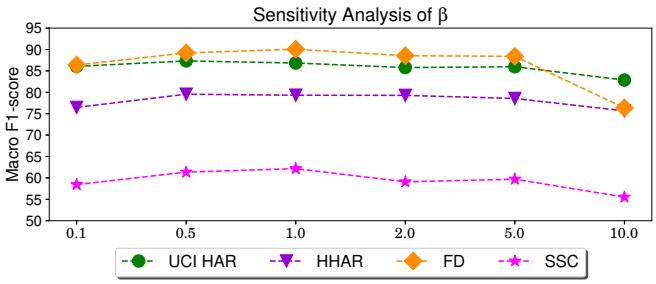

Figure 5: Performance with different β values.

indicating that those UDA methods might suffer from the inconsistency problem on different domain adaptation tasks. Secondly, the JKU method performs worse than AAD and MobileDA in most of transfer scenarios. The possible reason is that the teacher in JKU is trained together with the student, and it would lead to convergence problem for the student when progressively distilling the knowledge. Moreover, since the teachers used in AAD and MobileDA are only trained on source domain data, the knowledge from these teachers is very limited and biased to source domain, resulting performance degradation. Thirdly, those jointly optimizing KD and UDA methods even under-perform UDA methods like DIRT-T in most transfer scenarios, indicating that improperly transferring teacher's knowledge might decrease student's generalization capability on target data.

Lastly, our method consistently performs the best in terms of averaged macro F1-score over all the four datasets and outperforms other benchmarks on most of transfer scenarios. Moreover, compared with other joint KD and UDA methods, our UNI-KD can significantly reduce the performance gaps between teacher and student with the proposed universal and joint knowledge. This observation demonstrates the effectiveness of our method on the cross-domain model compression scenario. Via the evaluation on various time series domain adaptation tasks, our method can robustly compress the DL models with competitive performance as the complex teacher.

4.5 Ablation Study

There are three key components in our proposed approach: feature-domain discriminator D_f, data-domain discriminator D_d and joint knowledge distillation (JKD). To analyze the contribution of each component, we conduct the ablation study as Table 4 shows. Moreover, to validate the effectiveness of proposed JKD, we also include the standard KD (SKD) in Table 4. Some conclusions can be observed from Table 4 . First, applying universal feature-level KD via integrating D_f upon D_d could consistently improve student's performance over all datasets. However, integrating JKD upon D_d unexpectedly causes performance degradation in **HHAR** and **SSC** compared with only employing D_d. The possible reason is that logits contain less information than feature maps. Aligning teacher's and student's features could assist the student to learn more general representations than logits. Moreover, our UNI-KD suggests that these two types of knowledge are complementary to each other, and combining them can yield better performance as the last row shows. Furthermore, from the last two rows in Table 4, we can conclude that compared with standard KD, our proposed JKD is more effective in the cross-domain scenario.

D_d	D_f	JKD	SKD	HAR	HHAR	FD	SSC
Source Only				58.25	57.39	66.05	51.17
✓				82.42	76.03	83.45	60.12
✓	✓			85.99	78.11	87.97	61.79
✓		✓		86.48	73.15	69.38	58.21
✓	✓		✓	86.31	79.01	86.29	59.68
✓	✓	✓		**86.84**	**79.32**	**90.06**	**62.17**

Table 4: Ablation Study for the Proposed UNI-KD.

4.6 Sensitivity Analysis

There are two hyperparameters (*i.e.*, α and β) in our proposed approach as shown in Eq. (8). For α, we propose to gradually increase the importance of JKD loss during the training process as our method relies on an accurate data-domain discriminator. To validate its effectiveness, we compared our adaptive α method with fixed α values as illustrated in Fig. 4. We can see that the proposed adaptive α could consistently achieve better results than fixed α.

For the hyperparameter β, we utilize the grid search approach to identify the optimal values for different datasets. Fig. 5 illustrates the performance under different values of β. We can see that higher value of β will result in over-fitting to source data and would decrease the performance as expected. The optimal values for β is around $[0.5, 1.0]$. In all our experiments, we set $\beta = 0.5$ for dataset **UCI HAR** and **HHAR** and $\beta = 1.0$ for dataset **FD** and **SSC**.

5 Conclusion

In this paper, we propose an end-to-end framework for cross-domain knowledge distillation. Our method utilizes an adversarial learning paradigm with a feature-domain discriminator and a data-domain discriminator to improve student's generalization capability on target domain. With our proposed approach, the universal knowledge across both domains and the joint knowledge shared by both domains from a pre-trained teacher can be effectively transferred to a compact student. The experimental results show that the proposed UNI-KD can not only reduce the model complexity but also address domain shift issue.

Acknowledgments

This work is supported by the Agency for Science, Technology and Research (A*STAR) Singapore under its NRF AME Young Individual Research Grant (Grant No. A2084c1067) and A*STAR AME Programmatic Funds (Grant No. A20H6b0151).

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.

[Chen *et al.*, 2017] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017.

[Chen *et al.*, 2020] Chao Chen, Zhihang Fu, Zhihong Chen, Sheng Jin, Zhaowei Cheng, Xinyu Jin, and Xian-Sheng Hua. Homm: Higher-order moment matching for unsupervised domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3422–3429, 2020.

[Chung *et al.*, 2020] Inseop Chung, SeongUk Park, Jangho Kim, and Nojun Kwak. Feature-map-level online adversarial knowledge distillation. In *International Conference on Machine Learning*, pages 2006–2015. PMLR, 2020.

[Eldele *et al.*, 2021] Emadeldeen Eldele, Zhenghua Chen, Chengyu Liu, Min Wu, Chee-Keong Kwoh, Xiaoli Li, and Cuntai Guan. An attention-based deep learning approach for sleep stage classification with single-channel eeg. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:809–818, 2021.

[Elsken *et al.*, 2019] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *The Journal of Machine Learning Research*, 20(1):1997–2017, 2019.

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

[Gao *et al.*, 2019] Liang Gao, Haibo Mi, Boqing Zhu, Dawei Feng, Yicong Li, and Yuxing Peng. An adversarial feature distillation method for audio classification. *IEEE Access*, 7:105319–105330, 2019.

[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.

[Gou *et al.*, 2021] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[Granger *et al.*, 2020] Eric Granger, Madhu Kiran, Jose Dolz, Louis-Antoine Blais-Morin, et al. Joint progressive knowledge distillation and unsupervised domain adaptation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop*, 2015.

[Huang *et al.*, 2021] Zhen Huang, Xu Shen, Jun Xing, Tongliang Liu, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-Sheng Hua. Revisiting knowledge distillation: An inheritance and exploration framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3579–3588, 2021.

[Lessmeier *et al.*, 2016] Christian Lessmeier, James Kuria Kimotho, Detmar Zimmer, and Walter Sextro. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *PHM Society European Conference*, volume 3, 2016.

[Liang *et al.*, 2021] Tailin Liang, John Glossner, Lei Wang, Shaobo Shi, and Xiaotong Zhang. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, 461:370–403, 2021.

[Liu and Xue, 2021] Qiao Liu and Hui Xue. Adversarial spectral kernel matching for unsupervised time series domain adaptation. In *IJCAI*, pages 2744–2750, 2021.

[Liu *et al.*, 2021] Li Liu, Qingle Huang, Sihao Lin, Hongwei Xie, Bing Wang, Xiaojun Chang, and Xiaodan Liang. Exploring inter-channel correlation for diversity-preserved knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8271–8280, 2021.

[Long *et al.*, 2018] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[Maroto *et al.*, 2022] Javier Maroto, Guillermo Ortiz-Jiménez, and Pascal Frossard. On the benefits of knowledge distillation for adversarial robustness. *arXiv preprint arXiv:2203.07159*, 2022.

[Passalis and Tefas, 2018] Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 268–284, 2018.

[Ragab *et al.*, 2023] Mohamed Ragab, Emadeldeen Eldele, Wee Ling Tan, Chuan-Sheng Foo, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Adatime: A

benchmarking suite for domain adaptation on time series data. *ACM Trans. Knowl. Discov. Data*, mar 2023.

[Rahman *et al.*, 2020] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. On minimum discrepancy estimation for deep domain adaptation. In *Domain Adaptation for Visual Understanding*, pages 81–94. Springer, 2020.

[Romero *et al.*, 2014] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

[Ryu *et al.*, 2022] Minho Ryu, Geonseok Lee, and Kichun Lee. Knowledge distillation for bert unsupervised domain adaptation. *Knowledge and Information Systems*, 64(11):3113–3128, 2022.

[Shu *et al.*, 2018] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. *arXiv preprint arXiv:1802.08735*, 2018.

[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.

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

[Tzeng *et al.*, 2014] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

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

[Wilson *et al.*, 2020] Garrett Wilson, Janardhan Rao Doppa, and Diane J Cook. Multi-source deep domain adaptation with weak supervision for time-series sensor data. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1768–1778, 2020.

[Xu *et al.*, 2021] Qing Xu, Zhenghua Chen, Keyu Wu, Chao Wang, Min Wu, and Xiaoli Li. Kdnet-rul: A knowledge distillation framework to compress deep neural networks for machine remaining useful life prediction. *IEEE Transactions on Industrial Electronics*, 69(2):2022–2032, 2021.

[Xu *et al.*, 2022] Qing Xu, Zhenghua Chen, Mohamed Ragab, Chao Wang, Min Wu, and Xiaoli Li. Contrastive adversarial knowledge distillation for deep model compression in time-series regression tasks. *Neurocomputing*, 485:242–251, 2022.

[Yang *et al.*, 2020] Jianfei Yang, Han Zou, Shuxin Cao, Zhenghua Chen, and Lihua Xie. Mobileda: Toward edge-domain adaptation. *IEEE Internet of Things Journal*, 7(8):6909–6918, 2020.

[Zagoruyko and Komodakis, 2016] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016.

[Zhao *et al.*, 2019] Rui Zhao, Ruqiang Yan, Zhenghua Chen, Kezhi Mao, Peng Wang, and Robert X Gao. Deep learning and its applications to machine health monitoring. *Mechanical Systems and Signal Processing*, 115:213–237, 2019.