10-2021

Differentiated learning for multi-modal domain adaptation

Jianming LV

Kaijie LIU

Shengfeng HE
Singapore Management University, shengfenghe@smu.edu.sg

Differentiated Learning for Multi-Modal Domain Adaptation

Jianming Lv*
South China University of Technology
Guangzhou, Guangdong, China
jmlv@scut.edu.cn

Kaijie Liu
South China University of Technology
Guangzhou, Guangdong, China
cskaijieliu@gmail.com

Shengfeng He
South China University of Technology
Guangzhou, Guangdong, China
shengfenghe7@gmail.com

ABSTRACT

Directly deploying a trained multi-modal classifier to a new environment usually leads to poor performance due to the well-known domain shift problem. Existing multi-modal domain adaptation methods treated each modality equally and optimize the sub-models of different modalities synchronously. However, as observed in this paper, the degrees of domain shift in different modalities are usually diverse. We propose a novel Differentiated Learning framework to make use of the diversity between multiple modalities for more effective domain adaptation. Specifically, we model the classifiers of different modalities as a group of teacher/student sub-models, and a novel Prototype based Reliability Measurement is presented to estimate the reliability of the recognition results made by each sub-model on the target domain. More reliable results are then picked up as teaching materials for all sub-models in the group. Considering the diversity of different modalities, each sub-model performs the Asynchronous Curriculum Learning by choosing the teaching materials from easy to hard measured by itself. Furthermore, a reliability-aware fusion scheme is proposed to combine all optimized sub-models to support final decision. Comprehensive experiments based on three multi-modal datasets with different learning tasks have been conducted, which show the superior performance of our model while comparing with state-of-the-art multi-modal domain adaptation models.

CCS CONCEPTS

• **Information systems** → **Multimedia streaming**.

KEYWORDS

Differentiated learning; Multi-modal analysis; Domain adaptation

ACM Reference Format:
Jianming Lv, Kaijie Liu, and Shengfeng He . 2021. Differentiated Learning for Multi-Modal Domain Adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475660

*Corresponding Author

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475660

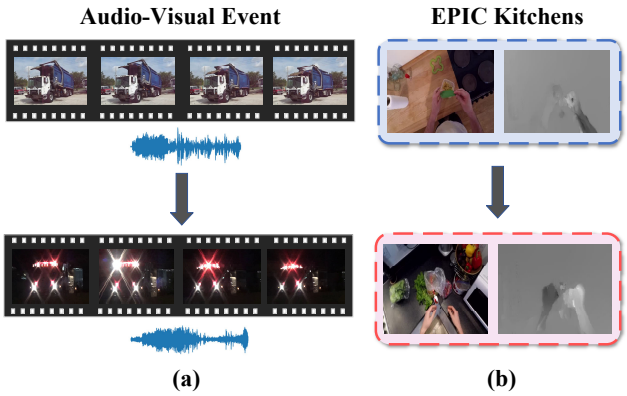

Figure 1: The domain shift in two multi-modal cross-domain scenarios. (a) The event recognition dataset (AVE) [37] with the image and audio modalities. (b) The action recognition dataset (EPIC Kitchens) [6] with the image and optical flow modalities.

Table 1: The accuracy of the classifiers on different modalities when transferring the models from the source domain to the target domain.

Dataset	Modal	Source	Target	Decline
AVE	Image	86.62	17.46	↓ 69.16
	Audio	85.79	42.83	↓ 42.96
EPIC Kitchens	RGB	63.32	35.33	↓ 27.99
	Flow	67.54	48.87	↓ 18.67

1 INTRODUCTION

Recently, multi-modal data mining [27] [2] [35] has gained more and more attention, which focuses on utilizing multiple modalities of data to improve the performance of pattern recognition. Different from the traditional supervised learning on single-modal data, multi-modal learning on huge amount of unlabeled data is more close to the real case of human learning, which brings new challenges.

Due to the well-known domain shift problem, directly applying a trained multi-modal classifier to a new environment usually leads to poor performance. How to make use of the unlabeled multi-modal data in the target domain to incrementally optimize the model is called as the multi-modal domain adaptation (MDA) problem, which is a natural extension of the the traditional single-modal domain adaptation (SDA) [9]. Compared with SDA, MDA has more potential to utilize the correlation between different modalities to enhance the performance.

Existing MDA methods can be roughly divided into two categories: *Adversarial Learning* and *Co-training*. Specifically, the *Adversarial Learning* [25, 29] aims to reduce the domain shift by utilizing the Generative Adversarial Networks to extract the domain-invariant representation of data. Since the idea of adversarial training to reduce domain shift has been well studied in traditional single-modal SDA [11, 13, 39], the research of [25, 29] can be viewed as an extension of these previous works by reducing the domain shift on each modality. Another way of MDA is the *Co-training* [3, 5, 40, 43], which constructs the classifier as a fusion of the sub-models from different modalities. Each sub-model is treated as a student model to learn knowledge from the others, which assign pseudo labels to the samples with the highest posterior probability and use them to train the student model. In above methods, all sub-models of different modalities are treated equally, and learn from the samples in the target domain in the same order without considering the diversity of different modalities.

However, according to our observation as shown in Fig.1, the degrees of domain shift in different modalities are usually diverse in the MDA cases, which lead to the diverse performance of the sub-models of different modalities in the target domain. Fig.1 (a) shows an example of the AVE dataset, where the truck image of source domain is clear and easy to distinguish, but in the target domain the poor lighting condition makes it difficult to judge. Compared with the image modality, the domain shift of audio is much smaller. The results in Table 1 further show that the accuracy of the image modality drops by 69.16% after cross-domain, while the audio modality only drops by 42.96%. Similar diversity can be observed in Fig.1 (b) about the EPIC Kitchens dataset, where the domain shift of the RGB modality is much larger than the Optical Flow. The results in Table 1 also confirm the diverse domain shift in different modalities.

This motivates us in following two aspects: 1) If we can accurately measure the reliability of the recognition results of each sub-model, we can achieve more precise pseudo labels for incremental learning; 2) Due to the diversity of the abilities of different sub-models, it is not optimal for all sub-models to learn synchronously. Just like a learning group of students with different knowledge levels, making personalized learning plan for each one may be a more proper way for efficient improvement.

Based on above thinking, we propose a novel *Differentiated Learning* framework for multi-modal domain adaptation, namely ***DLMM***, to organize an asynchronous learning group of the sub-models of different modalities. Each sub-model contributes to the teaching materials according to their reliability, which is measured by the similarity between the testing instance and the prototypes learned in the training dataset. Meanwhile, each sub-model performs the Curriculum Learning asynchronously by choosing teaching materials from easy to hard measured by itself. This is analog to the human learning principle of teaching students in accordance of their aptitude. Furthermore, all optimized sub-models are combined by a reliability-aware fusion scheme to make the final decision. Experiments based on three multi-modality datasets show the superior performance of this differentiated learning solution.

Main contributions of this paper are as follows:

(1) A novel *Differentiated Learning* framework is proposed for multi-modal domain adaptation, which organizes an asynchronous learning group of the sub-models of different modalities for incremental optimization on unlabeled data.

(2) A novel *Prototype based Reliability Measurement* is proposed to estimate the reliability of each transferred sub-model on unlabeled data in the target domain, and a *Reliability-aware Fusion* scheme is proposed to combine the sub-models to make the final decision. Experiments show that the *Prototype based Reliability Measurement* can significantly outperforms existing uncertainty estimation methods, such as the posterior probability-based [3, 43], entropy-based [14] and margin sampling methods [33].

(3) Distinct from the traditional synchronous optimization of all modalities, an *Asynchronous Curriculum Learning* strategy is adopted on each sub-model to choose teaching materials in accordance of their aptitude. Experiments on three multi-modal datasets with different learning tasks are conducted to verify the superior performance.

2 RELATED WORK

This section discusses related literatures including domain adaptation, uncertainty estimation, noisy label learning and curriculum learning.

2.1 Domain Adaptation

Single-modal Domain Adaptation (SDA). Most of the existing domain adaptation methods are designed for single-modal data, and can be summarized into three categories. (1) Discrepancy-based approaches [22, 23, 36]: matching mid-level representations of source and target domains by minimizing their discrepancy. (2) Adversarial-based approaches [11, 13, 39]: reducing the difference of the feature distributions between the source domain and target domain by introducing the adversarial learning with the domain classifiers. (3) Self-training-based approaches [31, 38, 44, 45]: selecting the samples in the target domain with higher confidence and assigning them with pseudo-labels for incremental training.

Multi-modal Domain Adaptation (MDA). There is no much research on multi-modal domain adaptation, but it has attracted more and more attention recently. Most of existing MDA methods are extended from the original SDA methods, and are roughly divided into two categories: Adversarial-based methods and Co-training. (1) Adversarial-based approaches [25, 29]. Different from the single-modal version in SDA, the extracted features of different modalities are used at the same time to obfuscate the domain classifier. In particular, in [25], the multi-modal alignment features are used as self-supervised signals. (2) Co-training approaches [3, 5, 40, 43]. In the original Co-training method [3], multi-view classifiers are applied, and each view improves the performance by learning the pseudo-labels made by others. The work [5] applies this method to multi-modal object detection. [43] further improves Co-training by drawing samples with high prediction probability without replacement. More recently, [40] introduces the Bayesian uncertainty as the weight of pseudo-labels in co-training.In above methods, all sub-models of different modalities are treated equally, and learn from the samples in the target domain in the same order without considering the diversity of different modalities.

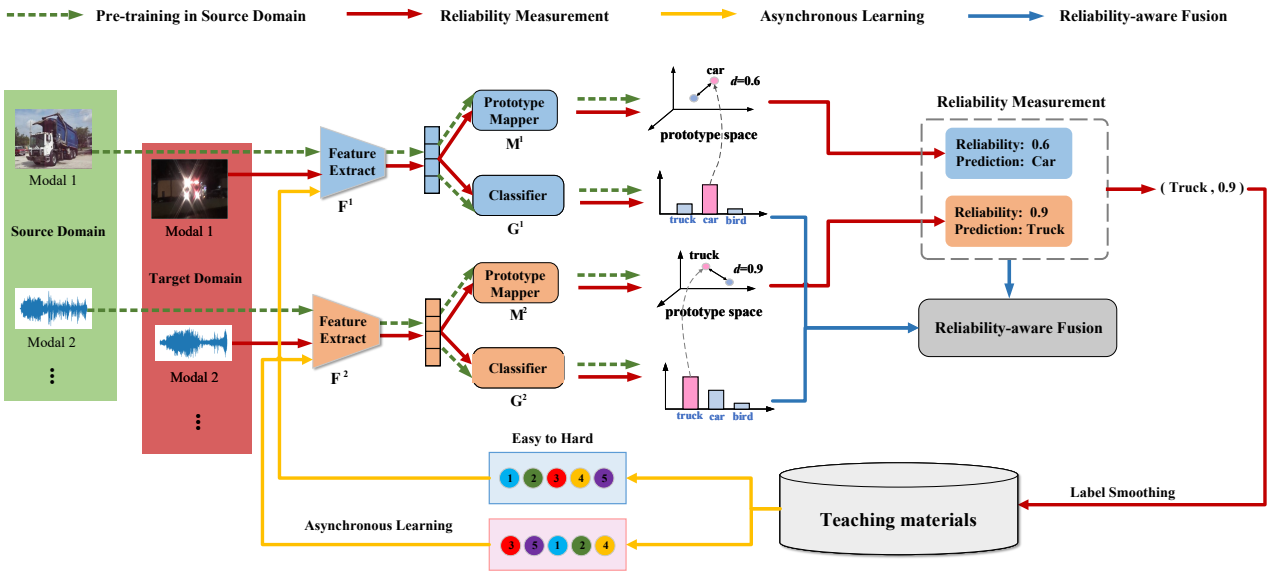

Figure 2: The Differentiated Learning framework for multi-modal domain adaptation, which contains four main stages: S1) Pre-training in the source domain to optimize the classifiers and achieve the prototypes. S2) Reliability Measurement of the pseudo labels in the target domain based on learnt prototypes, and the most reliable ones are selected as the teaching materials. S3) Asynchronous Learning is performed on each sub-model to learn the teaching materials from easy to hard. S4) Reliability-aware Fusion is adopted to output the final decision.

2.2 Uncertainty Estimation

Uncertainty estimation aims to measure the reliability of the model's decision-making. The most commonly used uncertainty estimation index is to use the probability value normalized by softmax [42]. However, the normalized probability may offer wrong estimation when measuring the transferred classifier in a new domain totally different from the training environment. In this case, the probability to all classes may be quite low due to the high uncertainty, but the normalized softmax can still achieve a relatively high score on a certain class. Another commonly used uncertainty estimation indicators include information entropy-based on posterior probability [14], interval-based on posterior probability [7], Bayesian uncertainty [17] and so on. Inspired by the prototype network [21, 34], we use the Gaussian kernel function representation of the similarity between a sample and the prototype of the predicted class as a measurement of the reliability of model decision. The superior performance of this reliability measurement will be shown in the following experiments comparing with other metrics.

2.3 Noisy Label Learning

The noisy label learning [26, 30, 41] is proposed to lower the risk of learning false pseudo-labels during incremental training on unlabeled dataset. An effective way to alleviate the damage caused by incorrect labels is to use label smoothing [44], which can prevent the model from making overconfident decisions after training. Different from the previous label-smoothing methods which require to manually set the smoothness in advance, the method proposed in this paper performs label smoothing based on the reliability measurement of recognition results.

2.4 Curriculum Learning

The core idea of curriculum learning [1, 16, 19] is that the model should start learning from simple samples and gradually transition to difficult samples, which make the training process more stable. That means the model should give priority to samples with high confidence in pseudo-labels when training on unlabeled data. Curriculum learning has been proven to be an effective method in the field of domain adaptation [44, 45] and semi-supervised learning [24].

3 DIFFERENTIATED LEARNING

Fig 2 shows the overview of our proposed *Differentiated Learning* scheme for multi-modal domain adaptation, namely ***DLMM***. *DLMM* is composed of four main steps. Firstly, **the Pre-training Stage** is to pre-train the multi-modal sub-models on the source domain on multi-tasks, including the original supervised classification task and an attached prototype extraction task to achieve the multi-modal prototypes of each class. Then in the **Prototype based Reliability Measurement Stage** on the target domain, the reliability of the prediction results of each sub-model on the unlabeled data is measured by the similarity with the pre-trained prototypes, and the most reliable samples with pseudo labels are picked up as teaching materials. Next, in the **Asynchronous Learning Stage**, each sub-model asynchronously selects the teaching materials from easy to hard based on its prediction loss self-adaptively. Finally, in the **Reliability-aware Fusion Stage**, the reliability-based weighted fusion of the multi-modal classifiers is adopted to output the final decision.

In the following section, we will firstly describe the problem definition of multi-modal domain adaptation **(MDA)**, and then detail each stage of the *Differentiated Learning* scheme.

3.1 Multi-modal Domain Adaptation

The dataset in the source domain can be formulated as a labeled collection: $\mathbf{S} = \{< X_{S_i}^{(1)}, X_{S_i}^{(2)}, ..., X_{S_i}^{(M)}, Y_{S_i} >\}$. Here $X_{S_i}^{(m)} (1 \leq m \leq M)$ indicates the input of the m^{th} modality of the i^{th} sample in the dataset, and M indicates the number of modalities. Y_{S_i} indicates the label of this sample. On the other hand, the unlabeled target domain is defined as the following collection without labels: $\mathbf{T} = \{< X_{T_i}^{(1)}, X_{T_i}^{(2)}, ..., X_{Ti}^{(M)} >\}$, where $X_{T_i}^{(m)} (1 \leq m \leq M)$ indicates the input of the m^{th} modality of the i^{th} sample in the target domain. The Multi-modal Domain Adaptation (MDA) problem is to transfer the model from S to T, and apply the unlabeled data in T to incrementally optimize the model.

3.2 Pre-training on Source Domain

As shown in Fig. 2, the model is composed of a group of sub-models corresponding to multiple modalities, and performs a later fusion of the decision made by all sub-models. To pre-train the models in the source domain, a multi-task scheme is adopted as shown in Fig. 2. Besides the original **Supervised Classification Task** performed on the classifier $G^m (1 \leq m \leq M)$, a new prototype mapper $M^m (1 \leq m \leq M)$ is attached to the model for the **Prototype Learning Task**, which aims to learn the prototype of each class. The prototype of a class indicates the typical representation of the samples belonging to the class in the source domain. For any new sample, the similarity between the sample and a prototype of a class indicates how close between the sample and the ones of the class. Thus higher similarity indicates higher confidence to map the sample to the class, which is the basic of the *Reliability Measurement* for the transferred models proposed in the next section. The detail of above two learning tasks are given as follows.

Supervised Classification Task. Each sub-model of a modality is trained independently for the classification task based on the labeled multi-modal data in source domain, where the loss is set as:

$$L_C^m = \sum_i -Y_{S_i} \log \sigma(G^m(F^m(X_{S_i}^{(m)}))) \tag{1}$$

where $F^m (1 \leq m \leq M)$ is the feature extractor of the m^{th} modality. G^m is the corresponding classifier. σ is the softmax function. Y_{S_i} is the one-hot label vector. Cross entropy loss is adopted here for the classification task.

Prototype Learning Task. It aims to learn the multi-modal representation vector of each class, namely a prototype. Specifically, the prototype of the $\mathrm{k}^{th} (1 \leq k \leq C)$ class in the $\mathrm{m}^{th} (1 \leq m \leq M)$ modality is noted as W_k^m, which is initialized randomly and optimized by the Prototype Learning Task. The similarity between an input sample x and W_k^m is defined as follows:

$$d(x, W_k^m) = e^{-\frac{\left\| M^m(F^m(x)) - W_k^m \right\|_2^2}{\gamma}} \tag{2}$$

It takes a Gaussian kernel transformation on the Euclidean distance metric to normalize the similarity measurement into the range of $[0, 1]$, which facilitates the comparison between different modalities. Here M^m is the prototype mapper of the m^{th} modality, which project the original feature vector into the prototype space to measure the similarity with the prototype W_k^m. γ is a scaling factor.

Based on Eq. (2), the prototypes are learned by minimizing the following multi-label classification loss by measuring the similarity between each input sample with the prototype of each class:

$$\begin{aligned} L_P^m = \sum_i \sum_{k=1}^{C} [-Y_{S_{i,k}} \log d(X_{S_i}^{(m)}, W_k^m) \\ - (1 - Y_{S_{i,k}}) \log(1 - d(X_{S_i}^{(m)}, W_k^m))] \end{aligned} \tag{3}$$

Here $Y_{S_{i,k}}$ is the k^{th} binary element of the one-hot label vector Y_{S_i}, and indicates whether the sample belongs to the k^{th} class.

As shown in Fig. 2, both the Supervised Classification Task and Prototype Learning Task are learned simultaneously on the source domain for each modality. The total loss of each modality is:

$$L_S^m = L_C^m + \lambda L_P^m \tag{4}$$

where λ is the weight to balance these two tasks. By minimizing L_S^m, we can achieve the prototypes $\{W_k^m\}$, and optimize the parameters in M^m, F^m, and G^m.

3.3 Prototype based Reliability Measurement

When transferring the classifier G^m to the target domain, we can use G^m to predict the pseudo labels of unlabeled data. How to measure the reliability of the pseudo labels is a critical problem of domain adaptation. We propose the *Prototype based Reliability Measurement* in this section. Specifically, Given a sample $X_{T_i} = \{X_{T_i}^m | 1 \leq m \leq M\}$ in the target domain, $X_{T_i}^m$ indicates the input of the m^{th} modality. The one-hot pseudo label vector $\hat{Y}_{T_i}^m$ can be calculated by G^m, where the element corresponding to the $\hat{\mathrm{k}}^{th}$ class with the highest probability output by G^m is set as one. The reliability of the pseudo label is measured based on the similarity between the sample and the prototype of the $\hat{\mathrm{k}}^{th}$ class:

$$R_{T_i}^m = d(X_{T_i}^m, W_{\hat{k}}^m) \tag{5}$$

where the function $d(.)$ is defined in Eq. (2). The measurement is based on the similarity between the *what you see* in the target domain and *what you learned* in the source domain. It is more possible for the classifier to offer reliable prediction when the input (*what you see*) is close to the prototype (*what you learned*) of the predicted class.

In this way, $< X_{T_i}^m, \hat{Y}_{T_i}^m, R_{T_i}^m > (1 \leq m \leq M)$ can be achieved on each modality. By comparing the reliability of all modalities, the most reliable modality to predict this sample is:

$$\bar{m} = \arg\max_m R_{T_i}^m \tag{6}$$

Thus, the most reliable pseudo label for X_{T_i} is $\hat{Y}_{T_i} = \hat{Y}_{T_i}^{\bar{m}}$, and the corresponding reliability is: $R_{T_i} = R_{T_i}^{\bar{m}}$.

Moreover, to reduce the harmful effect caused by the overconfident pseudo-labels, we further perform the label smoothing processing on pseudo-labels as follows:

$$\dot{Y}_{T_i} = R_{T_i} * \hat{Y}_{T_i} + \frac{(1 - R_{T_i})}{C} \tag{7}$$

where C is the number of class. As shown in Fig. 2, after achieving the pseudo-labels of all samples in the target domain, the most

reliable ones are selected to form the teaching materials for the subsequent incremental learning:

$$U_R = \{(X_{T_i}, \dot{Y}_{T_i}, R_{T_i}) | R_{T_i} > R_\lambda\} \tag{8}$$

R_λ is a constant threshold to the filter out the pseudo-labels with lower reliability. By default, we set R_λ as the top 50% reliability of the samples in the dataset.

3.4 Asynchronous Learning

After achieving the teaching materials U_R, the supervised incremental optimization based on pseudo labels can be performed on the sub-model of each modality. Distinct from traditional synchronous optimization of all modalities, we propose an asynchronous learning strategy to consider the diversity of the abilities of different sub-models. As shown in Fig. 2, each sub-model performs the Curriculum Learning [19] to choose the samples for learning from easy to hard in accordance of their aptitude. Specifically, the $m^{th}(1 \leq m)$ sub-model is optimized by minimizing the following loss:

$$\min_{F_m, G_m, v_i^m} [L_C^m + \beta \sum_{(X_{T_i}, \dot{Y}_{T_i}, R_{T_i}) \in U_R} v_i^m (L_T^m(X_{T_i}^m) - \tau_m)] \tag{9}$$

which utilizes both of the labeled data in the source domain and the unlabeled data in the target domain. Here L_C^m is defined in Eq. (1), which indicates the loss of training the labeled data in the source domain. $v_i^m \in \{0, 1\}$ is a binary variable to determine whether the sample X_{T_i} in the target domain is chosen to learn. $L_T^m(X_{T_i}^m)$ is the cross-entropy loss of the sub-model to predict the label of $X_{T_i}^m$:

$$L_T^m(X_{T_i}^m) = -\dot{Y}_{T_i} \log(\sigma(G^m(F^m(X_{T_i}^m)))) \tag{10}$$

where $\dot{Y}_{T_i}$ is the pseudo label of X_{T_i} defined in Eq. (7). τ_m in Eq. (9) is the threshold to filter out hard samples. β is a constant to balance the training in the source domain and target domain.

Eq. (9) can be solved by alternating optimization based on the following steps:

Step A) Sample selection. Fix F_m, G_m, and minimize Eq. (9) to optimize v_i^m as:

$$v_i^m = \begin{cases} 1, & L_T^m(X_{T_i}^m) < \tau_m \\ 0, & otherwise \end{cases} \tag{11}$$

That means the easier samples with the loss $L_T^m(X_{T_i}^m)$ less than τ_m are chosen for learning. τ_m can be used to control the learning difficulty.

Step B) Model optimization. Fix v_i^m, and minimize Eq. (9) to optimize F^m and G^m . Typical gradient descent methods can be used to perform the optimization.

By iteratively going through Step A and B, the sub-model G^m can be optimized by choosing the labels according to the loss. Like the traditional *Curriculum Learning* [19], we can slowly increase τ_m to learn from easy to hard. For the m^{th} modality, we rank the loss $L_T^m(X_{T_i}^m)$ of the samples in U_R in ascending order, and then set τ_m to the loss of the top $p * |U_R|$, where p is the ratio parameter. In particular, at the beginning we initialize p to 5%, and increase it by 5% in each iteration until it reaches 50%.

In this way, every sub-model can make differentiated learning schedule by choosing teaching materials from easy to hard measured by itself, which is quite analog to the human learning principle of teaching students in accordance of their aptitude.

3.5 Reliability-aware Fusion

After the *Asynchronous Learning* stage, each sub-model G^m gets optimized in the target domain. As shown in Fig 2, the fusion of the prediction results from all sub-models is adopted as the final decision. As analyzed before, the reliability of each modality can be calculated by the prototype based measurement according to Eq. (5), based on which the *Reliability-aware Fusion* of multi-modal results can be defined as follows:

$$\ddot{Y}_{T_i} = \frac{\sum_{m=1}^{M} R_{T_i}^m \ddot{Y}_{T_i}^m}{\sum_{m=1}^{M} R_{T_i}^m} \tag{12}$$

Here $\ddot{Y}_{T_i}$ indicates the final prediction vector of any sample X_{T_i} in the target domain. $\ddot{Y}_{T_i}^m (1 \leq m \leq M)$ is the prediction vector output by the sub-model G^m, and $R_{T_i}^m$ is the reliability of the m^{th} modality calculated by Eq. (5) .

4 EXPERIMENTS AND RESULTS

In this section, we first introduce the datasets and experimental setup in Sec. 4.1 and Sec. 4.2, and then present the results in Sec. 4.3. The ablation study is given in Sec. 4.4, and the parameter sensitivity analysis is offered in Sec. 4.5. Finally, the qualitative results are presented in Sec. 4.6.

4.1 Datasets

We will evaluate the *Differentiated Learning* framework, namely ***DLMM***, on three popular tasks: **event recognition**, **fatigue detection** and **action recognition**. The datasets are detailed as follows:

Event Recognition Dataset (AVE) [37]. The AVE dataset contains two modalities of image and audio, and there are 4,143 videos covering 28 event categories, where videos are labeled with audio-visual event boundaries. We divide the dataset into two sub-sets indicates two different domains. Specifically, the Resnet-50 network [12] pre-trained on Imagenet [8] is used to extract 1024-dimensional features of the image of each sample. Then the feature vectors of each category are clustered into two clusters by the K-Means algorithm [15]. In the end, we obtained 43,458 source domain samples and 113,829 target domain samples.

Fatigue Detection Dataset (CogBeacon) [28]. CogBeacon contains two usable modalities of EEG signal and facial keypoints. It consists of 76 sessions collected from 19 users performing three versions of cognitive tasks (namely V1,V2,V3), inspired by the principles of the Wisconsin Card Sorting Test. The number of samples corresponding to the cognitive task V1,V2,V3 are 2,259, 2,221 and 2,389 respectively.

Action Recognition Dataset (EPIC Kitchens) [6]. In this dataset, we adopt the same domain division configuration as the previous work [25], which contains three domains D1, D2, and D3 in EPIC Kitchens. Eight types of actions are analyzed, and each sample is represented as two modal forms of RGB image and Optical Flow.

The number of action segments in the three domains D1, D2, and D3 are 1978, 3245 and 4871 respectively.

More details about these datasets are presented in our attachment.

4.2 Configuration of Models

In the event recognition task, the Resnet-18 network [12] is used as the feature extraction F^m for both image and audio modal, and input samples are converted into 512-dimensional features through F^m. The classifier G^m is a single-layer fully connected layer. The prototype mapper M^m is implemented by a single-layer fully connected layer, which maps 512-dimensional features to a 128-dimensional prototype space.

In the fatigue detection task, the feature extraction F^m is implemented by three 1D convolutional layers and a single-layer fully connected layer for both EEG signals and facial keypoints. The 64-dimensional features are extracted after F^m. The classifier G^m is a single-layer fully connected layer. The prototype mapper M^m is implemented by one-layer fully connected layer, which maps 64-dimensional features to the 32-dimensional prototype space.

In the action recognition task, similar to previous works [25], the inflated 3D convolutional architecture (I3D) [4] is used as the feature extraction F^m for both modalities, which has the 1024 dimensional output vector. The classifier G^m is a single fully connected layer to predict class labels. The prototype mapper M^m is implemented by a single fully connected layer, which maps 1024-dimensional features to 256-dimensional prototype space.

The Adam optimization method [18] with learning rate e-4 is adopted. And the accuracy of a model is measured by the ratio of correctly classified samples in the target domain.

4.3 Comparison Results

Baseline Models. To verify the effectiveness of our proposed framework, the recently developed multi-modal domain adaptation methods are compared in the experiments. Specifically, Co-training approaches [3, 43] treat each modality equally, and assign pseudo labels to the samples with the highest posterior probability and use them to train the student models. *MDANN* [29] uses three levels of multi-modal fusion features to conduct adversarial learning between the source and target domains. *MM-SADA* [25] is state-of-the-art MDA method, which uses the consistency between modalities as self-supervised constraints and conducts adversarial learning between the source and target domains to reduce domain shift.

In addition, we also compare with several popular single-modal domain adaptation methods [10, 20, 22, 32, 45], and applying these methods on all modalities simultaneously. In particular, *MCD* [32] eliminates domain shift through classifier disagreement, and we adopt multi-modal classification heads here to combine the divergence between multiple modalities following the setting in [25]. *DANN* [10] confuses the recognition of the source and target domains to achieve feature alignment between domains. *CBST* [45] performs self-training based on class balance on each modal to reduce domain shift. *AdaBN* [20] updates batch Normalisation layers with target domain statistics. *MMD* [22] uses kernel transformation to align the features of the source and target domains.

Table 2: Performance comparison on the Event Recognition Dataset (AVE).

Method	Image	Audio	Fusion
Direct-Transfer	17.46	42.83	43.16
DANN[10]	20.83	44.21	45.78
CBST[45]	21.16	47.61	48.63
MCD[32]	27.15	39.29	40.35
CT[3]	31.52	36.67	37.18
eCT[43]	33.78	36.92	38.71
MDANN[29]	-	-	43.65
MM-SADA[25]	29.67	48.65	50.13
DLMM-prob	34.17	45.69	46.81
DLMM-entropy	36.88	47.57	48.76
DLMM-margin	36.41	46.82	48.05
DLMM-Seperate	22.83	46.33	47.59
DLMM	**42.58**	**52.37**	**55.02**
Supervised	67.82	79.23	83.15

Variation Models. In order to verify the effectiveness of the *Prototype based Reliability Measurement*, which plays a key role in judging the reliability of different modalities, we also compare *DLMM* with some variation models with different measurement. e.g. *DLMM-prob*, *DLMM-entropy* and *DLMM-margin* respectively represent the variation models by replacing the *Prototype based Reliability Measurement* with the posterior probability [3, 43], entropy of the prediction results [14] and margin sampling [33].

In addition, in order to explore the importance of multi-modal collaboration on generating shared teaching materials according to the reliability scores of all sub-models, we also compare *DLMM* with the variation model *DLMM-seperate*, where each sub-model generates the teaching materials individually according to its own reliability score.

Moreover, to test the performance of the backbone networks, we test the simplest variation model *Direct-Transfer* as the lower bound, which stands for directly migrating the model trained in the source domain to the target domain without any incremental optimization. Meanwhile, as the upper bound, we also report the results of supervised learning (denoted as *Supervised*) based on the ground-truth labels. In all above models, the reliability-aware fusion of the multi-modal classifiers are adopted as the final decision.

Experimental Results. The results of three tasks are shown in Table 2, 3, and 4 respectively. As shown in Table 2, in event recognition task, the accuracy of the image modal is much lower than that of the audio modal when directly transferring from the source domain to the target domain (*Direct-Transfer*). It's because the domain-shift of the image modal is much more serious than audio. In this case, treating each modality equally may even bring side-effect after domain adaptation, as shown in the results of baseline models in Table 2. In particular, in some models (e.g. *MCD* [32], CT [3], eCT [43]), the accuracy of the audio modal is even lower than before after the incremental learning on the target domain. Benefit from the reliability measurement of sub-models and the asynchronous learning, our method achieves much better performance than all baselines. It is worth mentioning that our method not only has the highest joint decision accuracy, but also improves

Table 3: Performance comparison on the Fatigue Detection Dataset (CogBeacon). $V1$,$V2$,and $V3$ indicate three domains.

Method	V1→V2			V2→V3			V3→V1			Mean		
	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion
Direct-Transfer	56.27	59.03	59.31	61.46	63.15	64.09	62.41	65.41	66.41	60.05	62.53	63.27
DANN[10]	57.81	60.12	61.52	63.02	64.87	66.16	64.95	67.77	69.81	61.93	63.31	65.83
CBST[45]	59.61	62.19	63.56	64.05	64.98	66.35	65.82	70.18	71.16	63.16	65.78	67.46
MCD[32]	58.04	61.75	62.83	65.34	66.71	67.65	65.67	69.13	70.25	63.02	65.83	66.91
CT[3]	57.98	60.42	61.93	62.97	64.47	65.13	65.33	67.01	69.34	62.09	63.97	65.47
eCT[43]	58.45	61.23	62.29	62.99	64.50	65.56	65.72	69.51	70.57	62.38	65.08	66.14
MDANN[29]	-	-	62.45	-	-	66.72	-	-	69.92	-	-	66.36
MM-SADA[25]	58.74	62.83	63.47	65.14	66.57	67.11	67.92	69.72	70.69	63.93	66.37	67.09
DLMM-prob	58.02	61.53	62.53	65.09	66.52	67.14	67.51	69.59	70.28	63.54	65.88	66.65
DLMM-entropy	58.17	61.86	63.07	65.26	67.07	68.32	68.47	71.26	72.47	63.97	66.73	67.95
DLMM-margin	58.06	61.62	62.86	65.37	69.17	68.84	68.36	70.89	72.03	63.93	67.23	67.91
DLMM-Separate	59.77	62.93	64.15	67.28	67.61	69.15	69.82	71.85	72.82	65.62	67.46	68.71
DLMM	**59.91**	**64.82**	**66.21**	**67.36**	**70.02**	**71.80**	**70.15**	**73.67**	**74.89**	**65.91**	**69.50**	**70.97**
Supervised	79.34	85.16	87.34	82.10	85.67	86.59	83.46	87.82	89.12	81.63	86.15	87.68

Table 4: Performance comparison on the Action Recognition Dataset (EPIC Kitchens). $D1$,$D2$,and $D3$ indicate three different domain.

Method	D1→D2			D2→D3			D3→D1			Mean		
	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion
Direct-Transfer	36.1	45.6	43.7	33.6	46.0	46.5	36.3	44.2	44.5	35.3	45.3	44.9
AdaBN[20]	44.1	46.5	47.0	44.8	48.3	48.7	41.5	45.0	47.8	43.5	46.6	47.8
MMD[22]	43.7	46.3	46.5	44.5	48.2	48.5	41.7	45.4	48.3	43.3	46.6	47.7
MCD[32]	43.5	46.3	46.4	45.8	50.7	51.0	42.0	45.0	47.9	43.8	47.3	48.4
CT[3]	43.7	46.0	46.1	45.3	50.2	50.7	41.6	44.8	47.3	43.5	47.0	48.0
eCT[43]	43.9	46.2	46.3	45.5	50.2	50.8	41.8	44.9	47.6	43.7	47.1	48.2
MDANN[29]	-	-	45.7	-	-	48.6	-	-	48.2	-	-	47.5
MM-SADA[25]	45.0	49.0	49.5	46.2	52.1	52.7	42.1	45.7	50.9	44.5	48.9	51.1
DLMM-prob	45.8	48.7	49.2	45.3	50.4	50.8	42.0	45.1	48.7	44.4	48.1	49.6
DLMM-entropy	46.0	48.8	49.6	45.7	51.2	51.6	42.2	45.3	49.3	44.6	48.4	50.2
DLMM-margin	45.6	48.7	49.1	45.3	50.7	51.2	42.1	46.1	49.5	44.3	48.5	49.9
DLMM-Separate	46.2	49.8	50.1	47.1	50.6	51.5	42.2	48.8	51.3	45.2	49.7	51.1
DLMM	**48.3**	**52.0**	**52.7**	**49.7**	**54.6**	**55.8**	**46.9**	**51.3**	**53.5**	**48.3**	**52.6**	**54.0**
Supervised	63.1	67.6	71.7	62.2	68.2	74.0	60.9	62.4	62.8	62.1	66.1	69.5

the accuracy of the image modal far beyond all baselines, showing the advantage of our framework to optimize the weaker sub-models.

In the fatigue detection and action recognition tasks, we respectively show the accuracy of transferring models among three given domains in Table 3 and Table 4. In particular, the performance of *Direct-Transfer* is not too bad, because of weaker domain-shift than the previous event recognition task in Table 1. In these cases, *DLMM* can still achieve much higher accuracy on each modality than other models. Due to the page limit, we only show the results of three configurations of source-target domain pairs. *DLMM* can achieve much higher accuracy on each modality than other models.

In addition, the accuracy of the variation models *DLMM-prob*, *DLMM-entropy* and *DLMM-margin* are significantly higher than *Direct-Transfer* in all above tasks, but much lower than our method *DLMM*. This shows the *Prototype-based Reliability Measurement* adopted in *DLMM* is a much more effective way to estimate the reliability of pseudo labels than the traditional methods such as the posterior probability [3, 43], entropy [14] and margin based metrics [33]. Besides, the accuracy of *DLMM-seperate* is significantly higher than *Direct-Transfer* in all tasks, but much lower than *DLMM*. This confirms the importance of multi-modal collaboration on generating shared teaching materials in *DLMM* to make full use of complementarity of different modalities.

4.4 Ablation Study

The most important components of *DLMM* include the *Prototype based Reliability Measurement* (Sec. 3.3), the *Asynchronous Learning* (Sec. 3.4), the *Reliability-aware Fusion* (Sec. 3.5), together with the *Label-smoothing* technique adopted in Eq. (7). To analyze the contribution of these components clearly, we conducted ablation studies on three datasets and show the results in Table 5. Specifically, *DLMM(RM)* indicates the basic model applying the *Prototype based Reliability Measurement* to pick out samples for incremental learning. *DLMM(RM+AL)* and *DLMM(RM+LS)* indicate the models adding

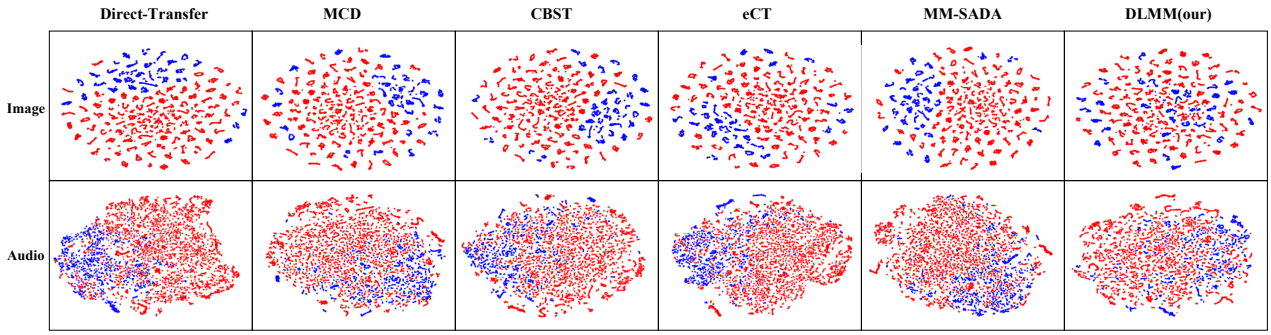

Figure 3: t-SNE plots of Image and Audio feature distribution in the AVE dataset produced by the baseline models and our proposed method *DLMM*. The source domain in blue and the target domain is shown in red.

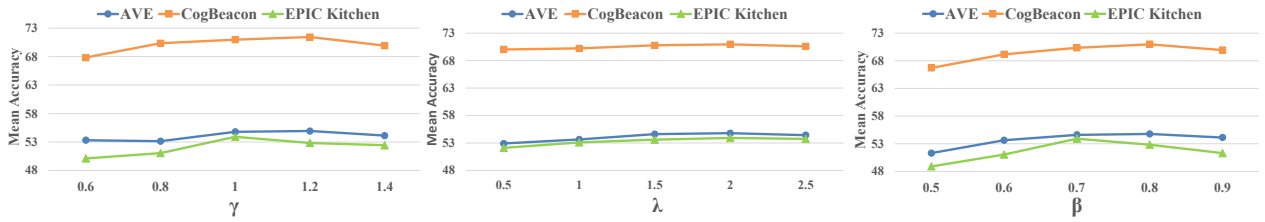

Figure 4: The accuracy of the *DLMM* model with different configurations of the hyper-parameters γ, λ, and β.

the *Asynchronous Learning* and *Label-smoothing* respectively. *RF* indicates that the *Reliability-aware Fusion* is used, while the multi-modal average fusion is used if without *RF*. *DLMM(RM+AL+LS+RF)* is the full model equipped with all components. As shown in Table 5, the accuracy of *DLMM(RM)* is significantly higher than *Direct-Transfer*, indicating that the reliability measurement can effectively pick out reliable pseudo-labels for incremental learning. *DLMM(RM+AL)* and *DLMM(RM+LS)* improve the performance further through asynchronous learning and label smoothing. And the accuracy will be higher when using Reliability-aware Fusion. The full model *DLMM(RM+AL+LS+RF)* achieves the highest accuracy by combining all components.

4.5 Parameter Sensitivity Analysis

The sensitivity of the hyper-parameter γ, λ and β applied in *DLMM* is tested and the results are shown in Fig. 4. In particular, γ means the scale parameter in the distance measurement in the prototype space. Too large or too small γ will reduce the accuracy, so we set γ to 1 in all three tasks according to Fig. 4. It can be also observed that the change of λ, which balances the prototype learning task and the classification task. And β balances the source domain loss and target domain loss.

4.6 Qualitative Results

In order to demonstrate the effect of *DLMM* in reducing domain-shift intuitively, we show the t-SNE visualisation of the feature vectors learned in the AVE dataset in Fig. 3. Comparing with other methods, *DLMM* is more effective to mix the feature vectors from the source domain and target domain. This proves that the *Differentiated Learning* is effective to reduce the domain-shift.

Table 5: Ablation study of *DLMM* with different configurations of key components.

Method	AVE	CogBeacon	EPIC Kitchen
Direct-Transfer	43.16	63.27	45.50
DLMM(RM)	49.42	66.79	46.52
DLMM(RM+RF)	49.48	67.19	46.83
DLMM(RM+AL)	51.83	68.48	47.35
DLMM(RM+AL+RF)	52.12	68.70	47.54
DLMM(RM+LS)	52.91	68.89	48.96
DLMM(RM+LS+RF)	53.25	69.18	49.24
DLMM(RM+AL+LS)	54.78	70.16	52.33
DLMM(RM+AL+LS+RF)	**55.02**	**70.32**	**52.68**

5 CONCLUSION

In this paper, we propose a novel *Differentiated Learning* framework, namely *DLMM*, for multi-modal domain adaptation. Comprehensive experiments on three multi-modal learning tasks show that *DLMM* can achieve much better performance than state-of-the-art MDA methods. In future work, we will extend the proposed method on the datasets with more than two modalities and combining *DLMM* with the adversarial learning to further improve the performance.

6 ACKNOWLEDGMENTS

The work described in this paper was supported by the grants from NSFC (No.61876065), Natural Science Foundation of Guangdong Province, China (No.2018A0303130022), Science and Technology Program of Guangzhou, China (No.201904010200), Water Conservancy Science and Technology Innovation Project of Guangdong Province, China (2020-03), and Guangzhou Science and Technology Program key projects (202007040002).

REFERENCES

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[2] Gaurav Bhatnagar, QM Jonathan Wu, and Zheng Liu. 2013. Directive contrast based multimodal medical image fusion in NSCT domain. *IEEE transactions on multimedia* 15, 5 (2013), 1014–1024.

[3] Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*. 92–100.

[4] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[5] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, Yong Rui, et al. 2016. Semi-Supervised Multimodal Deep Learning for RGB-D Object Recognition.. In *IJCAI*. 3345–3351.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 720–736.

[7] Sanjoy Dasgupta. 2006. Coarse sample complexity bounds for active learning. In *Advances in neural information processing systems*. 235–242.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[9] Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*. PMLR, 1180–1189.

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

[11] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. 2019. Dlow: Domain flow for adaptation and generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2477–2486.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*. PMLR, 1989–1998.

[14] Alex Holub, Pietro Perona, and Michael C Burl. 2008. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1–8.

[15] Anil K Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters* 31, 8 (2010), 651–666.

[16] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. 2015. Self-paced curriculum learning.. In *AAAI*, Vol. 2. 6.

[17] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977* (2017).

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

[19] M Kumar, Benjamin Packer, and Daphne Koller. 2010. Self-paced learning for latent variable models. *Advances in neural information processing systems* 23 (2010), 1189–1197.

[20] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. 2018. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition* 80 (2018), 109–117.

[21] Jinlu Liu, Liang Song, and Yongqiang Qin. 2020. Prototype rectification for few-shot learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 741–756.

[22] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.

[23] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*. PMLR, 2208–2217.

[24] Fan Ma, Deyu Meng, Qi Xie, Zina Li, and Xuanyi Dong. 2017. Self-paced co-training. In *International Conference on Machine Learning*. PMLR, 2275–2284.

[25] Jonathan Munro and Dima Damen. 2020. Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 122–132.

[26] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. In *Advances in neural information processing systems*. 1196–1204.

[27] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*.

[28] Michalis Papakostas, Akilesh Rajavenkatanarayanan, and Fillia Makedon. 2019. CogBeacon: A Multi-Modal Dataset and Data-Collection Platform for Modeling Cognitive Fatigue. *Technologies* 7, 2 (2019), 46.

[29] Fan Qi, Xiaoshan Yang, and Changsheng Xu. 2018. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM international conference on Multimedia*. 429–437.

[30] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. 2014. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596* (2014).

[31] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329* (2021).

[32] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3723–3732.

[33] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. 2001. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*. Springer, 309–318.

[34] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*. 4077–4087.

[35] Nitish Srivastava and Russ R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*. 2222–2230.

[36] Baochen Sun and Kate Saenko. 2016. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *Computer Vision – ECCV 2016 Workshops*, Gang Hua and Hervé Jégou (Eds.). Springer International Publishing, Cham, 443–450.

[37] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 247–263.

[38] Isaac Triguero, Salvador García, and Francisco Herrera. 2015. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information Systems* 42, 2 (2015), 245–284.

[39] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7167–7176.

[40] Yingda Xia, Dong Yang, Zhiding Yu, Fengze Liu, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. 2020. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical Image Analysis* 65 (2020), 101766.

[41] Zhiding Yu, Weiyang Liu, Yang Zou, Chen Feng, Srikumar Ramalingam, BVK Vijaya Kumar, and Jan Kautz. 2018. Simultaneous edge alignment and learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 388–404.

[42] Xiao Zhang, Rui Zhao, Yu Qiao, and Hongsheng Li. 2020. RBF-Softmax: Learning Deep Representative Prototypes with Radial Basis Function Softmax. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*. Springer, 296–311.

[43] Zixing Zhang, Fabien Ringeval, Bin Dong, Eduardo Coutinho, Erik Marchi, and Björn Schüller. 2016. Enhanced semi-supervised learning for multimodal emotion recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5185–5189.

[44] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. 2019. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5982–5991.

[45] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*. 289–305.