MODfinity: Unsupervised Domain Adaptation with Multimodal Information Flow Intertwining

Shanglin Liu[1] Jianming Lv[1,*] Jingdan Kang[1] Huaidong Zhang[1]
[1]South China University of Technology
slliuxjy@gmail.com, jmlv@scut.edu.cn, jingdankang6@gmail.com, huaidongz@scut.edu.cn

Zequan Liang[2]
[2]University of California, Davis
202121045193@mail.scut.edu.cn

Shengfeng He[3]
[3]Singapore Management University
shengfenghe@smu.edu.sg

Abstract

Multimodal unsupervised domain adaptation leverages unlabeled data in the target domain to enhance multimodal systems continuously. While current state-of-the-art methods encourage interaction between sub-models of different modalities through pseudo-labeling and feature-level exchange, varying sample quality across modalities can lead to the propagation of inaccurate information, resulting in error accumulation. To address this, we propose Modal-Affinity Multimodal Domain Adaptation (MODfinity), a method that dynamically manages multimodal information flow through fine-grained control over teacher model selection, guiding information intertwining at both feature and label levels. By treating labels as an independent modality, MODfinity enables balanced performance assessment across modalities, employing a novel modal-affinity measurement to evaluate information quality. Additionally, we introduce a modal-affinity distillation technique to control sample-level information exchange, ensuring reliable multimodal interaction based on affinity evaluations within the feature space. Extensive experiments on three multimodal datasets demonstrate that our framework consistently outperforms state-of-the-art methods, particularly in high-noise environments.

1. Introduction

Recent advancements in multimodal information fusion have greatly enhanced model understanding and reasoning across various domains by integrating data from multiple modalities. This progress is particularly evident in applications like multimodal image classification [10, 19, 33, 35], multimodal image segmentation [8, 30, 34, 38], and multimodal object detection [5, 29].

*Corresponding author.

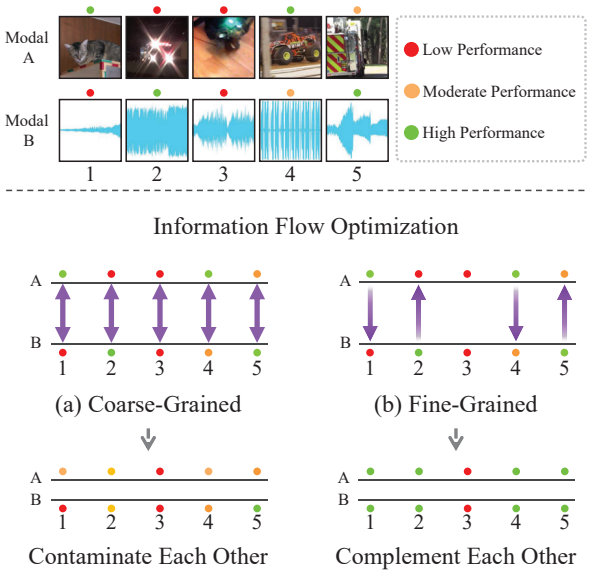


Figure 1. Comparison of Information Flow Optimization Approaches. (a) Coarse-grained optimization results in inaccurate information exchange, leading to cross-modal contamination. (b) Fine-grained optimization allows complementary interactions between sub-modalities, promoting mutual enhancement. Further details are available in the supplementary materials.

However, deploying multimodal models in new environments often results in substantial performance degradation due to distributional shifts between the source and target domains [2]. To mitigate this, recent research has introduced Multimodal Unsupervised Domain Adaptation (MUDA) methods, which leverage unlabeled data in the target domain to optimize models continuously. MUDA methods generally fall into two categories: Distribution Alignment and Co-learning.

Distribution Alignment techniques aim to bridge cross-domain distributional gaps [9, 20] but often overlook the

diverse generalization strengths across modalities [18, 23]. In contrast, Co-learning methods optimize models through structured inter-modal information exchange, establishing state-of-the-art performance [15, 28].

The core challenge for co-learning methods lies in effectively managing the learning process without ground-truth labels in the target domain. Traditional approaches generally treat all modalities equally, allowing sub-models to exchange information synchronously, with variable quality, and without specific constraints [3, 15, 22, 25, 26]. However, these straightforward co-learning schemes often fail to address the complexities of real-world data, where quality can vary significantly across modalities. As a result, low-quality data exchange can lead to error accumulation. To address this, some coarse-grained methods attempt to control inter-modal exchange by adjusting the degree of learning with hyperparameters or using curriculum learning to differentiate modality treatment in information flow [21, 28]. While these methods offer some control over the learning intensity of each modality, they do not fully address the critical issue that target domain quality often varies at the sample level, not just across modalities. Thus, they fall short of optimizing information flow for samples of differing quality, leading to suboptimal outcomes.

To overcome this limitation, we propose MODfinity, a fine-grained optimization method for multimodal information flow, as illustrated in Fig. 1 (b). Our approach introduces a modal-affinity measurement to evaluate pseudo-label quality and modality performance impartially by treating labels as an independent, continuously updated modality. This allows for balanced comparisons across modalities, with multimodal features evaluated against the feature vectors of the label modality, which serves as a neutral benchmark. Based on this measurement, we employ a modal-affinity distillation technique to guide information intertwining between modalities at a sample-specific level.

Extensive experiments demonstrate the efficacy of our method in assessing modality and sample performance accurately, effectively managing multimodal information flow. Our approach shows substantial performance gains across all modalities, especially on noisy datasets. On the AVE [32], Epic-Kitchens 55 [7], and CogBeacon [24] datasets, our framework consistently outperforms state-of-the-art methods. Notably, on the noisy AVE dataset, where other methods struggle or even degrade performance (e.g., contrastive learning [15] results in a 16.67% performance drop), our method provides reliable guidance, achieving a 5.79% performance improvement.

The main contributions of this paper are threefold:

- We introduce a multimodal domain adaptation framework that provides fine-grained control over information flow between sub-models at both feature and label levels.
- We propose a dynamic modal-affinity measurement that continuously evaluates sub-model performance and sample quality by treating labels as an independent modality, enabling balanced comparisons across modalities.
- We develop a modal-affinity distillation method, allowing the teacher modality to guide the student modality effectively in high-quality samples while minimizing its influence in low-quality samples.

2. Related Work

2.1. Multimodal Domain Adaptation

Domain adaptation strategies are essential for deploying learned models in new, previously unseen environments, particularly when labeled data in the target domain is scarce. This challenge is even more complex in multimodal settings, where each modality may require distinct adaptation techniques or adapt at different rates.

Traditional unsupervised domain adaptation (UDA) methods include minimizing domain discrepancies through distance-based approaches [6, 36, 37], adversarial strategies [9, 34], and adaptive normalization [16, 17]. However, these techniques are predominantly designed for unimodal scenarios. In Multimodal Unsupervised Domain Adaptation (MUDA), co-training methods leverage the complementary perspectives of multiple modalities to enhance model performance through inter-modal synergies [3, 13]. For instance, Munro et al. [22] employed adversarial techniques to align and synchronize features across modalities during adaptation. Kim et al. [15] tackled domain shift by aligning domain-specific feature distributions, while Planamente et al. [25, 26] improved cross-domain generalization by normalizing feature norms across modalities. Extending these approaches, Tang et al. [31] addressed target-domain adaptation without access to source-domain data, leveraging multimodal base models with memory-aware predictors and class-attention calibration mechanisms. Additionally, Ke et al. [14] explored missing modality completion in multimedia settings.

2.2. Differentiated Learning

Differentiated learning aims to address the unique characteristics and requirements of each modality within a multimodal system. This approach is crucial for optimizing performance across modalities that vary widely in nature and the type of data they process.

Lv et al. [21, 27] proposed methods for modality-specific treatment, significantly enhancing adaptation and learning efficiency across diverse modalities. Subsequent research has shown the effectiveness of applying tailored processing techniques to individual modalities [1]. Additionally, adaptive architectures and learning strategies have been explored to capitalize on each modality's strengths while compensating for its weaknesses [11, 18]. Meta-learning have been

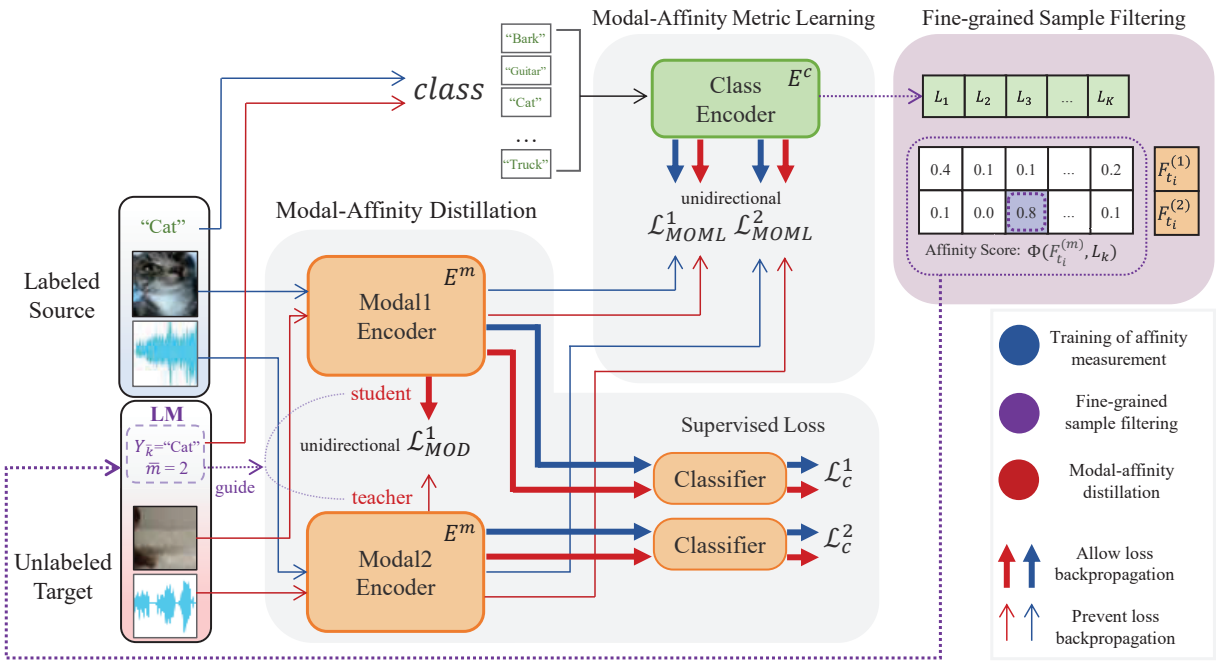


Figure 2. Our Modal-Affinity Multimodal Domain Adaptation framework consists of three main stages. **(S1) Training of Affinity Measurement:** The class encoder E^c is trained using Modal-Affinity Metric Learning (MOML). **(S2) Fine-Grained Sample Filtering:** Pseudo-labels $Y_{\bar{k}}$ are generated for each target domain sample t_i, the corresponding teacher modality $\bar{m}$ is identified, and this information is packaged as Learning Material (LM). **(S3) Modal-Affinity Distillation:** For each $LM_i \in LM$, the designated teacher modality $\bar{m}$ transfers information to the student modality via Modal-Affinity Distillation (MOD).

used to dynamically adjust learning parameters, allowing more capable modalities to guide less capable ones [28].

3. MODfinity

3.1. Problem Definition

Multimodal Unsupervised Domain Adaptation (MUDA) aims to transfer a multimodal system, which is trained on a labeled source domain, to an unlabeled target domain and leverage the unlabeled data to further optimize the model to increase its adaptability towards the new environment.

In particular, the source domain is represented as S , which is formulated as $S = \{s_1, s_2, \ldots, s_n\}$. Here $s_i = \langle X_{s_i}^{(1)}, X_{s_i}^{(2)}, \ldots, X_{s_i}^{(M)}, Y_{s_i} \rangle$ indicates a labeled sample with M modalities, where $X_{s_i}^{(m)}$ $(1 \leq m \leq M)$ is the input data of the m^{th} modality, and Y_{s_i} is the label of the sample. Similarly, the target domain dataset is defined as: $T = \{t_1, t_2, \ldots, t_{n'}\}$ containing n' unlabeled samples, where $t_i = \langle X_{t_i}^{(1)}, X_{t_i}^{(2)}, \ldots, X_{t_i}^{(M)} \rangle$. The objective is to leverage the unlabeled data $\{t_i\}$ to optimize the multimodal model on the target domain.

3.2. Model Overview

We propose the Modal-Affinity Multimodal Domain Adaptation framework, namely **MODfinity**, to achieve efficient filtering of pseudo labels and fine-grained control for the flow of information. As shown in Fig. 2, the pipeline of MODfinity contains three main stages: the Training of Affinity Measurement, the Fine-grained Sample Filtering, and the Modal-Affinity Distillation.

In particular, in the **Training of Affinity Measurement** stage, the labels are modeled as independent modality, the feature vectors of which are optimized through affinity learning in the source domain. In this way, the label modality can act as an independent measurement to judge the affinity of sub-models from different modalities for each sample. In the subsequent **Fine-grained Sample Filtering** stage, modal-affinity measurement is applied in the target domain to assess the performance of modalities and perform the filtering of samples, which are prepared for the following fine-grained co-learning. Finally, in the **Modal-Affinity Distillation** stage, for each individual sample, the better-performing modalities act as teachers to provide both label and feature-level guidance to other student modalities, which promotes the performance of the whole system through the differentiated training for different modalities. Furthermore, the Affinity Measurement module is continuously optimized through the Modal-Affinity Metric Learning loss, so as to remain adaptable to the change of the feature distribution in the target domain.

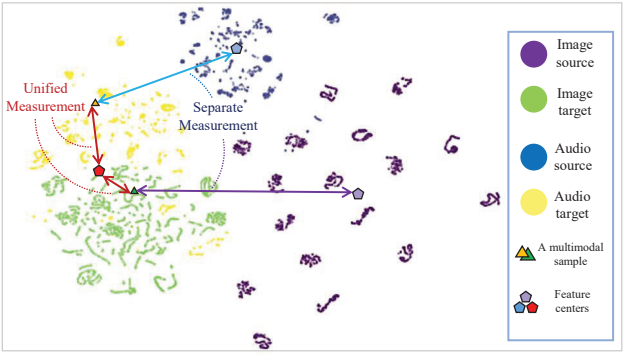


Figure 3. t-SNE plots of different measurement methods for the 'Church bell' category in the AVE dataset.

3.3. Affinity Measurement

To evaluate the pseudo labels of unlabeled samples, traditional methods typically use confidence scores based on the logits [3, 22, 28] or distances to cluster centers [21], which are measured in each modality independently.

However, due to the diverse feature distribution and different degrees of domain shift, the separate measurements from different modalities become incomparable as shown in Fig. 3. This inspires us to propose a unified measurement. The simplest way to achieve this goal is to project all of the feature vectors from different modalities into one unified feature space for the convenience of comparison, but this solution may lead to serious performance degradation in the unbalanced multimodal systems, where the worse modality may ruin the performance of others during the end-to-end supervised learning as shown in the following experiments. How to achieve a unified measurement of different modalities without projecting them into the same feature space is a challenging problem.

Motivated by the above analysis, we propose a more flexible affinity measurement to achieve a good trade-off between union and separation of different modalities. Specifically, we model the labels as an independent modality to construct the cross-modal measurement which is optimized by affinity learning between different modalities.

Fig. 4 shows the structure of the Affinity Measurement module, which is firstly trained in the labeled source domain. For each sample $s_i = \langle X_{s_i}^{(1)}, X_{s_i}^{(2)}, \ldots, X_{s_i}^{(M)}, Y_{s_i} \rangle$, the feature encoder E^m of the m^{th} modality transforms the input $X_{s_i}^{(m)}$ into the feature vector $F_{s_i}^{(m)}$, which are fed to the following classifiers to conduct the supervised learning based on the traditional cross-entropy loss $\mathcal{L}_c^m$ as shown in Fig. 2. Meanwhile, the labels are modeled as an independent modality in Fig. 4, where a class encoder E^c is applied to transform the one-hot label vector Y_{s_i} into the feature L_{s_i}. The affinity between the label Y_{s_i} and the input $X_{s_i}^{(m)}$

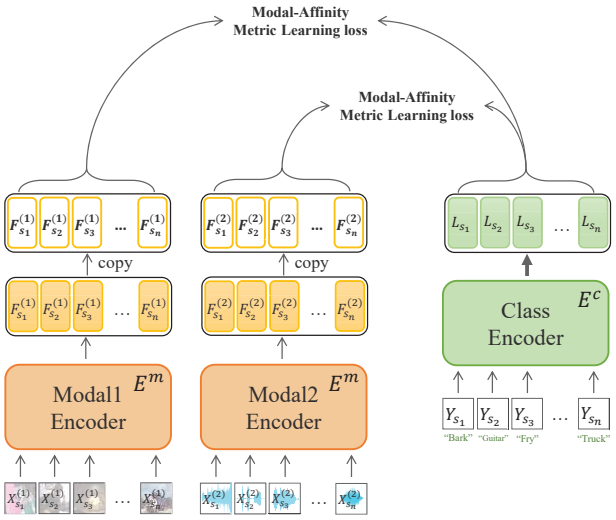


Figure 4. Architecture of the Modal-Affinity Metric Learning (MOML) process. MOML specifically enhances E^c by leveraging affinity loss between feature representations from modal encoders, Modal1 and Modal2, without directly affecting the operation of the individual modal encoders E^m.

is measured by the similarity of their feature vectors:

$$\phi(F_{s_i}^{(m)}, L_{s_i}) = exp(F_{s_i}^{(m)} \cdot L_{s_i}/\tau), \tag{1}$$

where τ is the temperature parameter used to control the gradient flow during back-propagation.

As shown in Fig. 4, we introduce the following Modal-Affinity Metric Learning (MOML) loss to train E^c:

$$\mathcal{L}_{MOML}^m = - \log \frac{\sum_{s_i \in S} \sum_{s_j \in s_i^+} \phi(\mathbf{F}_{s_i}^{(m)}, L_{s_j})}{\sum_{s_i \in S} \left(\sum_{s_j \in s_i^+} \phi(\mathbf{F}_{s_i}^{(m)}, L_{s_j}) + \sum_{s_k \in s_i^-} \phi(\mathbf{F}_{s_i}^{(m)}, L_{s_k}) \right)}. \tag{2}$$

Here s_i^+ indicates the positive samples in the batch sharing the same label Y with s_i while s_i^- for the negative ones with different labels. By applying affinity learning on all modalities, the encoder E^c will learn the multi-modal characteristics of each category. Furthermore, a copy of each $F_{s_i}^{(m)}$ without gradient propagation is adopted for affinity learning as shown in Fig. 4, ensuring the independent optimization of each encoder E^m so as to avoid the worse modality ruin the better one.

3.4. Fine-grained Sample Filtering

When transferring the model to the unlabeled target domain, the class encoder E^c and the feature encoder of each modality E^m are applied to generate the pseudo labels as

shown in stage S2 of Fig. 2. Given an unlabeled sample, $t_i = \langle X_{t_i}^{(1)}, X_{t_i}^{(2)}, \ldots, X_{t_i}^{(M)} \rangle$, the feature encoders E^m in the m^{th} modality is applied to extract the feature vector $F_{t_i}^{(m)}$. Meanwhile, for each label Y_k $(1 \leq k \leq C)$ where C indicates the number of the different labels, the corresponding one-hot label vector is input into the class encoder E^c to achieve the feature representations L_k. The similarity between the feature vector $F_{t_i}^{(m)}$ and the label Y_k is calculated as $\phi(F_{t_i}^{(m)}, L_k)$ according to Eq. (1).

Based on the affinity measurement, we can achieve the most confident pseudo label of the sample t_i:

$$(\bar{m}_i, \bar{k}_i) = \arg\max_{m,k} \phi(F_{t_i}^{(m)}, L_k), \tag{3}$$

where $\bar{k}_i$ indicates the index of the label, and $Y_{\bar{k}_i}$ is the best pseudo label predicted by the modality $\bar{m}_i$. The confidence of the pseudo label is:

$$R_{t_i} = \max_{m,k} \phi(F_{t_i}^{(m)}, L_k). \tag{4}$$

After collecting the pseudo labels of all samples, we rank the samples by the confidence R_{t_i} and select the ones with the top $\alpha\%$ confidence to form the labeled learning material LM:

$$LM = \{\langle t_i, R_{t_i}, Y_{\bar{k}_i}, \bar{m}_i \rangle\}. \tag{5}$$

LM provides a fine-grained teaching information about the learning object t_i, the teacher modality $\bar{m}_i$, the pseudo label $Y_{\bar{k}_i}$ from the teacher, and the confidence of the teacher R_{t_i}.

3.5. Modal-Affinity Distillation

Based on LM, we can organize multi-modal co-learning by deciding the teacher-student relationship for each sample, which allows for the control of the flow of information. In particular, for the m^{th} modality, the pseudo labels generated by other modalities form their own learning material:

$$LM^{(m)} = \{\langle t_i, \bar{m}_i \rangle | \langle t_i, R_{t_i}, Y_{\bar{k}_i}, \bar{m}_i \rangle \in LM, \bar{m}_i \neq m\}. \tag{6}$$

As shown in Fig. 5, for each sample t_i in $LM^{(m)}$ with the pseudo label $Y_{\bar{k}_i}$ generated by the modality $\bar{m}_i$, the Modal-Affinity Distillation (MOD) loss is introduced to finely distill the knowledge from the teacher modality $\bar{m}_i$ to the student modality m:

$$\mathcal{L}_{MOD}^m = -\log \frac{\sum_{\langle t_i, \bar{m}_i \rangle \in LM^{(m)}} \sum_{\langle t_j, \bar{m}_j \rangle \in t_i^+} \phi(F_{t_i}^{(m)}, \mathbf{F}_{t_j}^{(\bar{m}_j)})}{\sum_{\langle t_i, \bar{m}_i \rangle \in LM^{(m)}} \sum_{\langle t_j, \bar{m}_j \rangle \in LM^{(m)}} \phi(F_{t_i}^{(m)}, \mathbf{F}_{t_j}^{(\bar{m}_j)})}. \tag{7}$$

Here t_i^+ indicates the positive samples in $LM^{(m)}$ sharing the same pseudo label with t_i. Furthermore, we utilize a copy of the feature representation $\mathbf{F}_{t_j}^{\bar{m}_j}$ to enable unidirectional information flow from the teacher modality to the student while minimizing the loss $\mathcal{L}_{MOD}^m$.

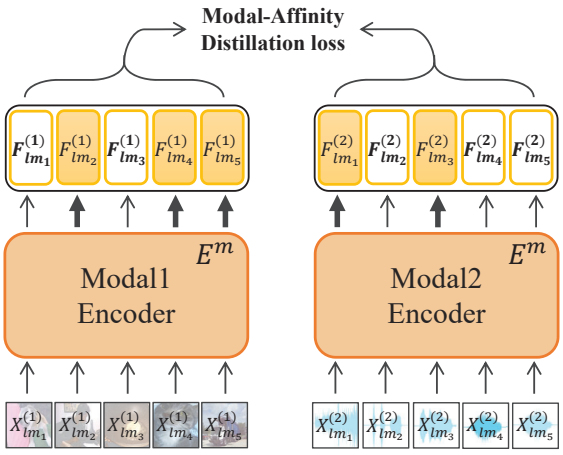


Figure 5. The process of Modal-Affinity Distillation. Each modality sends information flows to other modalities for samples it excels in, and receives information flows from other modalities for samples where it performs poorly.

3.6. Continuous Optimization

During the transition from the source domain S to the target domain T, feature distributions inevitably vary due to differences between domains. Consequently, continuing to use the class feature representation L extracted by the class encoder E^c trained in S may lead to inaccurate measurements. Therefore, it is essential to update E^c. Specifically, we employ the Modal-Affinity Metric Learning (MOML) loss of Eq. (2) to update E^c.

The overall loss function is formulated as follows:

$$\mathcal{L}_{all} = \sum_{m=1}^{M} \mathcal{L}_c^m + \lambda \mathcal{L}_{MOML}^m + \beta \mathcal{L}_{MOD}^m, \tag{8}$$

which combines the classification loss $\mathcal{L}_c^m$ for each modality m, MOML loss $\mathcal{L}_{MOML}^m$ weighted by λ, and MOD loss $\mathcal{L}_{MOD}^m$ weighted by β.

4. Experimental Results

In this section, we present an evaluation of the performance of our framework across various domain adaptation classification tasks, benchmarking it against several existing domain adaptation methodologies.

4.1. Datasets and Experimental Settings

Our evaluation employs three datasets covering diverse modalities. The source code will be released upon acceptance of the paper. For more details, please refer to the supplementary materials.

AVE [32] cover a broad spectrum of event types from animal sounds to human activities, along with diverse natural and man-made noises. The dataset features two modalities: image and audio. Following [21], this dataset was transformed into a collection of image-spectrogram pairs across 28 categories, comprising 41,728 source domain samples and 23,919 target domain samples. We utilize ResNet-18 [12] as the model backbone.

EPIC-Kitchens 55 [7]is a first-person perspective video dataset that includes fine-grained action categories. Video segments from participants P01, P08, and P22 represent three distinct domains: D1, D2, and D3, covering different kitchen layouts and lighting conditions. The number of action segments is 1,978, 3,245, and 4,871, respectively. We adopted the partitioning method from [22], categorizing tasks into eight action classes. The dataset features two modalities: RGB and optical flow. Following [22], we employ a dual-stream I3D network [4] as the model backbone.

CogBeacon [24] is a medium-sized collection dedicated to cognitive behavior analysis, featuring audio and video recordings of various human actions and interactions. The number of samples corresponding to the cognitive tasks for V1, V2, and V3 are 2,259, 2,221, and 2,389, respectively. The dataset features two modalities: EEG signals and facial key points. We utilize a three-layer one-dimensional ResNet network [12] as the model backbone.

4.2. Comparisons with State-of-the-Art Methods

We present the experimental results for the AVE, EPIC-Kitchens 55, and CogBeacon datasets in Tables 1, 2, and 3, respectively. These results are compared with traditional and contemporary domain adaptation methods, including DANN [9], CT [3], MM-SADA [22], , CL [15], RNA [25], C-RNA [26], DLMM [21], and MCT [28]. Additionally, MO [15] is a variant of CL [15], applies contrastive learning solely for cross-domain feature regularization.

In addition to comparisons with other methods, 'Source-only' serves as our experimental baseline, utilizing only labeled source domain data for training. 'Supervised-target,' which trains with target data that includes ground truth, represents the upper boundary for our experiments. All methods employ the same model architecture.

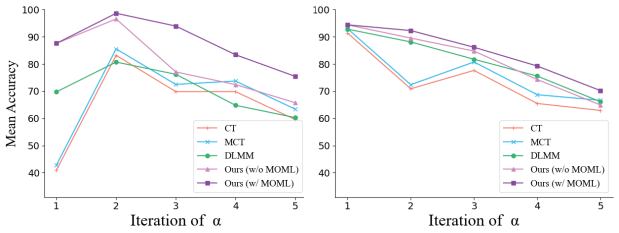
Furthermore, we explored several variants of our framework: Ours-unlock and Ours-logits. These versions correspond to employing Modal-Affinity Metric Learning (MOML) loss without unidirectional locking and using logits for evaluation outcomes. Experimental results across the AVE, EPIC-Kitchens 55, and CogBeacon datasets demonstrate that our framework consistently delivers superior performance across all modalities and fusion results.

Compared to methods like DANN [9] that optimize within individual modalities, or approaches like RNA [25] and C-RNA [26] that perform weak alignment of modal features, our proposed method enables substantial inter-modal information exchange. This enhances the distinct perspectives each modality brings to a common problem. Such information exchange is particularly evident in the AVE dataset, where the image modality, guided by the audio modality, improved its accuracy from 15.55% to 57.38%, an increase of 41.83%.

Furthermore, our proposed method provides fine-grained control to guide the flow of information at both the feature and label levels. In contrast to methods such as CT [3] , MMSADA [22], and CL [15] that treat all modalities equally without assessing their performances, our results are significantly better.

Table 1. Performance comparison on AVE.

Method	Image	Audio	Fusion
Source-only	15.55	43.08	44.13
CT [3]	47.42	49.12	49.47
DANN [9]	16.76	43.23	47.39
MM-SADA [22]	42.57	44.91	50.64
MO [15]	22.55	47.84	45.89
CL [15]	30.12	50.30	47.20
DLMM [21]	43.58	52.32	55.27
RNA [25]	16.62	44.09	43.52
C-RNA [26]	16.86	43.67	42.59
MCT [28]	52.38	52.45	54.56
Ours-unlock	53.17	54.55	55.74
Ours-logits	55.37	55.68	56.90
Ours	**57.38**	**57.66**	**58.11**
Supervised-target	66.83	79.79	83.55



(a) Image Modality (b) Audio Modality

Figure 6. Accuracy of Measurements for Image and Audio Modalities in the AVE Dataset.

4.3. More Results and Analysis

4.3.1. Accuracy of Affinity Measurements

To assess the reliability of our proposed affinity measurement, we tracked the accuracy of measurements across image and audio modalities during each iteration of domain adaptation on the AVE dataset. Our results were compared with those from CT [3], DLMM [21], and MCT [28]. Additionally, we evaluated the accuracy of our measurement without updating the class encoder (Ours w/o MOML). We set the learning threshold α to increment by 0.15 in each it-

Table 2. Performance comparison on EPIC-Kitchens 55. D1, D2, and D3 indicate three different domains.

Method	D1→D2			D2→D3			D3→D1			D2→D1			D3→D2			D1→D3			Mean		
	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion	RGB	Flow	Fusion
Source-only	36.7	46.4	47.7	35.1	46.9	46.9	38.1	43.8	44.0	37.6	45.1	43.1	45.7	53.8	55.9	36.3	42.0	42.7	38.3	46.3	46.7
CT [3]	46.8	48.8	49.0	46.1	50.4	51.6	41.7	45.3	46.7	41.2	45.1	45.6	46.4	55.7	56.4	37.4	41.1	42.3	43.3	47.7	48.6
DANN [9]	40.6	47.0	48.3	39.6	48.7	48.3	40.3	44.8	45.6	39.1	45.9	45.1	46.2	54.7	57.0	37.5	43.6	44.2	40.6	47.4	48.1
MM-SADA [22]	49.1	50.5	51.0	47.3	52.8	53.0	43.3	44.4	45.2	41.9	46.7	48.7	48.7	56.2	56.5	40.0	45.3	45.8	45.1	49.3	50.0
MO [15]	46.8	48.6	50.6	46.3	50.0	51.3	46.5	45.0	45.1	46.7	47.1	47.4	55.2	56.4	57.0	41.8	43.1	44.6	47.2	48.4	49.3
CL [15]	47.1	51.7	52.2	48.8	53.1	54.7	48.9	50.2	52.0	46.9	48.0	48.5	56.6	57.1	57.8	43.7	45.3	46.1	48.7	50.9	51.9
DLMM [21]	47.7	52.4	52.9	51.0	56.2	55.7	47.1	49.6	52.5	44.9	47.3	**50.4**	49.2	56.6	57.0	42.2	45.0	44.9	47.0	51.2	52.2
RNA [25]	43.3	47.3	51.5	47.2	49.1	56.2	42.6	43.0	42.8	45.4	46.2	50.0	47.6	54.5	54.7	42.5	44.9	44.3	44.8	47.5	49.9
C-RNA [26]	43.1	46.8	50.6	48.9	50.2	54.2	41.9	42.9	44.4	43.1	45.3	45.7	49.6	51.9	53.1	41.7	44.2	44.8	44.7	46.9	48.8
MCT [28]	48.3	49.9	52.7	51.4	53.1	53.7	44.6	49.1	50.3	43.2	46.3	46.7	48.3	56.4	57.1	40.8	45.6	46.4	46.1	50.1	51.2
Ours-unlock	53.8	53.6	54.6	55.5	57.3	58.3	51.3	53.0	53.4	44.5	47.9	47.5	50.4	54.3	55.8	44.3	46.5	47.4	50.0	52.1	52.8
Ours-logits	56.1	55.9	58.4	61.1	**60.3**	**61.4**	53.8	54.8	55.2	47.8	49.6	50.1	**57.1**	**57.3**	58.4	**45.6**	47.1	48.3	**53.6**	54.2	55.3
Ours	**56.2**	**56.0**	**58.6**	**61.2**	60.1	61.3	**53.9**	**55.1**	**55.6**	**48.1**	**49.6**	50.3	57.0	57.1	**58.7**	44.5	**48.2**	**49.0**	53.5	**54.4**	**55.6**
Supervised-target	66.8	72.3	73.6	70.5	72.4	74.5	61.6	63.9	65.2	61.6	63.9	65.2	66.8	72.3	73.6	70.5	72.4	74.5	66.3	69.5	71.1

Table 3. Performance comparison on CogBeacon. V1, V2, and V3 indicate three different domains.

Method	V1→V2			V2→V3			V3→V1			V2→V1			V3→V2			V1→V3			Mean		
	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion	FK	EEG	Fusion
Source-only	54.9	60.0	59.9	61.3	63.9	65.3	60.9	61.6	61.9	59.5	63.0	63.6	60.3	61.4	61.1	62.4	64.6	64.9	59.9	62.4	62.8
CT [3]	57.1	59.5	61.1	67.2	68.3	69.2	64.3	64.8	64.7	63.9	66.8	67.4	63.1	63.7	63.8	64.6	68.1	68.5	63.4	65.2	65.8
DANN [9]	54.5	62.8	64.3	62.3	65.5	67.4	64.0	63.5	64.1	62.3	63.7	64.8	61.1	63.3	63.3	64.5	67.2	67.9	61.5	64.3	65.3
MM-SADA [22]	56.0	63.8	64.6	63.8	67.4	68.3	62.2	63.8	64.5	64.3	66.0	66.8	64.1	66.4	65.0	66.7	68.4	70.1	62.8	66.0	66.6
MO [15]	56.9	64.9	61.9	63.0	66.2	68.4	61.8	64.3	64.9	62.6	66.4	67.0	62.5	63.7	64.2	65.0	69.8	71.1	61.9	65.9	66.3
CL [15]	57.8	65.7	65.8	62.5	70.0	71.4	61.7	63.4	64.3	63.1	66.8	67.1	62.3	64.1	63.8	66.1	70.1	71.0	62.3	66.7	67.2
DLMM [21]	60.3	63.4	65.0	67.5	70.2	71.5	63.1	65.1	65.4	64.6	67.1	67.2	62.2	64.3	65.6	66.6	69.2	70.2	64.1	66.6	67.5
RNA [25]	54.5	63.5	62.8	57.5	66.6	68.2	60.3	62.6	63.5	59.4	65.1	65.0	59.8	62.1	62.3	61.7	65.6	66.2	58.9	64.3	64.7
C-RNA [26]	54.2	63.5	63.1	62.0	69.6	70.1	60.6	63.4	63.9	59.0	65.4	65.1	59.9	62.9	62.5	61.1	65.8	67.0	59.5	65.1	65.3
MCT [28]	59.0	64.0	65.3	66.9	68.3	69.2	**64.7**	65.8	**66.7**	64.7	67.3	68.1	**64.3**	65.1	65.5	65.1	69.9	70.7	64.1	66.7	67.6
Ours-unlock	62.3	66.1	65.8	67.5	64.0	65.3	61.1	64.4	65.1	64.9	67.4	67.8	63.5	64.0	63.7	66.4	69.8	70.4	64.3	66.0	66.4
Ours-logits	62.8	67.4	67.5	68.4	69.1	69.1	62.5	65.3	66.0	67.1	68.7	69.0	64.0	65.1	65.6	67.5	71.3	72.6	65.4	67.8	68.3
Ours	**66.4**	**68.9**	**68.7**	**70.0**	**72.8**	**73.3**	64.3	**66.5**	66.4	**67.2**	**69.1**	**71.1**	63.9	**66.1**	**66.3**	**67.8**	**71.9**	**73.4**	**66.6**	**69.2**	**69.9**
Supervised-target	80.2	83.8	86.5	82.9	84.7	85.9	80.6	82.1	84.6	80.6	82.1	84.6	80.2	83.8	86.5	82.9	84.7	85.9	81.2	83.5	85.7

eration and utilize consistent modality encoders. As shown in Fig. 6, our affinity measurement consistently provided precise assessments in each modality. Furthermore, as E^c was continuously updated in the target domain, our measurement demonstrated a lesser reduction in accuracy over iterations compared to DLMM [21] and other methods that do not update their metrics.

4.3.2. Performance in Noisy Target Domains

As mentioned in the introduction, samples from target domain include those with significant domain shifts and noisy samples in certain modalities with unclear features that do not effectively represent their categories. The former allows the model to learn valuable features from the target domain, positively influencing both itself and other modalities during the domain adaptation process. The latter, being improperly handled during collection, turns into dirty samples. These not only fail to contribute positively to their modality but also negatively affect other modalities during the flow of information exchange.

In this experiment, we replaced 80% of the image samples in the AVE dataset with Gaussian noise images to test the efficacy of our proposed metric in distinguishing between noisy samples and domain-shifted but feature-rich samples across modalities. We compared our method with CL [15], DLMM[21], CT [3] and MCT [28]. The results, presented in Table 4, demonstrate that our method identifies quality samples among noisy ones at the modality level and submits them for learning by other modalities.

Table 4. Performance Comparison on AVE Dataset with Noise.

Method	Image	Audio	Fusion
Source-only	10.48	42.86	43.02
CL[15]	9.37	28.47	26.35
DLMM[21]	9.85	43.14	43.77
CT[3]	9.02	42.92	43.02
MCT[28]	10.85	43.29	43.53
Ours	**13.36**	**48.95**	**48.81**
Supervised-target	17.53	79.79	78.41

4.3.3. Cross-Modal Feature Distribution Analysis

Fig. 7 illustrates the t-SNE visualization results on the AVE dataset. Our method outperforms MCT [28] and DLMM [21] in metric accuracy and knowledge transfer, enabling the Image modality to acquire richer information from the Audio modality and form well-defined feature clusters. Compared to CL [15], our approach finely optimizes the flow of information and achieves more consistent alignment between source and target domain feature distributions, thereby enhancing cross-domain feature alignment.

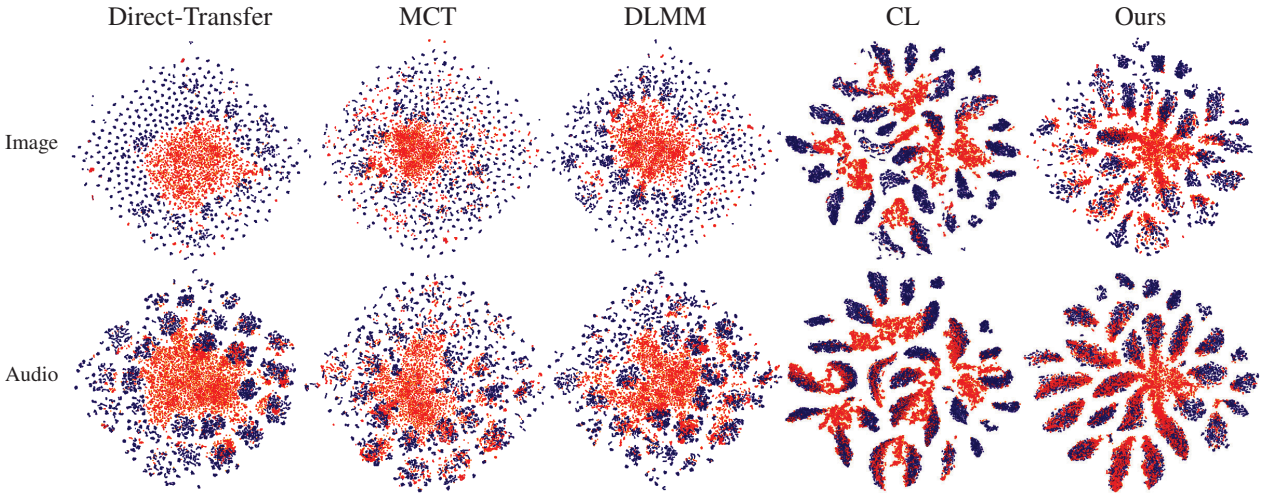


Figure 7. t-SNE plots of Image and Audio feature distribution in the AVE dataset produced by the baseline models and our proposed method. The source domain is shown in purple and the target domain is shown in orange.

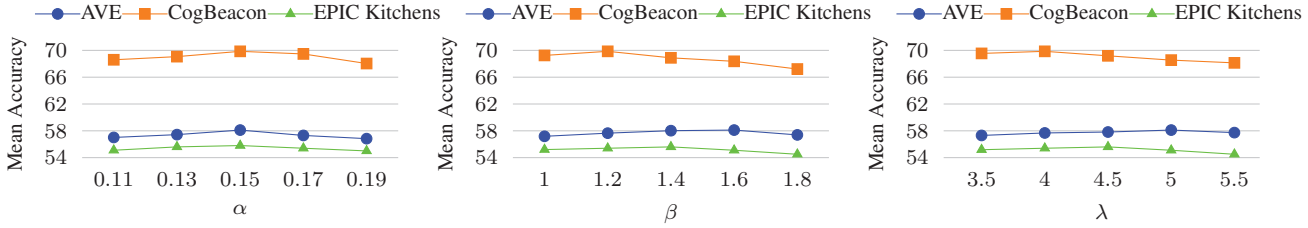


Figure 8. The accuracy of the *MODfinity* model with different configurations of the hyper-parameters α, β, and λ.

4.3.4. Parameter Sensitivity Analysis

The sensitivity of the hyper-parameters α, β, and λ applied in *MODfinity* is evaluated, with the results presented in Fig. 8. Specifically, α is the parameter that controls the threshold for the size of the learning material. Experimental results show that setting α to 0.15 has the best effect. β is the scale parameter used to adjust the intensity of affinity distillation. β that is too large can diminish accuracy in datasets of smaller sizes, leading us to set β to 1.2 in CogBeacon, as illustrated in Fig. 8. λ is the scale parameter for adjusting the intensity of affinity measurement training. If λ is set too high or too low, it can adversely affect accuracy, prompting us to set λ to 4.5 across all three tasks, as shown in Fig. 8.

4.3.5. Ablation Study

To systematically understand the contribution of each component within our method, we conducted an ablation study. As shown in Table 5, our proposed Modal-Affinity Distillation (MOD) enables differentiated information flow between modalities, substantially enhancing weaker modalities' performance under stronger ones' guidance. Meanwhile, Modal-Affinity Metric Learning (MOML) provides a fairer and more accurate affinity measurement that can be continuously updated during domain adaptation. This accurate metric offers superior target domain information for all modalities, significantly improving model accuracy.

Table 5. Ablation Study of Our Method on the AVE Dataset.

Method	Image	Audio	Fusion
Source-only	15.55	43.08	44.13
CE	49.90	47.83	49.65
CE+MOD	53.22	53.80	54.46
CE+MOML	51.17	53.60	54.55
CE+MOD+MOML	**57.38**	**57.66**	**58.11**
Supervised-target	66.83	79.79	83.55

5. Conclusion

In this paper, we introduce a modal-affinity distillation framework for multimodal domain adaptation. Extensive experiments across three multimodal learning tasks demonstrate that our framework consistently outperforms state-of-the-art multimodal domain adaptation methods. While our current approach primarily targets image and audio modalities, it establishes a robust foundation that can be readily extended to other data types.

One limitation is that our framework currently requires separate tuning for each modality pair, which may limit scalability as more modalities are added. In future work, we plan to explore training-free mechanisms to broaden the applicability and impact of our framework.

Acknowledgements

This work was supported by National Key R&D Program of China (2023YFA1011601), the Basic and Applied Basic Research Foundation of Guangdong Province (2024A1515012287), Science and Technology Key Program of Guangzhou (2023B03J1388), the Guangdong Natural Science Funds for Distinguished Young Scholars (Grant 2023B1515020097), the AI Singapore Programme under the National Research Foundation Singapore (Grant AISG3-GV-2023-011), the Lee Kong Chian Fellowships, and the National Natural Science Foundation of China (No.62302170), Guangdong Basic and the Applied Basic Research Foundation (No.2024A1515010187).

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 41(2):423–443, 2018. 2

[2] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010. 1

[3] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, 1998. 2, 4, 6, 7

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 6299–6308, 2017. 6

[5] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, pages 3339–3348, 2018. 1

[6] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. *NeurIPS*, 30, 2017. 2

[7] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, pages 720–736, 2018. 2, 6

[8] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pages 1180–1189. PMLR, 2015. 1

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(59):1–35, 2016. 1, 2, 6, 7

[10] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, pages 902–909. IEEE, 2010. 1

[11] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, pages 2827–2836, 2016. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[13] Xin Jiang, Hao Tang, Junyao Gao, Xiaoyu Du, Shengfeng He, and Zechao Li. Delving into multimodal prompting for fine-grained visual classification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2570–2578, 2024. 2

[14] Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing multi-modality completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[15] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *ICCV*, pages 13618–13627, 2021. 2, 6, 7

[16] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016. 2

[17] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 2

[18] Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, and Ying Gao. Diverse cotraining makes strong semi-supervised segmentor. *arXiv preprint arXiv:2308.09281*, 2023. 2

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[20] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *ICML*, pages 97–105. PMLR, 2015. 1

[21] Jianming Lv, Kaijie Liu, and Shengfeng He. Differentiated learning for multi-modal domain adaptation. In *ACM MM*, pages 1322–1330, 2021. 2, 4, 6, 7

[22] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, pages 122–132, 2020. 2, 4, 6, 7

[23] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011. 2

[24] Michalis Papakostas, Akilesh Rajavenkatanarayanan, and Fillia Makedon. Cogbeacon: A multi-modal dataset and data-collection platform for modeling cognitive fatigue. *Technologies*, 7(2):46, 2019. 2, 6

[25] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *WACV*, pages 1807–1818, 2022. 2, 6, 7

[26] Mirco Planamente, Chiara Plizzari, Simone Alberto Peirone, Barbara Caputo, and Andrea Bottino. Relative norm alignment for tackling domain shift in deep multi-modal classification. *IJCV*, pages 1–21, 2024. 2, 6, 7

[27] Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13325–13333, 2021. 2

[28] Jay C Rothenberger and Dimitrios I Diochnos. Meta co-training: Two views are better than one. *arXiv preprint arXiv:2311.18083*, 2023. 2, 3, 4, 6, 7

[29] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *CVPR*, pages 6956–6965, 2019. 1

[30] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *ICML*, pages 10096–10106. PMLR, 2021. 1

[31] Song Tang, Wenxin Su, Mao Ye, and Xiatian Zhu. Source-free domain adaptation with frozen multimodal foundation model. In *CVPR*, pages 23711–23720, 2024. 2

[32] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018. 2, 6

[33] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, pages 7472–7481, 2018. 1

[34] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *CVPR*, pages 7167–7176, 2017. 1, 2

[35] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, pages 2517–2526, 2019. 1

[36] Yifei Wang, Wen Li, Dengxin Dai, and Luc Van Gool. Deep domain adaptation by geodesic distance minimization. In *ICCV*, pages 2651–2657, 2017. 2

[37] Xuemiao Xu, Hai He, Huaidong Zhang, Yangyang Xu, and Shengfeng He. Unsupervised domain adaptation via importance sampling. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4688–4699, 2019. 2

[38] Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multimodal classification of alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011. 1