Adversarial Alignment with Anchor Dragging Drift (A^3D^2): Multimodal Domain Adaptation with Partially Shifted Modalities

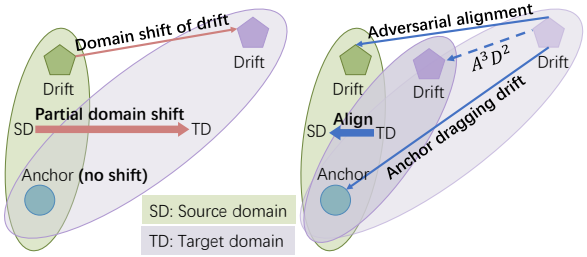
Jun Sun[1], Xinxin Zhang[2], Simin Hong[1]*, Jian Zhu[1,3], Lingfang Zeng[1]
[1]Zhejiang Lab, Hangzhou, China
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences
[3]University of Science and Technology of China
sunjun16sj@gmail.com, cliosimin@zhejianglab.org

Abstract

Multimodal learning has celebrated remarkable success across diverse areas, yet faces the challenge of prohibitively expensive data collection and annotation when adapting models to new environments. In this context, domain adaptation has gained growing popularity as a technique for knowledge transfer, which, however, remains underexplored in multimodal settings compared with unimodal ones. This paper investigates multimodal domain adaptation, focusing on a practical partially shifting scenario where some modalities (referred to as anchors) remain domain-stable, while others (referred to as drifts) undergo a domain shift. We propose a bi-alignment scheme to simultaneously perform drift-drift and anchor-drift matching. The former is achieved through adversarial learning, aligning the representations of the drifts across source and target domains; the latter corresponds to an "anchor dragging drift" strategy, which matches the distributions of the drifts and anchors within the target domain using the optimal transport (OT) method. The overall design principle features **A**dversarial **A**lignment with **A**nchor **D**ragging **D**rift, abbreviated as $\boldsymbol{A^3D^2}$, for multimodal domain adaptation with partially shifted modalities. Comprehensive empirical results verify the effectiveness of the proposed approach, and demonstrate that A^3D^2 achieves superior performance compared with state-of-the-art approaches. The code is available at: https://github.com/sunjunaimer/A3D2.git.

1 Introduction

Multimodal learning, which leverages heterogeneous and complementary signals to perform standard machine learning tasks, has risen to prominence in a broad spectrum of applications, such as medical analysis (Wang et al., 2024), social media (Yu et al., 2023), and affective computing



(a) Partial domain shift (b) Domain alignment with A^3D^2

Figure 1: Multimodal domain adaptation for partial domain shift.

(Guo et al., 2024). Nevertheless, constructing high-quality multimodal datasets for model training is prohibitively expensive, as data collection requires multiple devices or sensors, and annotation demands extensive manual effort. To this end, unsupervised domain adaptation, which aims to transfer knowledge from a label-rich source domain to a related but unlabeled target domain, is prevalent for mitigating the scarcity of annotated data. Domain adaptation typically aligns the target and source domains during training, thereby enabling the model supervised with only source domain labels to generalize well for the target domain.

Unsupervised domain adaptation has been extensively studied in the computer vision and natural language processing communities separately, showcasing impressive results in a multitude of unimodal tasks including image classification (Hoyer et al., 2023), object detection (Du et al., 2024), question answering (Zhang et al., 2024c), among others. In contrast, multimodal domain adaptation remains relatively underexplored and has garnered increasing research interest in recent years (Zhang et al., 2024a; Dong et al., 2025). A distinct characteristic of multimodal settings is that different modalities reside in separate physical spaces, and hence they may experience varying degrees of domain shift when the environment changes.

In this paper, we particularly consider the par-

*Corresponding author

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 19680–19690

tially shifting scenario, as illustrated in Figure 1(a), where all modalities are classified into anchor modalities (or anchors) and drift modalities (or drifts): the former stays unchanged, while the latter undergoes domain shifts. This phenomenon is commonly encountered in practical applications. For instance, a conversational robot is equipped with a multimodal emotion recognition system that utilizes acoustic, lexical, and visual information to detect a speaker's emotions. When the working scenario shifts from day to night, the changes in illumination conditions induce a domain shift in the visual modality, whereas the acoustic and lexical modalities remain stable. To the best of our knowledge, this work represents the first effort to identify and investigate this distinct yet practical case of multimodal domain adaptation.

Towards the goal of minimizing the discrepancy between the source domain and the partially shifted target domain, this work develops a multimodal domain adaptation (MMDA) framework, leveraging techniques from information bottleneck (IB) theory (Saxe et al., 2019; Kawaguchi et al., 2023), adversarial learning (Long et al., 2018; Chen et al., 2022a), and optimal transport (OT) (Fatras et al., 2021). Specifically, we first construct the model with pretrained backbones for each modality. In order to retain general knowledge while adapting to new tasks and domains, the pretrained backbones are partially finetuned with some layers frozen. Then, we apply IB theory to formulate the training objective, for the sake of attaining informative representations and promoting modality independence. Through the IB method, we enforce each modality to independently perform label prediction, thus preventing some "lazy" modalities from being under-trained (Sun et al., 2023).

Subsequently, in the representation space, as Figure 1(b) shows, domain gap is reduced using two strategies: 1) drift-drift alignment— matching the drift across the source and target domains; that is, we conduct adversarial alignment (AA) for each drift using its representation and label prediction information to achieve category-level alignment. 2) anchor-drift alignment— matching the anchor and drift within the target domain; namely, we develop an OT-based anchor-dragging-drift (ADD) approach to push the anchor and drift closer, which facilitates the promotion of domain alignment.

In summary, the present work proposes a novel approach, **A**dversarial **A**lignment with **A**nchor **D**ragging **D**rift ($\boldsymbol{A^3D^2}$), for MMDA with partially shifted modalities. The primary contributions are threefold.

1. We investigate, for the first time, a practical partial domain shift scenario in multimodal learning and propose a novel MMDA framework. Each modality learns its representation and predicts labels, both of which are used by the adversarial discriminator to align the source and target domains.
2. To boost the domain alignment, we exploit connections between the anchor and drift, and propose to match their representation distributions using the OT method.
3. Extensive experiments conducted on widely used benchmark datasets demonstrate the superior performance of A^3D^2 compared to competing approaches.

2 Related Works

2.1 Domain adaptation approaches

A plethora of works have been devoted to domain adaptation, which can be broadly grouped into three categories: statistical, adversarial, and optimal transport (OT) methods.

Statistical methods: Statistical methods usually learn domain-invariant representations via minimizing the moment-based distribution discrepancy of the target and source domains. Maximum mean discrepancy (MMD) based methods, such as DDC (Tzeng et al., 2014) and MK-MMD (Long et al., 2015), focus on aligning the first-order moment (i.e., the mean) of the representations. Coral (Sun et al., 2016) and JDDA (Chen et al., 2019) are typical second-order moment approaches, which matches the covariance of the representations. Furthermore, CMD (Zellinger et al., 2017) extends to high-order moments matching, aligning the central moments (mean, variance, skewness, etc.).

Adversarial methods: Starting from the pioneering work DANN (Ganin et al., 2016), numerous studies have applied adversarial methods to align the representation from the source and target domains. MDAN (Zhao et al., 2018) addresses multiple source domain adaptation and devises two versions of optimization strategies. Label prediction information is introduced as a condition for domain alignment in CDAN (Long et al., 2018), MADA (Pei et al., 2018) and CAN (Wu et al., 2021). A discriminator-free adversarial model is developed via reusing the task-specific classifier as a discriminator in DALN (Chen et al., 2022a). CDA (Yadav

et al., 2023) and LUHP (Zhang et al., 2024b) integrate contrastive learning into domain adaptation to achieve class-level alignment. f-DD (Wang and Mao) introduces a novel measure, f-domain discrepancy, for adversarial domain adaptation, and obtains new target error and sample complexity bounds. Other adversarial methods, PCL (Li et al., 2024) and DADA(Ren et al., 2024), incorporate data augmentation from the raw feature space and representation space, respectively.

Optimal transport (OT) methods: Optimal transport involves measuring the discrepancy between two distributions and matching them, which has gained popularity in domain adaptation recently due to its solid theoretical support. COT (Liu et al., 2023) formulates the domain alignment as an optimal transport problem to construct a mapping between clustering centers in the source and target domains. InfoOT (Chuang et al., 2023) propose an information theoretic extension of OT that maximizes the mutual information between domains while minimizing geometric distances. Work (Montesuma et al., 2024) devises an efficient OT-based domain adaptation method for Gaussian mixture models. To reduce the computational complexity of conventional OT, UMOT (Fatras et al., 2021) and POT (Nguyen et al., 2022) propose unbalanced and partial mini-batch optimal transport for domain adaptation, respectively, enabling OT methods to be applicable for large-scale data.

2.2 Multimodal domain adaptation

Compared with its unimodal counterpart, multimodal domain adaptation remains significantly less explored. MM-SADA (Munro and Damen, 2020) combines multimodal self-supervised alignment with within-modal adversarial alignment for MMDA. MD-DMD (Yin et al., 2022) proposes dynamically distilling knowledge across modalities in adversarial learning to boost adaptability. The study (Kim et al., 2021) develops a contrastive learning approach with properly designed sampling strategies to simultaneously regularize cross-modal and cross-domain feature representations. MC-TTA (Xiong et al., 2024) utilizes memory banks and self-assembled source-friendly feature reconstruction to enhance multimodal prototype alignment and cross-modal relative consistency. Novel language-guided domain divergence measurement losses are proposed in CLIP-Div (Zhu et al., 2024), which designs a language-guided pseudo-labeling strategy for calibrating the target pseudo labels in vision-language domain adaptation tasks. Additionally, the recent survey paper (Dong et al., 2025) provides a more comprehensive introduction to current research on MMDA.

In light of the aforementioned prior works, this paper investigates MMDA in a scenario of practical value, and proposes A^3D^2, building upon the adversarial and optimal transport methods.

3 Method: $\mathbf{A^3D^2}$

Before diving into the proposed method, A^3D^2, we give the definitions of some notations to be used.

Notations: For any positive integer I, let $[I]^+$ and $[I]$ denote the set $\{1, 2, \cdots, I\}$ and $\{0, 1, \cdots, I\}$, respectively. When $\boldsymbol{M}$ denotes a matrix, we use vector $[\boldsymbol{M}]_i$ to denote its i-th column. Let $\mathbf{1}$ denote any all-one vector of proper size. Without particular statement, all vectors in this paper are supposed to be column vector. Notation ":=" is used for definition.

Suppose that in the multimodal learning setting, there are $M+1$ independent modalities indexed by $0, 1, 2, \cdots, M$. For consistency of expression, we introduce an auxiliary modality, with index $M+1$, as the joint of all the $M+1$ independent modalities, which results in a total number of $M+2$ modalities. The training dataset is denoted as $\{\{\boldsymbol{x}_{m,n}\}_{m\in[M]}, \{\boldsymbol{y}_{n,m}\}_{m\in[M+1]}\}_{n\in[N]^+}$, where N is the number of samples, n indexes the samples, p_m-dimensional vector $\boldsymbol{x}_{m,n}$ represents the raw feature of modality m, and $\boldsymbol{y}_{n,m}$ is the label. In some cases where all modalities share the same label, $\boldsymbol{y}_{n,0} = \boldsymbol{y}_{n,1} = \cdots = \boldsymbol{y}_{n,M+1}$ holds. Suppose there are C categories in the classification task; then label $\boldsymbol{y}_{n,m}$ can be either a one-hot vector or a scalar in $[C]^+$, and we adopt either form as necessary in the rest of the paper. For consistency, we use $\boldsymbol{x}_{n,M+1} := [\boldsymbol{x}_{n,0}; \boldsymbol{x}_{n,1}; \cdots; \boldsymbol{x}_{n,M}]$ to aggregate all modalities. Let $\boldsymbol{X}_m$ and $\boldsymbol{Y}_m$ be general feature and label random variables for all $m \in [M+1]$, with $\boldsymbol{x}_{n,m}$ and $\boldsymbol{y}_{n,m}$ being their specific instances.

As mentioned above, in the considered partial domain shift scenario, some modalities remain stable across the source and target domains, and others undergo shift; the former is referred to as anchors and the latter is drifts. For brevity of description and without loss of generality, we assume there are one anchor and multiple drifts; that is, modality 0 represents the anchor, and all the rest M independent modalities are the drifts. We use superscript s

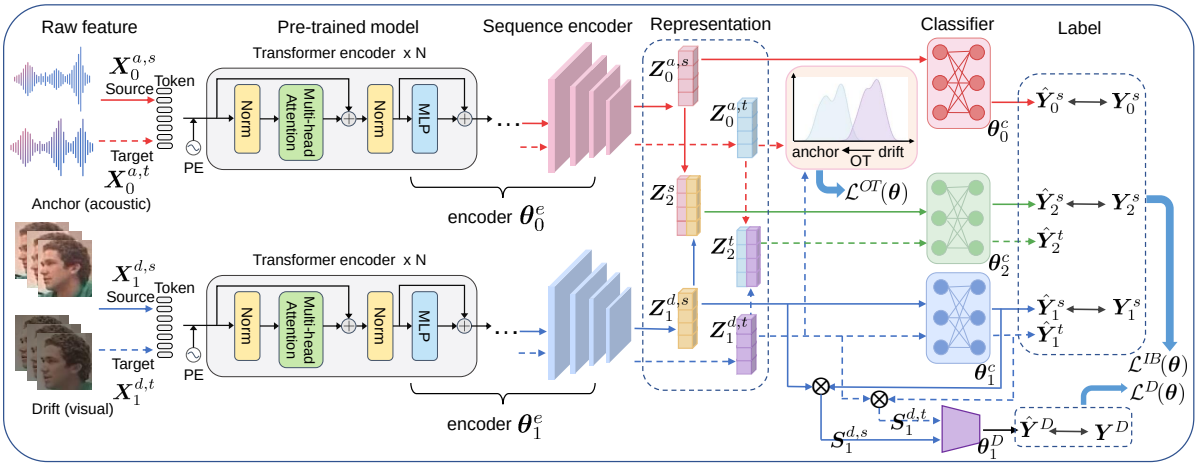


Figure 2: Model framework with 2 modalities as an example (multimodal representation $\boldsymbol{Z}_2$ is a concatenation of $\boldsymbol{Z}_0$ and $\boldsymbol{Z}_1$; solid and dashed regular arrows denote the flows of source and target domains, respectively; red, blue and green arrows represent the data flow of modalities 0, 1, and 2, respectively; double-headed arrows supervision signals, corresponding to the information bottleneck loss $\mathcal{L}^{IB}(\boldsymbol{\theta})$, domain discrimination loss $\mathcal{L}^{D}(\boldsymbol{\theta})$ and optimal transport loss $\mathcal{L}^{OT}(\boldsymbol{\theta})$).

and t to distinguish the source and target domains, and use a and d to distinguish the anchor and drift. For example, $\boldsymbol{X}_m^{d,s}, m \in [M]^+$ represents the feature of drift m from source domain. Each sample is associated with a domain label $\boldsymbol{Y}^D \in 0, 1$, indicating whether it belongs to the source domain (0) or the target domain (1).

In the sequel, we will present our model framework and derive the training objective function, including IB based representation learning, adversarial alignment (AA) of the source and target domains, and anchor-dragging-drift (ADD) strategy using optimal transport scheme.

3.1 Model architecture

In this section, we focus on the model framework and ignore the implementation details, which will be elaborated later in the Numerical Results section. Figure 2 illustrates the proposed multimodal domain adaptation framework in an example with two modalities: one anchor (acoustic) and one drift (visual), namely, $M = 1$. The raw features $\boldsymbol{X}_m, \forall m \in [M]$ are first tokenized and fed into the pretrained transformer-based models, of which the top layers will be finetuned. Following the pretrained models are sequence encoders which further encode the sequence features into a p-dimensional vector representation $\boldsymbol{Z}_m, \forall m \in [M]$. More formally, for each modality $m \in [M]$, the corresponding pretrained model and the sequence encoder can be summarized by a deterministic encoder function $f_m^e(\cdot; \boldsymbol{\theta}_m^e) : \mathbb{R}^{p_m} \to \mathbb{R}^p$ with trainable parameter $\boldsymbol{\theta}_m^e$; then we have $\boldsymbol{Z}_m = f_m^e(\boldsymbol{X}_m; \boldsymbol{\theta}_m^e)$ ($\boldsymbol{Z}_m$ is normalized with ℓ_2-norm being 1). The multimodal representation is denoted by $\boldsymbol{Z}_{M+1} := [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_M]$, a concatenation of the representations of all modalities.

Each modality $m \in [M+1]$ is associated with a classifier $f_m^c(\cdot, \boldsymbol{\theta}_m^c)$ with parameter $\boldsymbol{\theta}_m^c$ for label prediction; that is, $\hat{\boldsymbol{Y}}_m = f_m^c(\boldsymbol{Z}_m, \boldsymbol{\theta}_m^c)$. The multimodal prediction $\hat{\boldsymbol{Y}}_{M+1}$ is assigned to be the ultimate predicted label.

For each drift $m, m \in [M]^+$, we will train a domain discriminator m with parameter $\boldsymbol{\theta}_m^D$ to facilitate the domain alignment, which will be introduced latter. For brevity of expression, we use $\boldsymbol{\theta} := \{\boldsymbol{\theta}_{M+1}^c\} \cup \{\boldsymbol{\theta}_m^e, \boldsymbol{\theta}_m^c\}_{m\in[M]} \cup \{\boldsymbol{\theta}_m^D\}_{m\in[M]^+}$ to collect all model parameters, and $\boldsymbol{\theta}^e := \{\boldsymbol{\theta}_m^e\}_{m\in[M]}$ to collect the parameters of all encoders.

3.2 IB-based representation learning

With the above model framework, the information follows $\boldsymbol{X}_m \to \boldsymbol{Z}_m \to \boldsymbol{Y}_m, \forall m \in [M+1]$. Information bottleneck based representation learning aims to obtain representations $\boldsymbol{Z}_m^s, \forall m \in [M+1]$, such that they capture generalizable features, and thereby alleviate the difficulty in aligning the source and target representations.

Mathematically, the optimal representation is generated by minimizing the following IB loss on the source domain:

$$\mathcal{L}^{IB}(\boldsymbol{\theta}) := \sum_{m\in[M+1]} \beta I(\boldsymbol{X}_m^s, \boldsymbol{Z}_m^s) - I(\boldsymbol{Z}_m^s, \boldsymbol{Y}_m^s), \quad (1)$$

where $I(\cdot,\cdot)$ denotes the mutual information of any two random variables, and β is a predefined coefficient.

From the perspective of information theory, it is obvious that the resultant representation $\boldsymbol{Z}_m^s$ retains minimal information from the raw feature $\boldsymbol{X}_m^s$ yet maintains the maximal information of the label $\boldsymbol{Y}_m^s$. Therefore, $\boldsymbol{Z}_m^s$ is an optimal representation in the sense of information bottleneck theory (Saxe et al., 2019; Kawaguchi et al., 2023). Moreover, each individual modality m, for all $m \in [M]$, is enforced to generate its own optimal representation, which promotes modality independence and prevents some weak modalities from being dominated by strong ones.

Then, we specify how the two information terms in Eq. (1) are computed.

$$\begin{aligned} I(\boldsymbol{X}_m^s, \boldsymbol{Z}_m^s) &= H(\boldsymbol{Z}_m^s) - H(\boldsymbol{Z}_m^s|\boldsymbol{X}_m^s) \\ &= H(\boldsymbol{Z}_m^s) = \mathbb{E}_{\boldsymbol{Z}_m^s}[-\log p(\boldsymbol{Z}_m^s)], \end{aligned} \tag{2}$$

where $H(\cdot)$ represents entropy; and $H(\boldsymbol{Z}_m^s|\boldsymbol{X}_m^s) = 0$, since $\boldsymbol{Z}_m = f_m^e(\boldsymbol{X}_m; \boldsymbol{\theta}_m^e)$ is a deterministic function. It is a convention to assume that $\boldsymbol{Z}_m^s$ follows Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_m^s, \boldsymbol{\Sigma}_m^s)$ ($\boldsymbol{\mu}_m^s \in \mathbb{R}^p$, and $\boldsymbol{\Sigma}_m^s \in \mathbb{R}^{p\times p}$ is a diagonal matrix). As a result, we can estimate $\boldsymbol{\mu}_m^s$ and $\boldsymbol{\Sigma}_m^s$ with the representations $\boldsymbol{z}_{n,m}^s, n \in [N^s]^+$, and hence the entropy of $H(\boldsymbol{Z}_m^s)$ can be obtained as following:

$$H(\boldsymbol{Z}_m^s) = \frac{1}{2}\log|\boldsymbol{\Sigma}_m^s| + \frac{d}{2}(1+\log(2\pi)), \tag{3}$$

where $|\boldsymbol{\Sigma}_m^s|$ represents the determinant of $\boldsymbol{\Sigma}_m^s$.

Similarly, $I(\boldsymbol{Z}_m^s, \boldsymbol{Y}_m^s)$ can be written as:

$$\begin{aligned} I(\boldsymbol{Z}_m^s, Y_m^s) &= H(\boldsymbol{Y}_m^s) - H(\boldsymbol{Y}_m^s|\boldsymbol{Z}_m^s) \\ &= H_{Y,m}^s - H(\boldsymbol{Y}_m^s|\boldsymbol{Z}_m^s) \\ &= H_{Y,m}^s + \frac{1}{N^s}\sum_{n=1}^{N^s}\log p(\boldsymbol{y}_{n,m}^s|\boldsymbol{z}_{n,m}^s), \end{aligned} \tag{4}$$

where $H(\boldsymbol{Y}_m^s) = H_{Y,m}^s$ is a constant independent from the model parameter $\boldsymbol{\theta}$.

Combining Eqs. (1), (2), (3) and (4) gives the information bottleneck loss as follows (with constant terms omitted):

$$\begin{aligned} \mathcal{L}^{IB}(\boldsymbol{\theta}) = \sum_{m=0}^{M+1}\Big[&\frac{\beta}{2}\log|\boldsymbol{\Sigma}_m^s| \\ &- \frac{1}{N^s}\sum_{n=1}^{N^s}\log p(\boldsymbol{y}_{n,m}^s|\boldsymbol{z}_{n,m}^s)\Big], \end{aligned} \tag{5}$$

where the first term is a regularization for the representation that suppresses the noisy and ineffective information; and the second term corresponds to the negative log-likelihood of the prediction (equivalent to cross-entropy loss).

3.3 AA: drift-drift alignment

With the aforementioned IB-based learning approach, we can separately attain the representation ($\boldsymbol{Z}_m$) and label prediction ($\hat{\boldsymbol{Y}}_m$) of each modality $m, \forall m \in [M]$ separately. Then, for each drift $m, m \in [M]^+$, we align its source and target domains from the representation space using an adversarial method.

Taking inspiration from CDAN, we incorporate the category (i.e, label) information into the representation when aligning the source and target domains for the drifts. Specifically, as shown in Figure 2, the input to the domain discriminator $\mathcal{D}$ is the Kronecker product of the representation $\boldsymbol{Z}_m$ and the predicted label probability $\hat{\boldsymbol{Y}}_m$ for each drift $m, m \in [M]^+$; that is, $\boldsymbol{S}_m := \boldsymbol{Z}_m \otimes \hat{\boldsymbol{Y}}_m$. The output is the domain prediction, $\hat{\boldsymbol{Y}}^D$, which is

The adversarial alignment boils down to solving the following min-max problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}^{\mathcal{G}}}\max_{\boldsymbol{\theta}^{\mathcal{D}}}\ \mathcal{L}^D(\boldsymbol{\theta}) := \sum_{m=1}^{M}\Bigg[&\frac{1}{N^t}\sum_{n=1}^{N^t}\log\left(\mathcal{D}(\boldsymbol{s}_{n,m}^{d,t})\right) \\ &+ \frac{1}{N^s}\sum_{n=1}^{N^s}\log\left(1-\mathcal{D}(\boldsymbol{s}_{n,m}^{d,s})\right)\Bigg]. \end{aligned} \tag{6}$$

where we define $\boldsymbol{\theta}^{\mathcal{G}} := \{\boldsymbol{\theta}_m^e, \boldsymbol{\theta}_m^c\}_{m\in[M]^+}$ and $\boldsymbol{\theta}^{\mathcal{D}} := \{\boldsymbol{\theta}_m^D\}_{m\in[M]^+}$ for brevity.

Through adversarial alignment (AA), we achieve drift-drift alignment across the source and target domains. In what follows, we introduce anchor-dragging-drift (ADD) strategy for anchor-drift alignment within the target domain.

3.4 ADD: anchor-drift alignment

As shown in Figure 1, the drift in the target domain undergoes shifts relative to both the anchor and the drift in the source domain. In this section, we take advantage of the anchor to drag the drift closer in the target domain, which indirectly reduces the gap between the source and target domains. This is achieved using the optimal transport (OT) method, which is commonly adopted to match two distributions. Different from conventional OT methods that transfer the samples between domains, the proposed OT scheme in this work aligns two distri-

butions from the perspective of dimension-wise representation.

In specific, let $\tilde{\boldsymbol{Z}}_m$ collect the representations of all samples for modality m; namely, $\tilde{\boldsymbol{Z}}_m := [\boldsymbol{z}_{1,m}^T; \boldsymbol{z}_{2,m}^T; \cdots, \boldsymbol{z}_{N,m}^T] \in \mathbb{R}^{N\times p}$. The i-coordinate ($i = 1, 2, \cdots, p$) representation $[\tilde{\boldsymbol{Z}}_m]_i$ of all samples (or a batch of samples in stochastic gradient based training) is analogous to a sample in typical OT methods. Let $\mathcal{Z}_m^{d,t} := \{[\tilde{\boldsymbol{Z}}_m^{d,t}]_i\}_{i=1}^p$ and $\mathcal{Z}_0^{a,t} := \{[\tilde{\boldsymbol{Z}}_0^{a,t}]_i\}_{i=1}^p$ be the dimension-wise representations of the drift and anchor in the target domain, respectively. The corresponding empirical distributions are denoted by $\boldsymbol{\mu}_m^{d,t} := \frac{1}{p}\sum_{i=1}^p \delta_{[\tilde{\boldsymbol{Z}}_m^{d,t}]_i}$ and $\boldsymbol{\mu}_0^{a,t} := \frac{1}{p}\sum_{i=1}^p \delta_{[\tilde{\boldsymbol{Z}}_0^{a,t}]_i}$, where $\delta_{(\cdot)}$ is the Dirac at location $(\cdot)$.

The Kantorovich optimal transport (Courty et al., 2017) between $\mu_m^{d,t}$ and $\mu_0^{a,t}$ is formulated as a convex optimization problem:

$$\boldsymbol{\pi}_m^* := \arg\min_{\boldsymbol{\pi}_m \in \boldsymbol{\Pi}_m} \langle \boldsymbol{C}_m, \boldsymbol{\pi}_m \rangle, \tag{7}$$

where $\boldsymbol{C}_m := \mathbf{1}_{p\times p} - \frac{1}{N}\tilde{\boldsymbol{Z}}_m^T \cdot \tilde{\boldsymbol{Z}}_0$ is the dimension-wise distance (or cost) between $\tilde{\boldsymbol{Z}}_m$ and $\tilde{\boldsymbol{Z}}_0$, induced by similarity (recalling that $\boldsymbol{z}_{n,m}$ has ℓ_2-norm of 1, and $\mathbf{1}_{p\times p}$ represents a $p \times p$ all-one matrix); the solution $\boldsymbol{\pi}_m^*$ is known as transport plan given $\boldsymbol{C}_m$; $\boldsymbol{\Pi}_m$ as defined below represents a distribution space where the distribution is associated with marginals $\boldsymbol{\mu}_m^{d,t}$ and $\boldsymbol{\mu}_0^{a,t}$.

$$\boldsymbol{\Pi}_m = \{\boldsymbol{\pi} \in \mathbb{R}^{p\times p} | \boldsymbol{\pi} \cdot \mathbf{1} = \boldsymbol{\mu}_m^{d,t};\ \boldsymbol{\pi}^T \cdot \mathbf{1} = \boldsymbol{\mu}_0^{a,t}\}.$$

For any fixed $\boldsymbol{C}_m$, the Wasserstein distance between the drift distribution $\boldsymbol{\mu}_m^{d,t}$ and the anchor distribution $\boldsymbol{\mu}_0^{a,t}$ is defined as:

$$W_m := \min_{\boldsymbol{\pi}_m \in \boldsymbol{\Pi}_m} \langle \boldsymbol{C}_m, \boldsymbol{\pi}_m \rangle = \langle \boldsymbol{C}_m, \boldsymbol{\pi}_m^* \rangle. \tag{8}$$

During model training, we optimize $\boldsymbol{C}_m$ to enable the anchor to drag the drift closer; that is, the Wasserstein distance $W_m, m \in [M]^+$ between drift m and the anchor is minimized, which translates to minimizing the OT loss function:

$$\mathcal{L}^{OT}(\boldsymbol{\theta}) := \sum_{m=1}^{M} \langle \boldsymbol{C}_m, \boldsymbol{\pi}_m^* \rangle. \tag{9}$$

3.5 A^{3}D^2: bi-alignment for MMDA

With the above formulations, our bi-alignment approach, A^3D^2, incorporating AA and ADD, is tantamount to solving the optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{\theta}; \boldsymbol{\pi}_m \in \boldsymbol{\Pi}_m} \left[\mathcal{L}^{IB}(\boldsymbol{\theta}) + \gamma_1 \mathcal{L}^{OT}(\boldsymbol{\theta})\right] \\ &+ \gamma_2 \left[\min_{\boldsymbol{\theta}^{\mathcal{G}}} \mathcal{L}^D(\boldsymbol{\theta}) - \min_{\boldsymbol{\theta}^{\mathcal{D}}} \mathcal{L}^D(\boldsymbol{\theta})\right], \end{aligned} \tag{10}$$

Algorithm 1 A^3D^2: bi-alignment for MMDA

1: **Initialization:** initialize model parameter $\boldsymbol{\theta}^0$.
2: **for** $k = 0$ to $K - 1$ **do**
3: perform forward and calculate $\boldsymbol{C}_m$;
4: solve the problem in Eq. (7) using a covex problem solver to obtain $\boldsymbol{\pi}_m^*, \forall m \in [M]^+$;
5: calculate the overall objective as in Eq. (10) and perform backward pass to compute stochastic gradient;
6: update the model parameter using an optimizer (e.g., Adam): $\boldsymbol{\theta}^{k+1} \leftarrow \text{optimizer}(\boldsymbol{\theta}^k, \alpha)$.
7: **end for**
8: **Return:** Model parameter $\boldsymbol{\theta}^K$.

where γ_1 and γ_2 are constant coefficients balancing the losses.

The model training framework is summarized in **Algorithm 1**, where k indexes the iteration and α is the learning rate. In each iteration, we first calculate $\boldsymbol{C}_m$, and then employ an off-the-shelf convex optimization solver to solve problem in in Eq. (7) to obtain $\boldsymbol{\pi}_m^*$; with the transport plan $\boldsymbol{\pi}_m^*$, we compute the overall loss function as in Eq. (10), followed by the backward pass and model update.

4 Experiments

Benchmark datasets: We evaluate our method on two benchmark datasets, IEMOCAP (Busso et al., 2008) and MIntRec (Zhang et al., 2022), both containing acoustic, visual, and lexical modalities. IEMOCAP is for the emotion recognition task, composed of scripted and spontaneous dyadic conversations between actors. Following work (Zhao et al., 2021), we select samples from the four classes — neutral, happy, sad and angry, to construct the dataset for our experiments. MIntRec is a dataset collected from the TV series Superstore for intent recognition with 20 intent categories.

For each dataset, we split it evenly and randomly into two subsets. One subset is used directly as the source domain dataset, and the other, after some manipulations, serves as the target domain dataset. Specifically, for the target domain samples of the IEMOCAP dataset, we inject Gaussian noise with mean=0 and variance=0.01 into the drift acoustic modality; for the drift visual modality, the brightness of the video is reduced to 10 percent of its original level, and Gaussian noise with mean=0 and variance=0.1 is added to each frame that is

Methods	IEMOCAP				MIntRec			
	A	L	V	ave.	A	L	V	ave.
DT	54.15	58.67	57.59	56.80	47.84	61.42	44.66	51.31
DANN	62.20	65.17	63.38	63.58	49.10	64.32	49.89	54.44
CDAN	58.30	64.51	68.26	63.69	51.63	63.24	53.05	55.97
MADA	62.08	62.07	64.14	62.76	50.74	64.57	48.30	54.54
DALN	61.83	70.25	64.97	65.68	24.47	28.29	24.39	25.72
PCL	61.96	61.88	61.43	61.76	52.25	65.42	49.26	55.64
f-DD	61.59	64.60	59.67	61.95	51.28	59.49	47.96	52.91
LUHP	64.75	60.59	60.69	62.01	53.61	63.38	53.87	56.59
DADA	62.82	64.40	65.34	64.19	52.49	**66.37**	52.80	57.22
$\boldsymbol{A^3D^2}$	**65.15**	**70.84**	**69.20**	**68.40**	**54.46**	66.08	**53.96**	**58.17**

Table 1: The performance (in terms of F1 score) comparisons of A^3D^2 and the existing methods (the highest and second highest F1 scores in each column are highlighted with bold font and blue color, respectively).

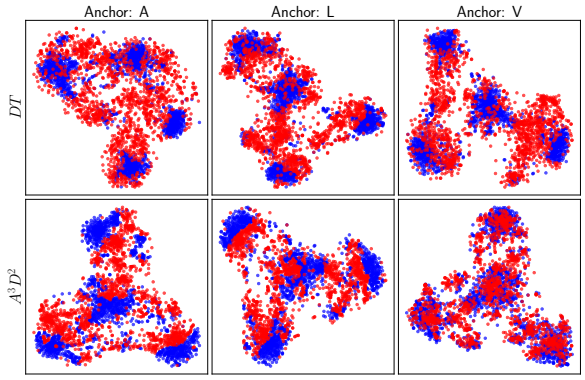


Figure 3: t-SNE projection of the representations of source (blue) and target (red) domain samples.

normalized with mean=(0.485, 0.456, 0.406) and variance=(0.229, 0.224, 0.225) for the three channels; for the drift lexical modality, 40 percent of the words in each utterance are randomly selected and masked. For the MIntRec dataset, Gaussian noise with mean=0 and variance=0.005 is injected to the acoustic modality; the brightness of the video is reduced to 30 percent of its original level, and Gaussian noise with mean=0 and variance=0.05 is added to each frame; 20 percent of the words in each utterance are randomly selected and masked.

Baseline methods: In the following Comparison Studies section, we compare our model, A^3D^2, with DANN (Ganin et al., 2016), CDAN (Long et al., 2018), MADA (Pei et al., 2018), DALN (Chen et al., 2022a), PCL (Li et al., 2024), f-DD (Wang and Mao), LUHP (Zhang et al., 2024b) and DADA(Ren et al., 2024), which are introduced in the Related Works section.

Implementation details: For the acoustic modality, WavLM (Chen et al., 2022b) followed by a TextCNN is employed as the feature encoder. For the visual modality, APViT pretrained on the RAF-DB (Li et al., 2017) database is utilized for sequence feature extraction, and then a one-layer LSTM is utilized to encode the sequence feature. Bert-base (Devlin et al., 2018) and TextCNN are adopted for the lexical modality. The parameters in the last three layers of the pretrained models are set to be trainable, with all other parameters frozen. The dimension of the representations $\boldsymbol{Z}_m, \forall m \in [M]$, is 256. The Adam optimizer is used for model training with learning rate 1×10^{-3}, momentum coefficient $(0.9, 0.999)$ and batch size 48. The hyperparameter settings are $\beta = 1 \times 10^{-3}$, $\gamma_1 = 10$, $\gamma_2 = 20$. More details of the implementation can be found from the appendix and the code in the supplementary material. We use the weighted F1 score as model performance metric, which is obtained by averaging the results from three repeated experiments, conducted on four Nvidia A100 GPUs with memory of 40GB.

4.1 Comparison studies

The F1 score comparisons are reported in Table 1, where DT refers to direct transfer, meaning the model is trained with only the source domain data and directly tested on the target domain samples. In the table, the columns "A", "L", and "V" correspond to the results of experiments using acoustic, lexical, and visual as anchor, respectively, while the column "ave." shows the average result of the three columns. On the IEMOCAP dataset, A^3D^2 improves the average result over the best baseline approach by a substantial margin of over

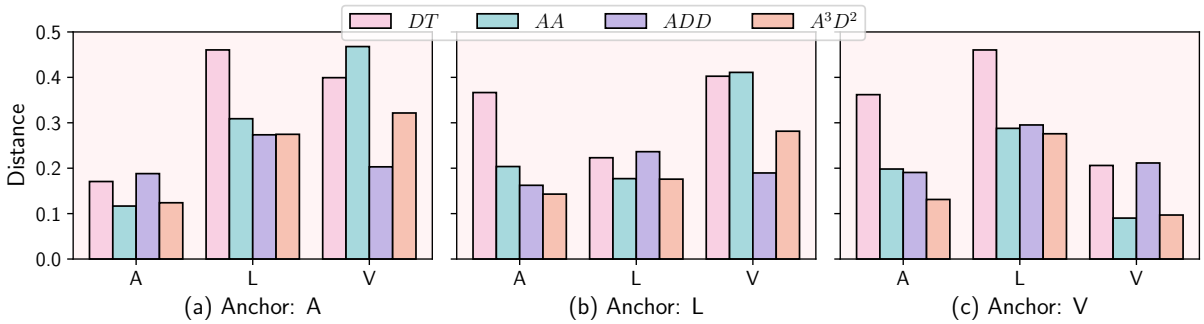


Figure 4: The distance between the source and target domains of different modalities.

AA	ADD	A	L	V	ave.
✗	✗	54.15	58.67	57.59	56.80
✓	✗	62.73	66.87	68.87	66.16
✗	✓	63.95	63.55	**69.48**	65.66
✓	✓	**65.15**	**70.84**	69.20	**68.40**

Table 2: Results of ablation studies on IEMOCAP.

2.5 percentage points. For any specific modality as an anchor, A^3D^2 surpasses other methods by at least 0.4 percentage points. On the MIntRec dataset, A^3D^2 generally outperforms other methods, with only one exception: DADA slightly exceeds A^3D^2 by 0.29 percentage points when L is the anchor. Note that while DALN demonstrates satisfactory performance on IEMOCAP, it falls short on MIntRec. These comparisons validate the superior performance of A^3D^2 over current competing approaches.

Figure 3 displays the t-SNE visualization of the representation distributions on the IEMOCAP dataset. The upper panel shows the results of DT, where the two domains exhibit significant mismatch, indicating that the model generalizes poorly on the target domain. With A^3D^2, the two domains show significantly improved category-level alignment, as illustrated in the lower panel of Figure 3.

4.2 Ablation studies

In this section, we present the ablation studies on the IEMOCAP dataset for the two primary components, i.e., adversarial alignment (AA) and anchor dragging drift (ADD). As shown in Table 2, on average, both AA and ADD individually improve the F1-score by at least 8 percentage points compared to the baseline, which does not employ any adaptation technique. With the joint contributions of AA and ADD, A^3D^2 achieves further improvement, reaching an average F1 score of 68.40. Regarding the specific cases where the modalities A, L, and V serve as anchors, A^3D^2 outperforms AA and ADD in the cases of A and L, and is only marginally weaker than ADD in the case of V.

In order to verify that AA, ADD, and A^3D^2 can reduce the gap between the source and target domains, we calculate the average category-wise distance between the two domains using their representations. Specifically, the center for class c and modality m is computed as: $\bar{\boldsymbol{z}}_{m,c} = \frac{1}{\sum_{n=1}^{N} 1_{y_{n,m}=c}} \sum_{n=1}^{N} \boldsymbol{z}_{n,m} \cdot 1_{y_{n,m}=c}$. Then, the distance between the source and target domains for modality m is defined as: $D_m := \frac{1}{C} \sum_{c=1}^{C} ||\bar{\boldsymbol{z}}^s_{m,c} - \bar{\boldsymbol{z}}^t_{m,c}||_2$. A large (small) distance indicates a large (resp. small) gap between two domains. Figure 4 exhibits the distance for different modalities resulting from all approaches, where the distance corresponding to DT represents the original gap between the source and target domains.

Figures 4(a), 4(b), and 4(c) show that the distance of the anchor is naturally smaller than that of the drifts. ADD results in a negligible increase in distance for the anchor compared with DT, since "anchor dragging drift" in the target domain causes the anchor to deviate from its original position. This small deviation is insignificant because A^3D^2, with the assistance of AA, can significantly reduce the gap of the anchor between the source and target domains. The reason can be that when AA aligns the drifts, it implicitly aligns the anchor as well.

As for the drifts, AA largely decreases the distance for A and L compared with DT, but slightly increases the distance for V. In comparison, it is interesting that although ADD is designed to directly bring the anchor and drift in the target domain closer, it indirectly reduces the distance of all drifts across the source and target domains. Via combining AA and ADD, A^3D^2 effectively diminishes the domain gap.

From the above in-depth analysis on the IEMOCAP dataset, we draw two conclusions: 1) for the anchor modality, AA dominates the distance reduction. 2) for the drifts, ADD consistently reduces the distance, while AA can lead to a small increase for V. When AA and ADD are employed collaboratively, the overall distance is reduced. Therefore, AA and ADD complement each other and both contribute to the success of A^3D^2.

5 Conclusions

In this paper, we investigate a practical and unique multimodal domain adaptation problem, where some modalities (i.e., anchors) remain stable, while others (i.e., drifts) undergo shifts. Building upon adversarial learning and optimal transport methods, we propose a bi-alignment strategy that performs drift-drift alignment across domains and anchor-drift alignment within the target domain. Extensive experimental results and a detailed analysis corroborate the effectiveness of the proposed approach.

6 Limitations

The main limitation of this work is the lack of theoretical analysis for the OT-based cross-modality alignment (i.e., ADD) in the context of MMDA, which will be investigated in our future work.

Acknowledgments

The work by Jun Sun is supported by the National Natural Science Foundation of China (Grant No. 62306289) and the National Key R&D Program of China (Grant No. 2024YFB4505602). The work by Lingfang Zeng is supported by the National Key R&D Program of China (2022YFB4500405), the Regional Innovation and Development Joint Fund of National Natural Science Foundation of China (U22A6001), and the Zhejiang provincial "Ten Thousand Talents Program" (2021R52007).

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Chao Chen, Zhihong Chen, Boyuan Jiang, and Xinyu Jin. 2019. Joint domain alignment and discriminative feature learning for unsupervised deep domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3296–3303.

Lin Chen, Huaian Chen, Zhixiang Wei, Xin Jin, Xiao Tan, Yi Jin, and Enhong Chen. 2022a. Reusing the task-specific classifier as a discriminator: Discriminator-free adversarial domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7181–7190.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022b. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Ching-Yao Chuang, Stefanie Jegelka, and David Alvarez-Melis. 2023. Infoot: Information maximizing optimal transport. In *International Conference on Machine Learning*, pages 6228–6242. PMLR.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. 2017. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. 2025. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *arXiv preprint arXiv:2501.18592*.

Zhipeng Du, Miaojing Shi, and Jiankang Deng. 2024. Boosting object detection with zero-shot day-night domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12666–12676.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. 2021. Unbalanced minibatch optimal transport; applications to domain adaptation. In *International Conference on Machine Learning*, pages 3186–3197. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

Zirun Guo, Tao Jin, and Zhou Zhao. 2024. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1726–1736.

Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. 2023. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11721–11732.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. 2023. How does information bottleneck help deep learning? In *International Conference on Machine Learning*. PMLR.

Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. 2021. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627.

Junjie Li, Yixin Zhang, Zilei Wang, Saihui Hou, Keyu Tu, and Man Zhang. 2024. Probabilistic contrastive learning for domain adaptation. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 1001–1009.

Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861.

Yang Liu, Zhipeng Zhou, and Baigui Sun. 2023. Cot: Unsupervised domain adaptation with clustering and optimal transport. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19998–20007.

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

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

Eduardo Fernandes Montesuma, Fred Maurice Ngolè Mboula, and Antoine Souloumiac. 2024. Optimal transport for domain adaptation through gaussian mixture models. *arXiv preprint arXiv:2403.13847*.

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*, pages 122–132.

Khai Nguyen, Dang Nguyen, Tung Pham, Nhat Ho, et al. 2022. Improving mini-batch optimal transport via partial transportation. In *International Conference on Machine Learning*, pages 16656–16690. PMLR.

Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Li Ren, Chen Chen, Liqiang Wang, and Kien Hua. 2024. Towards improved proxy-based deep metric learning via data-augmented domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14811–14819.

Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. 2019. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020.

Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023. Layer-wise fusion with modality independence modeling for multi-modal emotion recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 658–670.

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

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Yuan Zhong, Xiaokun Zhang, Yaqing Wang, Parminder Bhatia, Cao Xiao, and Fenglong Ma. 2024. Unity in diversity: Collaborative pre-training across multimodal medical sources. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3644–3656.

Ziqiao Wang and Yongyi Mao. On f-divergence principled domain adaptation: An improved framework. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuan Wu, Diana Inkpen, and Ahmed El-Roby. 2021. Conditional adversarial networks for multi-domain text classification. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 16–27.

Baochen Xiong, Xiaoshan Yang, Yaguang Song, Yaowei Wang, and Changsheng Xu. 2024. Modality-collaborative test-time adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26732–26741.

Nishant Yadav, Mahbubul Alam, Ahmed Farahat, Dipanjan Ghosh, Chetan Gupta, and Auroop R Ganguly. 2023. Cda: Contrastive-adversarial domain adaptation. *arXiv preprint arXiv:2301.03826*.

Yuehao Yin, Bin Zhu, Jingjing Chen, Lechao Cheng, and Yu-Gang Jiang. 2022. Mix-dann and dynamic-modal-distillation for video domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3224–3233.

Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded multimodal named entity recognition on social media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9141–9154.

Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central moment discrepancy (cmd) for domain-invariant representation learning. *arXiv preprint arXiv:1702.08811*.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. 2022. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1688–1697.

Xinxin Zhang, Jun Sun, Simin Hong, and Taihao Li. 2024a. Amanda: Adaptively modality-balanced domain adaptation for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 14448–14458.

Xinyu Zhang, Meng Kang, and Shuai Lü. 2024b. Low category uncertainty and high training potential instance learning for unsupervised domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16881–16889.

Yufeng Zhang, Jianxing Yu, Yanghui Rao, Libin Zheng, Qinliang Su, Huaijie Zhu, and Jian Yin. 2024c. Domain adaptation for subjective induction questions answering on products by adversarial disentangled learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9074–9089.

Han Zhao, Shanghang Zhang, Guanhang Wu, José MF Moura, Joao P Costeira, and Geoffrey J Gordon. 2018. Adversarial multiple source domain adaptation. *Advances in neural information processing systems*, 31.

Jinming Zhao, Ruichen Li, and Qin Jin. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2608–2618.

Jinjing Zhu, Yucheng Chen, and Lin Wang. 2024. Clip the divergence: Language-guided unsupervised domain adaptation. *arXiv preprint arXiv:2407.01842*.