Amanda: Adaptively Modality-Balanced Domain Adaptation for Multimodal Emotion Recognition

Anonymous ACL submission

Abstract

This paper investigates unsupervised multimodal domain adaptation for multimodal emotion recognition, which is a solution for data scarcity yet remains under studied. Due to the varying distribution discrepancies of different modalities between source and target domains, the primary challenge lies in how to balance the domain alignment across modalities to guarantee they are all well aligned. To achieve this, we first develop our model based on the information bottleneck theory to learn optimal representation for each modality independently. Then, we align the domains via matching the label distributions and the representations. In order to balance the representation alignment, we propose to minimize a surrogate of the alignment losses, which is equivalent to adaptively adjusting the weights of the modalities throughout training, thus achieving balanced domain alignment across modalities. Overall, the proposed approach features **A**daptively **m**odality-bal**an**ced **d**omain **a**daptation, dubbed as **Amanda**, for multimodal emotion recognition. Extensive empirical results on commonly used benchmark datasets demonstrate that Amanda significantly outperforms competing approaches. The code is submitted as supplementary material, and the extracted features of the datasets will be made publicly available upon the publication of the paper.

1 Introduction

Emotion recognition has gained increasing attention in recent years in a wide spectrum of applications, including emotional support (Tu et al., 2022), conversation system (Shi and Huang, 2023) and healthcare (Zanwar et al., 2023). Multimodal emotion recognition which takes advantage of heterogeneous and complementary signals, such as acoustic, visual, lexical information, has demonstrated superior performance to its unimodal counterpart (Zhu et al., 2022; Zhang and Li, 2023). Nevertheless, one of the notable drawbacks of multimodal learning is that collecting and annotating data of multiple modalities is much more expensive than one single modality (Lian et al., 2023). Thus, the importance of the ability of a model to transfer knowledge from annotated datasets to unannotated but related ones is manifested in the context of multimodal emotion recognition.

In this regard, unsupervised domain adaptation techniques are popular for promoting the generalization capability of a model from a labeled source domain to an unlabeled target domain. Domain adaptation typically fills the model's performance gap between the target and source domains via matching the data distributions of the two domains with sample-based, feature-based and inference-based approaches (Kouw and Loog, 2021). Accordingly, numerous schemes have been developed for various tasks in the fields of computer vision (Li et al., 2021; Liu et al., 2023) and natural language processing (Calderon et al., 2022; Dua et al., 2023). In contrast, domain adaptation in multmodal learning settings remains relatively less researched, not to mention in multimodal emotion recognition.

The previous literature on multimodal domain adaptation broadly falls into two categories, multiple visual modalities and general multiple modalities. The former tackles multimodal computer vision tasks, where different modalities correspond to RGB and optical flows (Munro and Damen, 2020), CT and MRI images (Kruse et al., 2021), or 2D image and 3D point cloud (Xing et al., 2023). In these scenarios, the modalities are similar and share the same environment, suggesting a close distribution gap between source and target domains of different modalities. Therefore, no specific effort is required to address modality differences when aligning the domains. The latter category focuses on more general multimodal domain adaptation approaches, applicable to text/image and video/audio applications. However, in these studies, the source and target domains of different modalities are aligned uniformly

without recognizing modality disparity (Qi et al., 2018), or they are not considered jointly, leading to the situation where some modalities are well aligned while others not (Yuan et al., 2022).

In multimodal emotion recognition tasks, the commonly utilized modalities—linguistic, visual and acoustic, exhibit high heterogeneity. Moreover, these modalities live in decoupled spaces, as opposed to the visual modalities mentioned above. Consequently, from the source to target domains, different modalities experience varying degree of distribution shift. For example, consider a shift in the working scene of a multimodal emotion recognition system from the day (source domain) to the night (target domain). In this scenario, the distribution of visual features noticeably shifts due to the variation of illumination conditions, while that of acoustic features remains relatively unchanged.

Henceforth, directly applying existing domain adaptation approaches to multimodal emotion recognition might result in an imbalanced alignment of different modalities. The model may then rely heavily on the well aligned modalities in the target domain and under-utilize others; in other words, well aligned modalities dominate others, causing the latter to be under-trained.

With the above analysis, in this paper, we advocate modality independence (Sun et al., 2023a; Qu et al., 2021) and align the source and target domains of different modalities, taking their varying distribution gaps into consideration. To be specific, we design our model based on the information bottleneck (IB) theory (Saxe et al., 2019; Kawaguchi et al., 2023), which enforces each modality to perform label prediction, thereby encouraging each to obtain its optimal representation independently. As for the domain alignment, we first introduce label distribution alignment under the practical assumption that the label distributions remain consistent across the source and target domains. We then employ correlation alignment (Sun et al., 2016; Sun and Saenko, 2016) to match the optimal representations in the two domains for each modality.

To balance the representation alignment, we minimize a surrogate of the alignment losses rather than minimizing a weighted sum of the losses with fixed weights. Via judiciously devising the surrogate function, minimizing it is tantamount to minimizing the weighted sum of the losses with the weights being adaptively tuned throughout the training progress. Concretely, the modalities with larger (resp. smaller) losses receive proportionally larger (resp. smaller) weights, which achieves dynamically balanced domain alignment across modalities.

In summary, our work features **A**daptively **m**odality-bal**an**ced **d**omain **a**daptation (abbreviated as **Amanda**) for multimodal emotion recognition. The contributions are primarily threefold.

- We develop a mulitmodal emotion recognition model which learns the representations of modalities independently and aligns the source and target domain via matching the representations and the labels.
- We propose a paradigm for alignment loss surrogate function design, which adaptively balances all modalities during training.
- Empirical results verify the effectiveness of the proposed method, and demonstrate that Amanda outperforms the compared schemes.

2 Related Works

2.1 Domain adaptation

There are an enormous number of prior works on domain adaptation, for which interested readers can refer to survey papers (Wang and Deng, 2018; Kouw and Loog, 2021; Yu et al., 2023) and references therein. We only cover the most relevant works, which can be classified into two branches, i.e., the adversarial learning methods and moment matching methods. Starting from the pioneering work DANN (Ganin et al., 2016), a vast amount of adversarial learning methods emerge. MDAN (Zhao et al., 2018) investigates domain adaptation with multiple source domains and devises two versions of optimization strategies. CDAN (Long et al., 2018), MADA (Pei et al., 2018) and CAN (Wu et al., 2021) introduce label prediction information as conditioning for domain alignment. DADA (Tang and Jia, 2020) integrates domain and category classifiers as a shared classifier to encourage a mutually inhibitory relation between domain and category predictions. CDA (Yadav et al., 2023) incorporates contrastive learning into domain adaptation to achieve class-level alignment.

As for the moment matching branch, maximum mean discrepancies (MMD) (Tzeng et al., 2014) and its variants, such as MK-MMD (Long et al., 2015), RTN (Long et al., 2016) are typical first order moment approaches which match the mean of the representations. Coral (Sun et al., 2016; Sun and Saenko, 2016) and JDDA (Chen et al., 2019) represent second moment approaches, matching the covariance of the representations.

2.2 Balanced multimodal learning

Another line of relevant studies that inspire this work are devoted to balancing the convergence of different modalities to prevent some modalities being overfitting while others being underfitting. The learning rates of different modalities are dynamically regulated via tracking the label prediction losses of all modalities in studies (Sun et al., 2021; Peng et al., 2022). Work (Wu et al., 2022) proposes a scheme to estimate the model's dependence on each modality, based on which an algorithm to balance the learning speeds of all modalities is introduced. In study (Wang et al., 2020), the overfitting behaviors of the modalities are evaluated, and accordingly, an optimal blending of gradients is computed for model updates. The relative advantage of each modality is defined during model training, with which a bi-level optimization problem is formulated to re-weight the loss terms of all modalities in work (Sun et al., 2023b). These approaches are effective yet involve a delicate heuristic design in the light of some observations during training.

In our work, we adopt the second moment matching method for representation alignment, thus focusing on balancing modalities and circumventing the difficulty in balancing the competing generative and discriminative components in adversarial-based methods. We propose a paradigm for alignment loss surrogate function design, enabling adaptive balancing of alignment losses across modalities without extra effort to tune the learning rates.

3 Method: Amanda

Prior to delving into our method, Amanda, we introduce the notations and assumptions below.

Notations: Suppose the multimodal training dataset contains N samples, each with M modalities. For ease of expression, let us define an auxiliary modality as a union of all modalities, and thus the total number of modalities is $M+1$. Let $[P]$ for any positive integer P denote the set $\{1, 2, \cdots, P\}$. The training samples are denoted by $(\{\boldsymbol{x}_{n,m}\}_{m\in[M]}, \{\boldsymbol{y}_{n,m}\}_{m\in[M+1]})$, where $n \in [N]$ indexes the samples, $\boldsymbol{x}_{n,m} \in \mathbb{R}^{d_m}$ represents the d_m-dimensional feature vector (the feature can also be vector sequence) of modality $m, \forall m \in [M]$, and $\boldsymbol{y}_{n,m}$ represents the label corresponding to modality $m, \forall m \in [M+1]$ (for datasets where all modalities share a common label, $\boldsymbol{y}_{n,1} = \boldsymbol{y}_{n,2} = \cdots = \boldsymbol{y}_{n,M+1}$ holds). Suppose the number of emotion categories is C; then the label $\boldsymbol{y}_{n,m}$ can be a one-hot vector or a scalar in [C], and we adopt either of these two forms when necessary in the rest of the paper. For the consistency of expression, we use $\boldsymbol{x}_{n,M+1} := [\boldsymbol{x}_{n,1}; \boldsymbol{x}_{n,2}; \cdots; \boldsymbol{x}_{n,M}]$ to collect all features of sample n.

Let $\boldsymbol{X}_m$ and $\boldsymbol{Y}_m$ represent general feature and label random variables for all $m \in [M+1]$, with $\boldsymbol{x}_{n,m}$ and $\boldsymbol{y}_{n,m}$ as their realizations. Let vector $\boldsymbol{Z}_m \in \mathbb{R}^d$, a map of $\boldsymbol{X}_m$, denote the representation of modality m, and $\boldsymbol{z}_{n,m}$ is a realization of $\boldsymbol{Z}_m$ (for brevity, we assume the representations of all modalities are d-dimensional vectors). We use superscript s and t to distinguish variables of source and target domains. For instance, $\boldsymbol{X}_m^s$ and $\boldsymbol{X}_m^t$ denote the features of modality m from source and target domains, respectively.

Assumptions: In this paper we consider unsupervised domain adaptation problem for multimodal learning, for which the following assumptions are satisfied: 1) the label target domain data is inaccessible; 2) the feature distributions shift with the domains, yet the label distributions remain unchanged, meaning that $p(\boldsymbol{X}_m^t) \neq p(\boldsymbol{X}_m^s)$ and $p(\boldsymbol{Y}_m^t) = p(\boldsymbol{Y}_m^s)$ hold for any $m \in [M+1]$, where $p(\cdot)$ represents the distribution of a random variable. The second assumption holds true when the domain changes the feature but is not a causal factor of the considered event. For example, although the illumination (the feature) of the vision system varies between the day and the night (different domains), one's emotion (the label) distribution remains relatively stable with the day and the night.

3.1 Model design

A. Overview of model design

Figure 1(a) visualizes the architecture of our model, Amanda, in an example with two modalities. As illustrated, to map the feature $\boldsymbol{X}_m$ to the representation $\boldsymbol{Z}_m$, we employ a deterministic feature encoder $f_m(\cdot; \boldsymbol{\theta}_m^f) : \mathbb{R}^{d_m} \to \mathbb{R}^d$ with model parameter $\boldsymbol{\theta}_m^f$, which means $\boldsymbol{Z}_m = f_m(\boldsymbol{X}_m; \boldsymbol{\theta}_m^f), \forall m \in [M]$. For different modalities, $f_m(\cdot)$ can take different forms; for instance, in our model framework, we utilize TextCNN for the acoustic and lexical modalities, and LSTM for the visual modality. Let $\boldsymbol{Z}_{M+1} := [\boldsymbol{Z}_1; \boldsymbol{Z}_2; \cdots; \boldsymbol{Z}_M]$ concatenate the representation of all modalities. For each modality $m \in [M+1]$, an MLP $g_m(\cdot; \boldsymbol{\theta}_m^g)$ is adopted to predict the label using the corresponding representation, that is, $\hat{\boldsymbol{Y}}_m = g_m(\boldsymbol{Z}_m; \boldsymbol{\theta}_m^g)$. The multimodal prediction $\hat{\boldsymbol{Y}}_{M+1}$ is admitted as the final

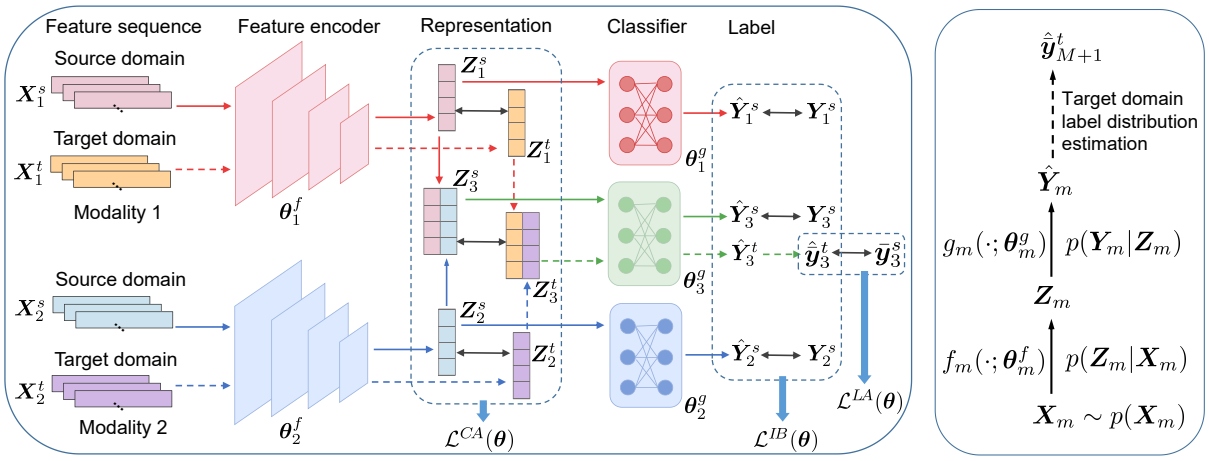


Figure 1: (a) Model architecture with 2 modalities as an example (multimodal representation $\boldsymbol{Z}_3$ is a concatenation of $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$; solid and dashed regular arrows represent the flows of source and target domains, respectively; double-headed arrows represent alignment or supervision signals, corresponding to the information bottleneck loss $\mathcal{L}^{IB}(\boldsymbol{\theta})$, label alignment loss $\mathcal{L}^{LA}(\boldsymbol{\theta})$ and correlation alignment loss $\mathcal{L}^{CA}(\boldsymbol{\theta})$). (b) Information flow of modality m.

predicted label. For ease of expression, we use $\boldsymbol{\theta} = \{\boldsymbol{\theta}^g_{M+1}\} \cup \{\boldsymbol{\theta}^f_m, \boldsymbol{\theta}^g_m\}_{m\in[M]}$ to collect all model parameters.

B. IB based representation learning

With the above model framework, the information chain follows $\boldsymbol{X}_m \to \boldsymbol{Z}_m \to \boldsymbol{Y}_m, \forall m \in [M+1]$, as is shown in Figure 1(b). A model with good generalization performance should be able to generate representation $\boldsymbol{Z}_m$ which maintains task relevant information and discards the rest in $\boldsymbol{X}_m$. To achieve this, the labeled source domain data is utilized to minimize the following information bottleneck (IB) loss:

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

where $I(\cdot,\cdot)$ represents the mutual information of two random variables, and γ is a coefficient balancing the two terms.

From the perspective of information theory, it is evident that minimizing $\mathcal{L}_{IB}(\boldsymbol{\theta})$ leads to a representation $\boldsymbol{Z}^s_m$ that retains minimal information from the original feature $\boldsymbol{X}^s_m$ while capturing the maximal information of the label $\boldsymbol{Y}^s_m$. Henceforth, $\boldsymbol{Z}^s_m$ is an optimal representation in the sense of information bottleneck theory (Saxe et al., 2019; Kawaguchi et al., 2023). Moreover, not only is the joint modality $M+1$ enforced to learn the task relevant information $I(\boldsymbol{Z}^s_{M+1}, \boldsymbol{Y}^s_{M+1})$, but each individual modality m, for all $m \in [M]$, is also required to maximize their corresponding $I(\boldsymbol{Z}^s_m, \boldsymbol{Y}^s_m)$ even if all modalities share a common label. This promotes modality independence and prevents some weak modalities from being 'lazy' and being dominated by strong modalities.

Next, we elaborate on how the two information terms in Eq. (1) are calculated.

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

where $H(\cdot)$ denotes entropy, and $H(\boldsymbol{Z}^s_m|\boldsymbol{X}^s_m) = 0$ since $\boldsymbol{Z}_m = f_m(\boldsymbol{X}_m; \boldsymbol{\theta}^f_m)$ is a deterministic function. Upon assuming that $p(\boldsymbol{Z}^s_m)$ follows Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}^s_m, \boldsymbol{\Sigma}^s_m)$ ($\boldsymbol{\mu}^s_m \in \mathbb{R}^d$, and $\boldsymbol{\Sigma}^s_m \in \mathbb{R}^{d\times d}$ is a diagonal matrix), we can estimate $\boldsymbol{\mu}^s_m$ and $\boldsymbol{\Sigma}^s_m$ with the representations $\boldsymbol{z}^s_{n,m}, n \in [N^s]$. The the entropy of $H(\boldsymbol{Z}^s_m)$ is

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

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

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

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

where we use the fact that $H(\boldsymbol{Y}^s_m) = H^s_{Y,m}$ is a constant independent from parameter $\boldsymbol{\theta}$.

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

$$\begin{aligned} \mathcal{L}^{IB}(\boldsymbol{\theta}) = \sum_{m=1}^{M+1} \Big[&\frac{\gamma}{2}\log|\boldsymbol{\Sigma}^s_m| \\ &- \frac{1}{N^s}\sum_{n\in[N^s]} \log p(\boldsymbol{y}^s_{n,m}|\boldsymbol{z}^s_{n,m}) \Big]. \end{aligned} \quad (5)$$

C. Label alignment

As assumption 2) states, the label distributions of the target and source domains remain the same. We capitalize on this assumption to exploit the unlabeled target data. For the target domain sample $\boldsymbol{X}_m^t$, we can obtain its label $\hat{\boldsymbol{Y}}_{M+1}^t = p(\boldsymbol{Y}_{M+1}^t|\boldsymbol{X}_{M+1}^t;\boldsymbol{\theta})$. Although no label can be used as supervision signal for each individual target domain sample, the label distributions of the target and source domains can be aligned. The label distribution of source domain can be immediately computed from the labels as following:

$$p(\boldsymbol{Y}_{M+1}^s) = \bar{\boldsymbol{y}}_{M+1}^s = \frac{1}{N^s}\sum_{n\in[N^s]} \boldsymbol{y}_{n,M+1}^s \quad (6)$$

The predicted label distribution of target domain is

$$\hat{\bar{\boldsymbol{y}}}_{M+1}^t = \frac{1}{N^t}\sum_{n\in[N^t]} \hat{\boldsymbol{y}}_{n,M+1}^t \quad (7)$$

Label alignment (LA) is achieved by minimizing the following cross entropy loss between the target and source label distributions:

$$\begin{aligned}\mathcal{L}^{LA}(\boldsymbol{\theta}) &= -\mathbb{E}_{\bar{\boldsymbol{y}}_{M+1}^s\sim p(\boldsymbol{Y}_{M+1}^s)}[\log\hat{\bar{\boldsymbol{y}}}_{M+1}^t] \\ &= \sum_{c\in[C]} -(\bar{\boldsymbol{y}}_{M+1}^s)_c \log(\hat{\bar{\boldsymbol{y}}}_{M+1}^t)_c, \end{aligned} \quad (8)$$

where $(\cdot)_c$ is the c-th element of the vector.

D. Modality-wise representation alignment

We align the optimal representations of different modalities across the target and source domains, following the idea of matching the distributions by aligning the second order statistics. In specific, we first calculate the variance of $\boldsymbol{Z}_m^s$ and $\boldsymbol{Z}_m^t$, and denote them as $\boldsymbol{C}_m^s$ and $\boldsymbol{C}_m^t, \forall m \in [M+1]$, respectively. Then, the representation is aligned by minimizing the following correlation alignment (CA) loss (Sun et al., 2016):

$$\mathcal{L}_m^{CA}(\boldsymbol{\theta}) = ||\boldsymbol{C}_m^t - \boldsymbol{C}_m^s||_F^2, \quad (9)$$

where $||\cdot||_F$ represents the Frobenius norm.

However, as mentioned above, directly applying correlation alignment to multimodal domain adaptation faces the difficulty in balancing the modalities, since the gaps between target and source distributions of different modalities vary. To this end, we propose to minimize a surrogate function of $\boldsymbol{\mathcal{L}}^{CA}(\boldsymbol{\theta}) := [\mathcal{L}_1^{CA}(\boldsymbol{\theta}), \mathcal{L}_2^{CA}(\boldsymbol{\theta}), \cdots, \mathcal{L}_{M+1}^{CA}(\boldsymbol{\theta})]$, $h(\cdot) : \mathbb{R}^{M+1} \rightarrow \mathbb{R}$. The goal of minimizing $h(\boldsymbol{\mathcal{L}}^{CA}(\boldsymbol{\theta}))$ is to dynamically balancing different modalities during the optimization procedure. The details of how to determine $h(\cdot)$ is postponed to the next subsection.

With the above model and loss functions, the overall model training loss follows:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}^{IB}(\boldsymbol{\theta}) + \alpha_1\mathcal{L}^{LA}(\boldsymbol{\theta}) + \alpha_2 h(\boldsymbol{\mathcal{L}}^{CA}(\boldsymbol{\theta})), \quad (10)$$

where α_1 and α_2 are the constant coefficients weighting the three loss terms which are also shown in Figure 1. In the sequel, we present our approach for the design of the surrogate function $h(\cdot)$.

3.2 Adaptive modality balancing

In this subsection, we develop a surrogate function — modality balanced alignment loss (MBAL) function $h(\boldsymbol{a}(\boldsymbol{\theta})), \forall \boldsymbol{a}(\boldsymbol{\theta}) = [a_1(\boldsymbol{\theta}), a_2(\boldsymbol{\theta}), \cdots, a_{M+1}(\boldsymbol{\theta})], a_m(\boldsymbol{\theta}) \geq 0, \forall m \in [M+1]$, such that minimizing $h(\boldsymbol{a}(\boldsymbol{\theta}))$ can adaptively balance the minimization of all elements of $\boldsymbol{a}(\boldsymbol{\theta})$. Note that here we use $\boldsymbol{a}(\boldsymbol{\theta})$ for brevity and generality, and substituting $\boldsymbol{a}(\boldsymbol{\theta})$ with $\boldsymbol{\mathcal{L}}^{CA}(\boldsymbol{\theta})$ in $h(\boldsymbol{a}(\boldsymbol{\theta}))$ directly gives the alignment loss term in Eq. (10).

A. A general design of the MBAL function

We first propose that $h(\boldsymbol{a}(\boldsymbol{\theta}))$ in general takes the following form:

$$h(\boldsymbol{a}(\boldsymbol{\theta})) = \phi^{-1}\Big(\sum_{m\in[M+1]} \phi(a_m(\boldsymbol{\theta}))\Big), \quad (11)$$

where $\phi(\cdot)$ is a convex and monotonically increasing function, and $\phi^{-1}(\cdot)$ denotes the inverse function of $\phi(\cdot)$.

Applying the chain rule of derivative, the gradient of $h(\boldsymbol{a}(\boldsymbol{\theta}))$ is derived:

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} h(\boldsymbol{a}(\boldsymbol{\theta})) &= \frac{\sum_{m=1}^{M+1}\phi'(a_m(\boldsymbol{\theta}))\cdot\nabla_{\boldsymbol{\theta}} a_m(\boldsymbol{\theta})}{\phi'\big(\phi^{-1}(\sum_{m=1}^{M+1}\phi(a_m(\boldsymbol{\theta})))\big)} \\ &= \sum_{m=1}^{M+1}\psi_m(\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\theta}} a_m(\boldsymbol{\theta}), \end{aligned} \quad (12)$$

where $\phi'(\cdot)$ is the derivative of function $\phi(\cdot)$, and $\psi_m(\boldsymbol{\theta})$ is defined as:

$$\psi_m(\boldsymbol{\theta}) = \frac{\phi'(a_m(\boldsymbol{\theta}))}{\phi'\big(\phi^{-1}(\sum_{m=1}^{M+1}\phi(a_m(\boldsymbol{\theta})))\big)}. \quad (13)$$

From Eq. (12), it is obvious that the gradient $\nabla_{\boldsymbol{\theta}} h(\boldsymbol{a}(\boldsymbol{\theta}))$ corresponds to the weighted sum of $\nabla_{\boldsymbol{\theta}} a_m(\boldsymbol{\theta}), m \in [M+1]$ with weight coefficient $\psi_m(\boldsymbol{\theta})$. For the brevity of expression, we drop the variable $\boldsymbol{\theta}$ when no ambiguity occurs.

Next, we analyze the properties of ψ_m with $\phi(\cdot)$ elaborated as broadly used convex functions.

B. Two families of MBAL functions

We will show that when $\phi(\cdot)$ takes the form of power and exponential functions, the corresponding surrogate function $h(\boldsymbol{a})$ is consolidated as norm

and log-exp functions, respectively. The weights $\psi_m, m \in [M+1]$ are then properly bounded and positively correlated to $a_m, \forall a_m \geq 0$. This implies that with properly chosen learning rate, the convergence of the learning can be guaranteed, and meanwhile larger losses enjoy larger weights.

Norm functions For any $p \geq 1$, choosing $\phi(a) = a^p$ immediately gives that $h(\boldsymbol{a}) = ||\boldsymbol{a}||_p := (\sum_{m=1}^{M+1} a_m^p)^{1/p}$, which means $h(\boldsymbol{a})$ is the p-norm of $\boldsymbol{a}$. Then, ψ_m can be attained as:

$$\psi_m = \frac{a_m^{p-1}}{\left(\sum_{m=1}^{M+1} a_m^p\right)^{\frac{p-1}{p}}}, \forall m \in [M+1]. \quad (14)$$

Three cases come in order based on the value of p.

1) $p = 1$: $h(\boldsymbol{a})$ is a direct summation of a_m, and $\psi_m = 1, m \in [M+1]$ hold. This case corresponds to the imbalanced version of Amanda.

2) $1 < p < +\infty$: Eqs. (14) and (12) indicate that the gradient $\nabla_{\boldsymbol{\theta}} a_m(\boldsymbol{\theta})$ associated with larger a_m is highlighted with larger weight ψ_m. This implies that during training, the equivalent alignment loss weights of different modalities is adaptively regulated according to the corresponding losses, which pays more attention to larger losses.

3) $p = +\infty$: Eq. (14) reduces to $\psi_m = 1$, if $m = \text{argmax}_{m \in [M+1]} a_m$; otherwise, $\psi_m = 0$. Consequently, only the largest alignment loss among all modalities counts during training in terms of the gradient in Eq. (12).

Log-exp functions For any $t > 0$, choosing $\phi(a) = \exp(ta)$ leads to log-exp function: $h(\boldsymbol{a}) = \frac{1}{t}\ln(\sum_{m=1}^{M+1} \exp(ta_m))$. The weight ψ_m writes as:

$$\psi_m = \frac{\exp(ta_m)}{\sum_{m=1}^{M} \exp(ta_m)}, \forall m \in [M+1]. \quad (15)$$

Similarly, two cases follows:

1) $0 < t < +\infty$: Similar to the analysis of case 2) in the above norm function part, conclusion can be drawn by combining Eq. (12) and Eq. (14) that gradient-based training algorithms will "take more care of" the larger alignment losses.

2) $t = +\infty$: This case is exactly the same as case 3 in the above norm function part.

C. Theoretical properties and insights

Now we present theoretical properties of the weight $\boldsymbol{\psi} := [\psi_1, \psi_2, \cdots, \psi_{M+1}]$ and MBAL function $h(\boldsymbol{a})$.

Lemma 1. *The norm of the weight $\boldsymbol{\psi}$ satisfies ($p \geq 1$, and $1/p + 1/q = 1$):*

$$||\boldsymbol{\psi}||_q = 1, \text{if } h(\boldsymbol{a}) = ||\boldsymbol{a}||_p; \quad (16a)$$

$$||\boldsymbol{\psi}||_1 = 1, \text{if } h(\boldsymbol{a}) = \frac{1}{t}\ln(\sum_{m=1}^{M+1} \exp(ta_m). \quad (16b)$$

Eqs. (16a) and (16b) can be verified via calculating the q-norm and 1-norm of $\boldsymbol{\psi}$ using ψ_m in Eqs. (14) and (15), respectively.

Theorem 1. *The MBAL function $h(\boldsymbol{a})$ is an upper bound of the weighted sum of a_m with weights $\psi_m, m \in [M+1]$, which translates to the following inequalities:*

$$\sum_{m=1}^{M+1} \psi_m a_m \leq ||\boldsymbol{a}||_p = h(\boldsymbol{a}); \quad (17a)$$

$$\sum_{m=1}^{M+1} \psi_m a_m \leq \frac{1}{t}\ln(\sum_{m=1}^{M+1} \exp(ta_m)) = h(\boldsymbol{a}). \quad (17b)$$

Proof. For any $p \geq 1$, and $1/p + 1/q = 1$, the inequality (i.e., Eq. (17a)) below follows from Hölder's inequality and Eq. (16a).

$$\sum_{m=1}^{M+1} \psi_m a_m = \boldsymbol{\psi}^T \boldsymbol{a} \leq ||\boldsymbol{\psi}||_q \cdot ||\boldsymbol{a}||_p = ||\boldsymbol{a}||_p.$$

Since $\ln(\cdot)$ is a concave function, the following inequality (i.e., Eq. (17b)) is a result of Jensen's inequality and Eq. (16b).

$$\begin{aligned}\frac{1}{t}\ln(\sum_{m=1}^{M+1} \psi_m \exp(ta_m)) &\geq \frac{1}{t}\sum_{m=1}^{M+1} \psi_m \ln(\exp(ta_m)) \\ &= \sum_{m=1}^{M+1} \psi_m a_m,\end{aligned}$$

which finishes the proof. □

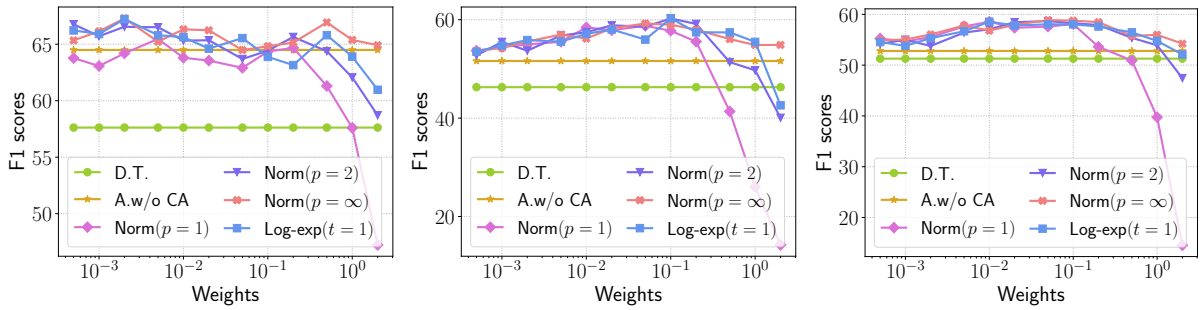
To sum up, we propose a paradigm for the alignment loss surrogate function design, under which two families of surrogate functions, norm functions and log-exp functions are analyzed. Theoretical results show that with the developed approach, the representation alignment losses of different modalities are adaptively balanced during training using gradient-based algorithms. Furthermore, minimizing the surrogate function boils down to minimizing the upper bound of the weighted sum of the alignment losses, where the bounded weights always correlate positively to the losses in the training progress.

4 Numerical Results

Benchmark datasets: We assess our method on four widely used benchmark multimodal emotion recognition datasets, IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), CMU-MOSEI (Zadeh et al., 2018), and MSP-IMPROV (Busso et al., 2016), which all contain acoustic, visual and lexical modalities. IEMOCAP and MSP-IMPROV

Method	IE.→MS.	IE.→MO.	ME.→IE.	ME.→MS.	MO.→IE.	MO.→MS.	MS.→IE.
D.T.	57.62	33.39	51.28	47.23	46.29	48.64	59.95
DANN	58.83	36.70	52.43	49.36	50.62	45.73	61.46
CDAN	60.57	37.50	55.84	49.28	51.01	46.86	63.56
CDAN+E	61.26	37.31	55.04	49.94	51.01	49.33	63.56
MADA	62.83	36.76	54.62	49.91	50.98	46.88	63.73
A-N$(p=1)$	64.35	38.31	58.09	53.44	57.75	52.31	62.67
A-N$(p=2)$	64.43	39.10	58.27	57.46	60.25	53.99	63.61
A-N$(p=\infty)$	64.82	38.38	58.77	54.46	58.98	54.39	64.30
A-L$(t=1)$	64.33	38.68	57.52	54.44	60.00	55.24	64.05

Table 1: F1 scores of the compared approaches. Abbreviations: D.T.: Direct transfer, A-N: Amanda with norm surrogate function, A-L: Amanda with log-exp surrogate function, IE.: IEMOCAP, MS.: MSP-IMPROV, ME.: MELD, MO.: CMU-MOSEI; the arrow "→" means from source to target domains.



(a) IEMOCAP → MSP (b) MOSEI → IEMOCAP (c) MELD → IEMOCAP

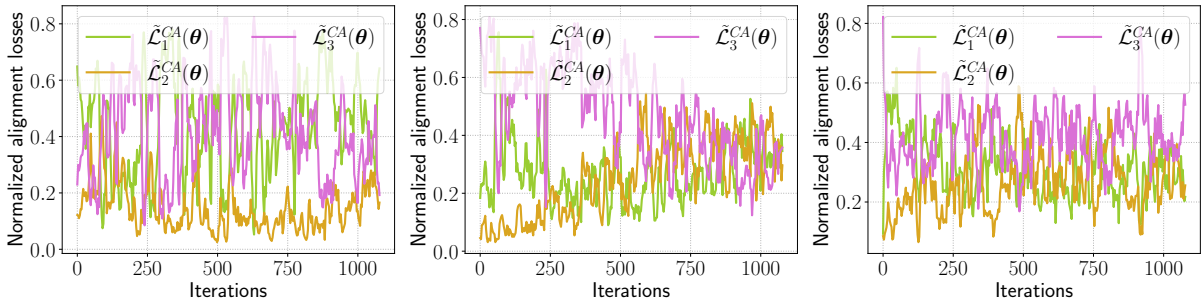
Figure 2: F1 scores v.s. varying weight (α_2) of the surrogate functions. Weights {0.0005, 0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, 2} are tested, and the x-axis is with log scale.

are composed of dyadic conversations collected in the laboratory setting, and the latter is of higher recording quality. CMU-MOSEI gathers monologue videos from more than 1000 speakers on YouTube over various topics. MELD consists of fragments from the TV series "Friends", which contains multi-party conversations with over two participants. These datasets are collected from different scenarios and exhibit different characteristics, and hence represent different domains. Following work (Zhao et al., 2021), we select samples in the four classes— neutral, happy, sad and angry, to construct datasets for our experiments.

Feature extraction: For the visual modality, we first sample each video uniformly to obtain 64 frames. Then, the frames are processed with S3FD(Zhang et al., 2017) to attain the speaker's faces which are then fed into vision model APViT (Xue et al., 2022) pretrained with dataset RAF-DB(Li et al., 2017), resulting in 64×768 sequential feature. BERT-base(Devlin et al., 2018) and Wav2Vec2(Baevski et al., 2020) are employed to extract lexical and acoustic features, respectively. To retain the feature of different levels, the outputs from the 1st, 7th, and 12th transformer blocks are concatenated as the final feature. The generated feature sequences are of dimension 2304, and their lengths are determined by the lengths of the text and audio, respectively.

Baseline models: We compare our model, Amanda, with DANN, CDAN, MADA and CDAN-E, of which the first three are introduced in the related works section, and CDAN+E is an extension of CDAN with the incorporation of entropy-aware reweighting for the domain discrimination loss.

Implementation details: The multimodal emotion recognition model involves three modalities, in which we employ one-layer LSTM for visual modality, and TextCNN for acoustic and lexical modalities as study (Zhao et al., 2021). The dimension of the representations is chosen as 128. We adopt optimizer Adam with learning rate 5×10^{-4}, momentum coefficient $(0.9, 0.999)$ and batch size 128 for model training. The parameter settings are $\gamma = 5 \times 10^{-4}$, $\alpha_1 = 0.08$; and $\alpha_2 = 0.1$ is selected for the comparison studies, and we will investigate how α_2 impacts the model performance in the ablation studies. More details of the implementation can be found from the code in the supplementary material. Throughout this section, we use



(a) $p = 1$ (b) $p = 2$ (c) $p = +\infty$

Figure 3: Normalized alignment losses of different modalities during training with A-N (IEMOCAP → MSP).

weighted F1 score as model performance metric. The reported F1 scores are obtained by averaging results from 3 repeated experiments, conducted on 2 Nvidia A100 GPUs with 40GB memory.

4.1 Comparison studies

We denote Amanda with 1-norm, 2-norm and ∞-norm, log-exp($t = 1$) surrogate functions by A-N($p = 1$), A-N($p = 2$), A-N($p = \infty$) and A-L($t = 1$), respectively. Table 1 reports the F1 scores of the baseline models and different versions of Amanda. In this table, direct transfer means the target domain data is not used for training, and directly be tested with the model trained on source domain. The results indicate that Amanda with p-norm ($p > 1$) log-exp ($t > 0$) surrogate functions performs on par with the baseline models on dataset MSP → IEMOCAP; and for all other datasets, it achieves substantial improvement. Due to the space limitation, more results (A-N($p = 3$) and A-L($t = 0.5$)) are included in the appendix.

4.2 Ablation studies

In this part, we conduct ablative studies on the two critical designs in Amanda, the label alignment and balanced representation alignment. Figure 2 illustrates the model performance with varying weight (α_2) of the surrogate functions. Comparing Amanda without correlation alignment (A. w/o CA) to direct transfer (D.T.), it is clear that label alignment enhances the knowledge transferring capability of the model significantly. For weight $\alpha_2 \leq 0.2$, Amanda with p-norm ($p > 1$) and log-exp surrogate functions can further improve the model performance over Amanda without CA (A. w/o CA). Particularly, when weight α_2 grows larger than 0.2, the balanced versions of Amanda, A-N($p = 2$), A-N($p = \infty$) and A-L($t = 1$), experience less performance drop compared to its imbalanced counterpart, A-N($p = 1$). The above results corroborate that both the label alignment and the adaptive domain alignment contribute to the success of Amanda.

In order to demonstrate that the proposed p-norm surrogate functions ($p > 1$) are able to balance the domain alignment of different modalities, we show the normalized alignment losses ($\tilde{\mathcal{L}}_m^{CA}(\boldsymbol{\theta}) := \mathcal{L}_m^{CA}(\boldsymbol{\theta}) / \sum_{m=1}^{M} \mathcal{L}_m^{CA}(\boldsymbol{\theta})$) in Figure 3, where the target and source datasets are IEMOCAP and MSP, respectively. Consistent with our analysis in section 3.2, the losses are not balanced with $p = 1$, and hence the losses exhibit large discrepancy among modalities throughout the training, as illustrated in Figure 3(a). In contrast, with $p = 2$ as shown in Figure 3(b), the losses are adaptively balanced, leading to closer gaps among modalities (when the three normalized losses are all 1/3, perfect balance is achieved). Figure 3(c) displays the case of $p = +\infty$, where the losses are also more balanced than that of the case $p = 1$. Due to the space limitation, we show the unweighted losses and the losses corresponding to log-exp surrogate functions in the appendix. These results validate that the proposed surrogate functions succeed in balancing the domain alignment of different modalities.

5 Conclusions

In this work, we devise a multimodal domain adaptation approach for multimodal emotion recognition. In order to close the gap between the target and source domains, we propose to match the label distributions of the two domains and to align the optimal representations for different modalities. Towards the objective of balancing the representation alignment, a general alignment loss surrogate function design paradigm is developed. Furthermore, we present the theoretical analysis of two families of surrogate functions which achieve adaptively modality-balanced domain adaptation. The effectiveness of the proposed approach is corroborated by extensive comparison and ablation studies.

6 Limitations

In light of the future work, the limitations of the present work are mainly twofold. 1) Although our method is applicable to more general multimodal supervised learning problems, we only validate it on emotion recognition tasks. 2) We have not established the theoretical upper bound of the target domain risk for the proposed approach.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

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.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. Docogen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7727–7746.

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.

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*.

Dheeru Dua, Emma Strubell, Sameer Singh, and Pat Verga. 2023. To adapt or to annotate: Challenges and interventions for domain adaptation in open-domain question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14429–14446, Toronto, Canada. Association for Computational Linguistics.

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.

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

Wouter M Kouw and Marco Loog. 2021. A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(03):766–785.

Christian N Kruse, Lasse Hansen, and Mattias P Heinrich. 2021. Multi-modal unsupervised domain adaptation for deformable registration based on maximum classifier discrepancy. In *Bildverarbeitung für die Medizin 2021: Proceedings, German Workshop on Medical Image Computing, Regensburg, March 7-9, 2021*, pages 192–197. Springer.

Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, and Han Zhao. 2021. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1104–1113.

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.

Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. 2023. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10):1440.

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.

Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2016. Unsupervised domain adaptation with residual transfer networks. *Advances in neural information processing systems*, 29.

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.

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.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

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

Shuhui Qu, Yan Kang, and Janghwan Lee. 2021. Efficient multi-modal fusion with diversity analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2663–2670.

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.

Tao Shi and Shao-Lun Huang. 2023. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766.

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.

Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer.

Jun Sun, Shoukang Han, Yu-Ping Ruan, Xiaoning Zhang, Shu-Kai Zheng, Yulong Liu, Yuxin Huang, and Taihao Li. 2023a. 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.

Jun Sun, Xinxin Zhang, Shoukang Han, Yu-ping Ruan, and Taihao Li. 2023b. Redcore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. *arXiv preprint arXiv:2312.10386*.

Ya Sun, Sijie Mai, and Haifeng Hu. 2021. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28:1650–1654.

Hui Tang and Kui Jia. 2020. Discriminative adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5940–5947.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.

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

Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705.

Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR.

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.

Bowei Xing, Xianghua Ying, Ruibin Wang, Jinfa Yang, and Taiyan Chen. 2023. Cross-modal contrastive learning for domain adaptation in 3d semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2974–2982.

Fanglei Xue, Qiangchang Wang, Zichang Tan, Zhongsong Ma, and Guodong Guo. 2022. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*.

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*.

Zhiqi Yu, Jingjing Li, Zhekai Du, Lei Zhu, and Heng Tao Shen. 2023. A comprehensive survey on source-free domain adaptation. *arXiv preprint arXiv:2302.11803*.

Junkun Yuan, Xu Ma, Defang Chen, Kun Kuang, Fei Wu, and Lanfen Lin. 2022. Label-efficient domain generalization via collaborative exploration and generalization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2361–2370.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Sourabh Zanwar, Xiaofei Li, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2023. What to fuse and how to fuse: Exploring emotion and personality fusion strategies for explainable mental disorder detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8926–8940.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. 2017. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201.

Xiaoheng Zhang and Yang Li. 2023. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13099–13110.

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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618.

Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, and Xiao Xiao. 2022. Multimodal emotion classification with multi-level semantic reasoning network. *IEEE Transactions on Multimedia*.

Appendix

The numbers of the tables and figures in the appendix follow those in the paper.

A Supervised learning baselines of the four datasets

Table 2 shows the baselines of the four datasets, which is obtained with standard supervised learning on each dataset itself.

B Alignment losses of different modalities

Figures 4 and 5 illustrate the alignment losses on dataset IEMOCAP → MSP, with norm and log-exp surrogate functions, respectively. Figure 6 shows the losses on dataset MOSEI → IEMOCAP. It can be concluded from these results that the designed surrogate functions indeed balance the domain alignment losses of different modalities.

Datasets	F1 score	Accuracy
IEMOCAP	79.72	79.95
MSP	79.68	79.87
MELD	55.84	56.39
MOSEI	54.89	55.58

Table 2: Supervised learning baselines of the four datasets.

C More results of the comparison studies

Table 3 is an extended version of Table 1, additionally including the results of Amanda with 3-norm and log-exp(t=0.5) surrogate functions.

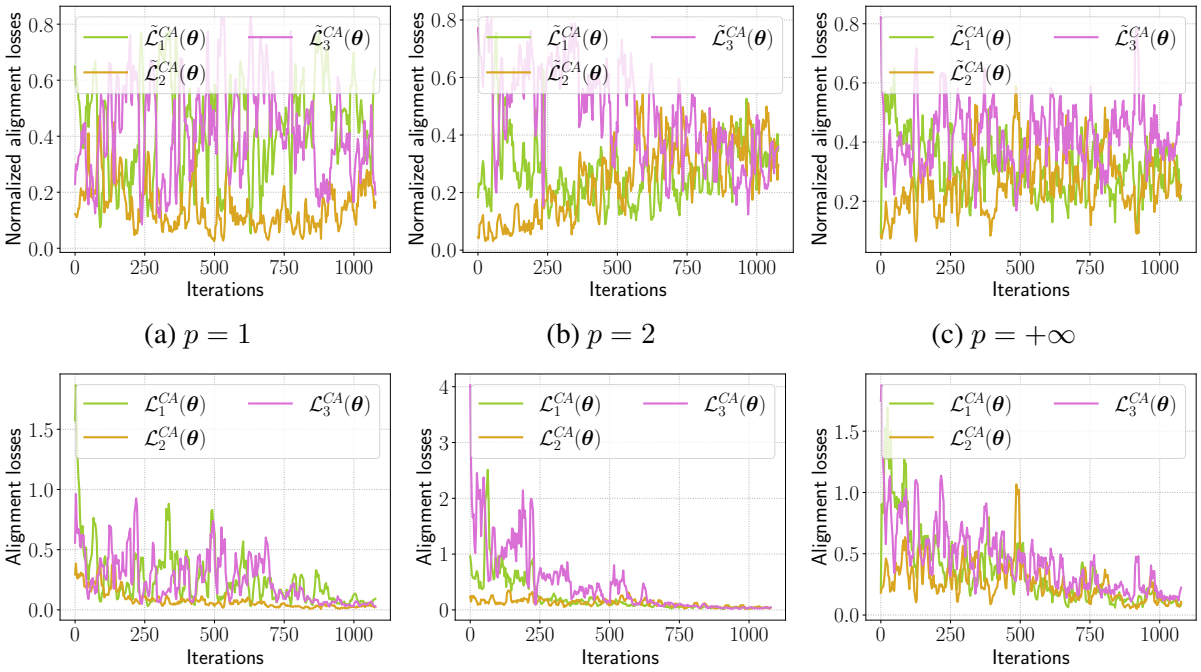
D Model performance with varying weights of the surrogate function

Tables 4, 5 and 6 report the model performance with varying weights of the surrogate function, which provided the details of Figure 2.

E Statistics of the datasets & the labels of the CMU-MOSEI dataset

Table 7 reports the detailed numbers of samples in each emotion category for the used four datasets, CMU-MOSEI, MELD, IEMOCAP and MSP-IMPROV.

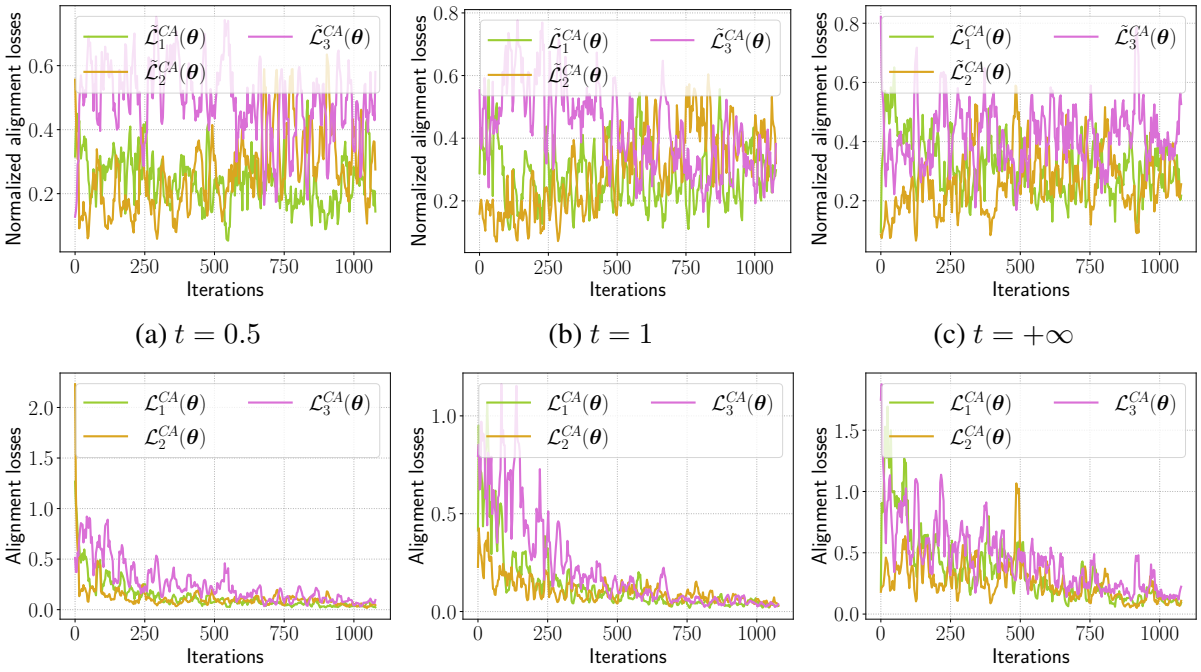
The original CMU-MOSEI dataset is annotated with a sentiment score and six emotion scores for emotion categories {happy, sad, angry, fear, disgusted, surprised}, which indicate the intensity of the sentiment and emotions, respectively. We categorize samples with sentiment score 0 and all emotion scores 0, to be neutral. For samples with a unique highest emotion score, the corresponding emotion label is assigned. We discard samples with multiple highest emotion scores to guarantee all selected ones are with a distinguishable emotion.



(a) $p = 1$ (b) $p = 2$ (c) $p = +\infty$

(d) $p = 1$ (e) $p = 2$ (f) $p = +\infty$

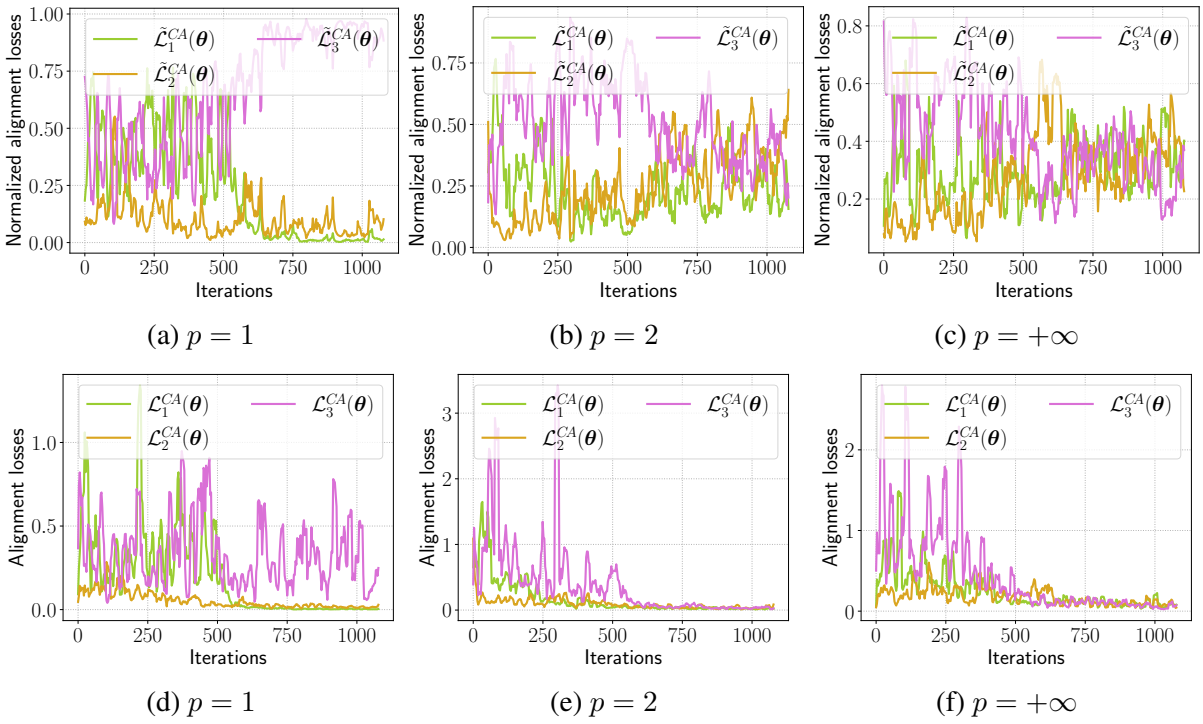
Figure 4: Alignment losses of different modalities during training by Amanda with norm surrogate functions (IEMOCAP $\rightarrow$ MSP). The upper and lower panels correspond to the normalized and unnormalized losses, respectively.



(a) $t = 0.5$ (b) $t = 1$ (c) $t = +\infty$

(d) $t = 0.5$ (e) $t = 1$ (f) $t = +\infty$

Figure 5: Alignment losses of different modalities during training by Amanda with log-exp surrogate functions (IEMOCAP $\rightarrow$ MSP). The upper and lower panels correspond to the normalized and unnormalized losses, respectively.



(a) $p = 1$ (b) $p = 2$ (c) $p = +\infty$

(d) $p = 1$ (e) $p = 2$ (f) $p = +\infty$

Figure 6: Alignment losses of different modalities during training by Amanda with norm surrogate functions (MOSEI → IEMOCAP). The upper and lower panels correspond to the normalized and unnormalized losses, respectively.

Method	IE.→MS.	IE.→MO.	ME.→IE.	ME.→MS.	MO.→IE.	MO.→MS.	MS.→IE.
D.T.	57.62	33.39	51.28	47.23	46.29	48.64	59.95
DANN	58.83	36.70	52.43	49.36	50.62	45.73	61.46
CDAN	60.57	37.50	55.84	49.28	51.01	46.86	63.56
CDAN+E	61.26	37.31	55.04	49.94	51.01	49.33	63.56
MADA	62.83	36.76	54.62	49.91	50.98	46.88	63.73
A-N($p = 1$)	64.35	38.31	58.09	53.44	57.75	52.31	62.67
A-N($p = 2$)	64.43	39.10	58.27	57.46	60.25	53.99	63.61
A-N($p = 3$)	64.98	39.10	58.82	54.82	59.23	56.28	63.80
A-N($p = \infty$)	64.82	38.38	58.77	54.46	58.98	54.39	64.30
A-L($t = 0.5$)	65.11	37.21	57.90	54.26	58.56	54.90	63.89
A-L($t = 1$)	64.33	38.68	57.52	54.44	60.00	55.24	64.05

Table 3: F1 scores of the compared approaches. Abbreviations: D.T.: Direct transfer, A-N: Amanda with norm surrogate function, A-L: Amanda with log-exp surrogate function, IE.: IEMOCAP, MS.: MSP-IMPROV, ME.: MELD, MO.: CMU-MOSEI.

Weight (α_2)	**0.0005**	**0.001**	**0.002**	**0.005**	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.5**	**1.0**	**2.0**
D.T.	57.62	57.62	57.62	57.62	57.62	57.62	57.62	57.62	57.62	57.62	57.62	57.62
A. w/o CA	64.49	64.49	64.49	64.49	64.49	64.49	64.49	64.49	64.49	64.49	64.49	64.49
A-N($p = 1$)	63.77	63.07	64.19	65.47	63.80	63.55	62.91	64.35	64.68	61.30	57.60	47.24
A-N($p = 2$)	66.77	65.71	66.54	66.5	65.32	65.35	63.69	64.43	65.64	64.38	62.06	58.71
A-N($p = 3$)	65.71	66.62	67.42	66.09	65.98	65.10	64.61	64.98	64.72	65.37	62.87	59.43
A-N($p = \infty$)	65.34	66.14	67.29	65.22	66.32	66.23	64.48	64.82	65.14	66.93	65.38	64.91
A-L($t = 0.5$)	65.67	66.75	66.51	66.45	66.65	66.12	66.56	65.11	64.47	64.99	64.21	64.24
A-L($t = 1$)	66.21	65.73	66.46	65.63	65.60	64.70	64.74	64.33	64.12	66.06	64.36	61.15

Table 4: F1 scores with varying weight (α_2) of the surrogate functions (IEMOCAP→ MSP).

Weight (α_2)	**0.0005**	**0.001**	**0.002**	**0.005**	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.5**	**1.0**	**2.0**
D.T.	46.29	46.29	46.29	46.29	46.29	46.29	46.29	46.29	46.29	46.29	46.29	46.29
A. w/o CA	51.61	51.61	51.61	51.61	51.61	51.61	51.61	51.61	51.61	51.61	51.61	51.61
A-N($p = 1$)	53.71	54.49	54.96	55.52	58.44	57.87	58.76	57.75	55.55	41.36	26.03	14.23
A-N($p = 2$)	52.75	55.48	53.70	56.78	57.45	58.86	58.57	60.25	59.14	51.38	49.72	40.06
A-N($p = 3$)	53.37	54.71	55.74	57.88	56.30	58.14	58.27	59.23	57.57	52.19	49.10	44.74
A-N($p = \infty$)	53.62	54.27	55.57	56.98	56.17	58.25	59.23	58.98	57.97	56.06	54.86	54.88
A-L($t = 0.5$)	53.57	53.67	54.81	54.70	54.92	55.81	57.11	58.56	58.55	57.10	56.27	54.72
A-L($t = 1$)	54.01	53.88	55.59	55.49	57.38	57.44	56.35	60.00	58.43	56.81	57.02	41.10

Table 5: F1 scores with varying weight (α_2) of the surrogate functions (MOSEI→ IEMOCAP).

Weight (α_2)	**0.0005**	**0.001**	**0.002**	**0.005**	**0.01**	**0.02**	**0.05**	**0.1**	**0.2**	**0.5**	**1.0**	**2.0**
D.T.	51.28	51.28	51.28	51.28	51.28	51.28	51.28	51.28	51.28	51.28	51.28	51.28
A. w/o CA	52.79	52.79	52.79	52.79	52.79	52.79	52.79	52.79	52.79	52.79	52.79	52.79
A-N($p = 1$)	55.24	54.64	55.53	57.62	58.54	57.36	57.59	58.09	53.61	50.91	39.75	14.52
A-N($p = 2$)	54.21	55.00	53.77	56.41	57.06	58.45	58.67	58.27	57.90	55.37	53.91	47.44
A-N($p = 3$)	54.01	55.40	55.64	58.11	58.49	57.92	59.50	58.82	56.44	56.93	55.09	49.40
A-N($p = \infty$)	54.90	55.05	56.04	57.80	56.80	57.94	58.87	58.77	58.47	55.73	55.98	54.23
A-L($t = 0.5$)	53.43	54.69	54.97	54.70	55.81	57.19	57.74	57.90	59.39	57.95	58.23	53.55
A-L($t = 1$)	54.24	53.83	55.50	55.89	58.02	58.12	58.58	57.52	57.67	57.38	53.29	52.61

Table 6: F1 scores with varying weight (α_2) of the surrogate functions (MELD→ IEMOCAP).

Emotion	**CMU-MOSEI**				**MELD**				**IEMOCAP**				**MSP-IMPROV**			
	train	val	test	sum	train	val	test	sum	train	val	test	sum	train	val	test	sum
Neutral	1128	136	338	1602	1021	109	270	1400	1221	145	333	1699	830	114	256	1200
Happy	1119	137	346	1602	956	119	325	1400	1119	115	351	1585	846	88	266	1200
Sad	780	74	216	1070	650	93	189	932	751	86	238	1075	587	62	151	800
Angry	758	98	214	1070	676	67	189	932	777	109	216	1102	564	68	160	792

Table 7: Statistics of the datasets.