Information Fusion 117 (2025) 102862

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Multi-source multi-modal domain adaptation

Sicheng Zhao [a,*,1], Jing Jiang [b,1], Wenbo Tang [c,d], Jiankun Zhu [b], Hui Chen [a], Pengfei Xu [c,d], Björn W. Schuller [e], Jianhua Tao [a,f], Hongxun Yao [b], Guiguang Ding [a,g]

[a] BNRist, Tsinghua University, Beijing, 100084, China
[b] Faculty of Computing, Harbin Institute of Technology, Harbin, 150001, China
[c] Didi Chuxing, 101300, Beijing, China
[d] NavInfo, 100028, Beijing, China
[e] Department of Computing, Imperial College London, London, SW7 2AZ, UK
[f] Department of Automation, Tsinghua University, Beijing, 100084, China
[g] School of Software, Tsinghua University, Beijing, 100084, China

ARTICLE INFO

Keywords:
Domain adaptation (DA)
Multi-modal DA
Multi-source DA
Contrastive learning
Adversarial learning
Sample selection

ABSTRACT

Learning from multiple modalities has recently attracted increasing attention in many tasks. However, deep learning-based multi-modal learning cannot guarantee good generalization to another target domain, because of the presence of domain shift. Multi-modal domain adaptation (MMDA) addresses this issue by learning a transferable model with alignment across domains. However, existing MMDA methods only focus on the single-source scenario with just one labeled source domain. When labeled data are collected from multiple sources with different distributions, the naive application of these single-source MMDA methods will result in sub-optimal performance without considering the domain shift among different sources. In this paper, we propose to study multi-source multi-modal domain adaptation (MSMMDA). There are two major challenges in this task: modal gaps between multiple modalities (*e.g.*, mismatched text-image pairs) and domain gaps between multiple domains (*e.g.*, differences in style). Therefore, we propose a novel framework, termed Multi-source Multi-modal Contrastive Adversarial Network (M2CAN), to perform alignments across different modalities and domains. Specifically, M2CAN consists of four main components: cross-modal contrastive feature alignment (CMCFA) to bridge modal gaps, cross-domain contrastive feature alignment (CDCFA), cross-domain adversarial feature alignment (CDAFA), and uncertainty-aware classifier refinement (UACR) to bridge domain gaps. CMCFA, CDCFA, and CDAFA aim to learn domain-invariant multi-modal representations by conducting feature-level alignments for each modality, within each domain, and on the fused representations, respectively. UACR performs label space-level alignment by progressively selecting confident pseudo labels for the unlabeled target samples to conduct self-learning and participate in alignment. After such feature-level and label space-level alignments, different source and target domains are mapped into a shared multi-modal representation space, and the task classifiers are adapted to both the source and target domains. Extensive experiments are conducted on sentiment analysis and aesthetics assessment tasks. The results demonstrate that the proposed M2CAN outperforms state-of-the-art methods for the MSMMDA task by 2.8% and 2.1% in average accuracy, respectively. The code is available at https://github.com/jingjiang02/M2CAN.

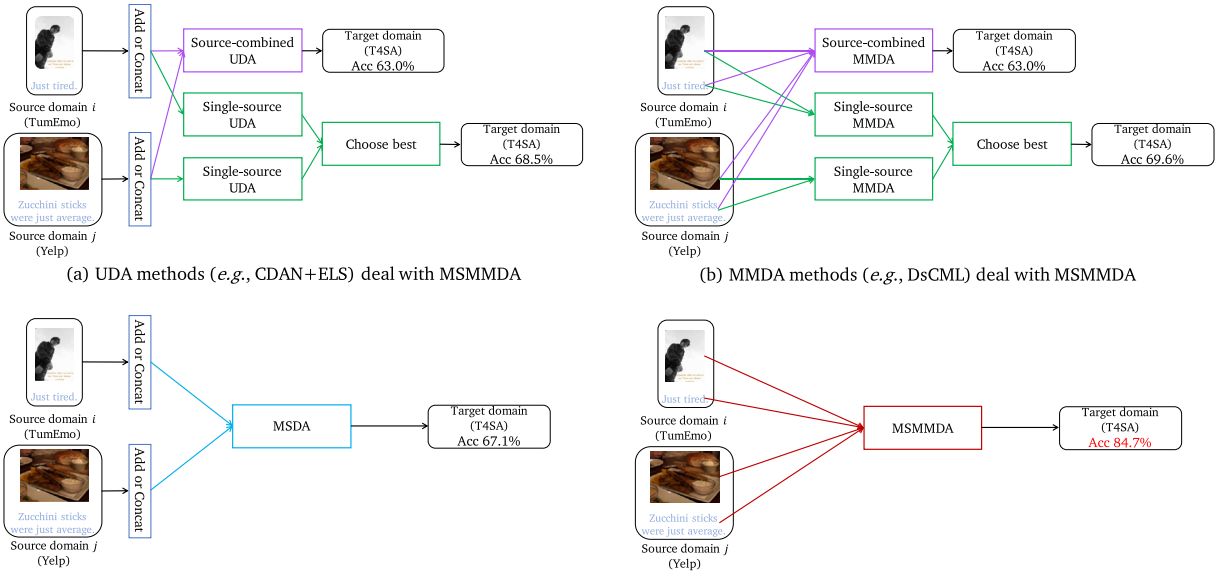
1. Introduction

In recent years, the evolution of web technologies has led to a significant increase in the availability of multi-modal information [1]. Typically, different modalities complement each other, and this phenomenon has attracted research interest in integrating multi-modal feature spaces for a more comprehensive representation of data objects [2–6]. While multi-modal methods have been explored in various fields over many years and have achieved commendable performance, they still require a large amount of multi-modal labeled data. Since the annotation of multi-modal data is computationally time-consuming, training a model on an existing labeled source domain and then transferring the model to the desired target domain has become a highly valuable alternative.

Due to the presence of domain shift [10], *i.e.*, the distributions of observed multi-modal data and labels differ between the source

* Corresponding author.
E-mail address: schzhao@tsinghua.edu.cn (S. Zhao).
[1] Equal contribution

https://doi.org/10.1016/j.inffus.2024.102862
Received 18 September 2024; Received in revised form 7 November 2024; Accepted 4 December 2024
Available online 11 December 2024



(a) UDA methods (*e.g.*, CDAN+ELS) deal with MSMMDA

(b) MMDA methods (*e.g.*, DsCML) deal with MSMMDA

(c) MSDA methods (*e.g.*, M^{3}SDA) deal with MSMMDA

(d) M2CAN (ours) deals with MSMMDA

Fig. 1. Comparison of different types of domain adaptation methods to deal with the MSMMDA task: (a) unsupervised domain adaptation (UDA), (b) multi-modal domain adaptation (MMDA), (c) multi-source domain adaptation (MSDA), and (d) multi-source multi-modal domain adaptation (MSMMDA), taking CDAN [7]+ELS [8], DsCML [3], M^{3}SDA [9], and the proposed M2CAN as examples, respectively.

and target domains, direct transfer cannot guarantee optimal generalization, often leading to a large performance decline [10,11]. To mitigate domain shift, domain adaptation (DA) [10,12,13] aims to train a model on the labeled source domain that can effectively generalize to the target domain through specific domain alignments, such as discrepancy-based [14,15], adversarial discriminative [16,17], adversarial generative [18,19], and self-supervision-based methods [20,21].

Current multi-modal DA (MMDA) methods focus solely on the single-source unsupervised setting [2,3,22,23], which assumes that the labeled multi-modal source data is collected from a single distribution. However, in practice, it is more common for the labeled source data to come from different distributions [24]. One straightforward method is to combine different sources into one source and then directly apply existing single-source MMDA methods. Because of the neglect of multiple source domain gaps, such methods may lead to sub-optimal results (see the comparison between Single-best and Source-combined results in Fig. 1(b)).

Consequently, effective multi-source domain adaptation (MSDA) techniques [25–34] are essential for sufficiently leveraging the discriminative information from different sources. Based on different alignment strategies, MSDA methods can be categorized into three groups: latent space transformation [9,35,36], intermediate domain generation [37–39], and task classifier refinement [40–42]. While these methods consider data from multiple domains, they overlook the multi-modal data scenario. Specifically, when adapted to multi-modal settings, their performance tends to be unsatisfactory, primarily due to their inability to bridge modal gaps, *e.g.*, the heterogeneous differences among modalities in the feature space [43]. Therefore, ineffective alignment of feature representations and inadequate mining of cross-modal information may result in interference among various modalities, causing the model to struggle to capture precise and consistent patterns.

In this paper, we extend the single-source MMDA and single-modal MSDA tasks to the multi-source multi-modal domain adaptation (MSMMDA) task. The comparison of these methods for handling multi-source and multi-modal data is illustrated in Fig. 1. As observed, the UDA methods either add or concatenate the features of each modality, followed by the single-source DA process. MMDA methods either merge the source domains into a single domain to perform source-combined DA or apply each single-source DA process and select the best one. MSDA methods first fuse the features of each modality before proceeding with the multi-source DA process. In contrast, the proposed MSMMDA leverages multi-source multi-modal data, facilitating knowledge transfer from the source domains with multi-modal samples to the target domain without compromise.

The MSMMDA task involves two major challenges: modal gaps between multiple modalities and domain gaps between multiple domains. Specifically, we propose a Multi-source Multi-modal Contrastive Adversarial Network, termed M2CAN, a novel framework that performs four alignments to alleviate the modal and domain gaps between multiple source and target domains with multi-modal data. As illustrated in Fig. 2, these alignments are conducted at two levels: feature level and label space level. (1) Cross-modal contrastive feature alignment (CMCFA) works on the representations within multi-modality in each domain. To mitigate the effects of multi-modal mismatch during adaptation, CMCFA dynamically adjusts the alignment based on the predicted differences between the individual headers of each modality. (2) Cross-domain contrastive feature alignment (CDCFA) works on the representations within multi-source for each modality. (3) Cross-domain adversarial feature alignment (CDAFA) aims to align the fused multi-modal representations. (4) Uncertainty-aware classifier refinement (UACR) conducts label space-level alignment, which adopts the predicted differences of target samples across multiple task classifiers and the aggregated output to select pseudo labels with high confidence. Only the selected samples in the target domain are involved in task classifier training and alignment to avoid negative optimization.

We conduct extensive experiments on two tasks: sentiment analysis and aesthetics assessment. For sentiment analysis, we utilize a combined dataset consisting of three domains, *i.e.*, TumEmo [44], T4SA [45], and Yelp [46]. Similarly, the performance of cross-domain aesthetics assessment is evaluated on a combined dataset with three domains, *i.e.*, AVA [47], PCCD [48], and RPCD [49]. The experimental results demonstrate that M2CAN outperforms the state-of-the-art DA methods for MSMMDA. In summary, the contributions of this paper are threefold:

- We introduce a novel and practical DA setting, namely multi-source multi-modal domain adaptation (MSMMDA). To the best of our knowledge, this is the first investigation into multi-modal domain adaptation with multiple source domains.

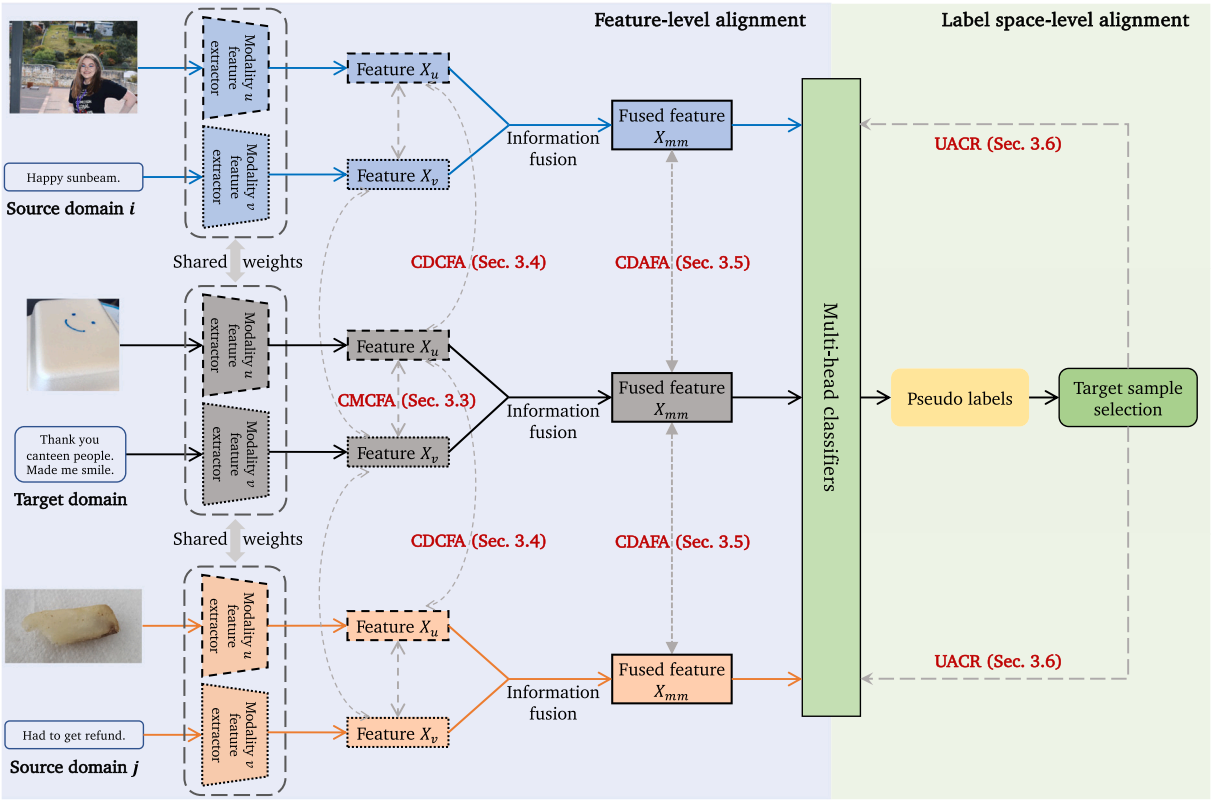


Fig. 2. The framework of the proposed Multi-source Multi-modal Contrastive Adversarial Network (M2CAN). In order to reduce the modal gap and domain gap in MSMMDA, the source and target domains are aligned on both the feature level and label space level. Feature-level alignment considers the individual features for each modality and the fused features for multiple modalities. Label space-level alignment is based on the pseudo labels of target samples, of which only the selected highly confident ones participate in the training procedure of the classifiers.

- We propose a novel framework, M2CAN, to perform MSMMDA using both feature-level alignments and label space-level alignments to reduce the modal and domain gaps in the MSMMDA task. M2CAN conducts three types of feature-level alignments: CMCFA within each domain, CDCFA for each modality, and CDAFA on the fused multi-modal representations. Additionally, UACR selects highly confident target pseudo labels for conducting label space-level alignment.
- We conduct extensive experiments on two benchmark datasets: one for sentiment analysis and the other for aesthetics assessment. Compared to the prior best methods, the proposed M2CAN achieves an average accuracy improvement of 2.8% and 2.1% for the two tasks.

The rest of this paper is structured as follows. Section 2 reviews the work related to MSMMDA. Section 3 introduces the proposed M2CAN in detail, including problem setup, overview, and different alignment strategies. Section 4 presents the experimental settings, results, ablation study, and visualization, followed by a conclusion in Section 5.

2. Related work

2.1. Multi-modal learning

Multi-modal tasks entail the processing and fusion of information from multiple modalities [50]. Within the realms of machine learning and artificial intelligence, these modalities typically encompass various types of data inputs, such as text, images, audio, and video, among others. Based on the methodology of feature fusion, two distinct strategies are identified [4]: model-free fusion and model-based fusion..

Model-free fusion, characterized by its operation independent of specific learning algorithms, has been extensively utilized for decades. As for the stage of fusion, multi-modal methods are generally classified into three categories [51]: early fusion [52,53], late fusion [2,3], and hybrid fusion [54,55] methods. **Model-based fusion** integrates fusion processes directly into the construction of the models. This approach typically employs straightforward techniques that are not specifically tailored for multi-modal data, such as attention mechanisms [56–58].

While these multi-modal methods have delivered impressive results, they are all supervised methods that require vast amounts of annotated samples. Especially when a new domain (such as a new scene) appears, it is necessary to annotate the samples of the new domain for the model to train. Therefore, adapting the model to unlabeled samples in a new domain, by leveraging labeled samples from other domains, would greatly reduce the labeling cost.

2.2. Unsupervised domain adaptation

The Unsupervised Domain Adaptation (UDA) task involves transferring knowledge from the annotated source domain to the unlabeled target domain. A typical UDA model includes two primary components: task-specific loss functions for the labeled source domain, and alignment loss functions between the source and target domains. UDA methods vary in their alignment strategies, and they can be categorized into different types [10]: discrepancy-based, adversarial discriminative, adversarial generative, and self-supervision-based methods.

Discrepancy-based methods. These methods [14,15,59–61] primarily utilize Maximum Mean Discrepancy (MMD) [62] to measure the distance between the source and target distribution. In addition to MMD, Correlation Alignment (CORAL) [63] is also a popular choice [64–67]. HoMM [68] enables arbitrary-order moment matching. This approach has shown that the first-order HoMM is equivalent to MMD and the second-order HoMM corresponds to CORAL. The Contrastive Domain Discrepancy (CDD) [69] provides a class-aware

approach to UDA. It focuses on minimizing the intra-class discrepancy and maximizing the inter-class margin. **Adversarial discriminative methods** typically utilize discriminators tasked with classifying source and target samples [17,70,71]. By fostering domain misclassification through an adversarial objective function, they aim to decrease the distance between source and target domain distributions. **Adversarial generative methods** bridge domain gaps by generating realistic images while aligning features across domains. SimGAN [72] and CyCADA [18] improve realism and consistency through feature alignment and cycle-consistency, while other methods use attention alignment [73] and multi-domain discriminators [19] to enhance domain-invariant representations. **Self-supervision-based methods** aim to decouple domain-specific features and domain-invariant features, typically using sample reconstruction as a self-supervised loss function, such as DRCN [74], DSN [20], and Fido [21].

While these DA methods have made good progress in various tasks concentrated on single-source and single-modal settings, they have not taken into account the multi-modal scenarios. Directly applying these DA methods to a multi-modal setting might result in unsatisfactory performance due to differences in the data distributions between modalities.

2.3. Multi-modal domain adaptation

Multi-modal Domain Adaptation (MMDA) presents greater complexity than UDA due to the need to account for distinct modality structures and the varying domain shifts associated with each modality [75]. Most MMDA methods can be divided into three categories based on the location of feature alignment: alignment before feature fusion, alignment after feature fusion, and mixed alignment.

Alignment before feature fusion refers to aligning each modality separately or aligning between modalities before conducting multi-modal feature fusion. [2] propose xMUDA where modalities learn from each other through mutual mimicking, disentangled from the segmentation objective. DsCML [3] enhances the efficacy of multi-modal information interactions for domain adaptation. [23] propose MMADT, which consists of DFB to recalibrate depth information and DAT to compensate for depth differences between the source and target domains. **Alignment after feature fusion** means aligning the fused multi-modal features. In Social Media Event Rumor Detection, MDDA [76] decomposes the multimedia posts into the event content information and the rumor writing style information, and then removes the event-specific features to obtain event-invariant rumor style features. In Audio-Visual Emotion Recognition, [77] combine the gradient reversal technique with an entropy loss as well as a soft-label loss on the fused multi-modal features. **Mixed alignment** refers to both alignment before and after feature fusion. MDANN [78] comprehensively learns domain-invariant features by constraining single-modal features, fused features, and attention scores. For fine-grained action recognition, [79] use a domain discriminator per modality that penalizes domain-specific features from each modality's stream.

These MMDA methods combine the advantages of multi-modal learning and UDA, enabling effective knowledge transfer from multi-modal data in the source domain to the target domain. However, there are generally multiple source domains in practice. Thus, extending these MMDA methods to multi-source settings could further improve their effectiveness.

2.4. Multi-source domain adaptation

Multi-source Domain Adaptation (MSDA) is a powerful extension of UDA, where the labeled data are collected from multiple sources with different distributions. Depending on the alignment strategies employed, MSDA can be broadly classified into three categories [80]: latent space transformation, intermediate domain generation, and task classifier refinement.

Latent space transformation aims to align the latent spaces (such as features) across different domains. This is typically achieved by optimizing either the discrepancy loss or the adversarial loss. Some methods use MMD [28,32], $\mathcal{L}2$ distance [26], and moment distance [9]. Some methods use adversarial loss for latent space transformation, including GAN loss [25], H-divergence [81], and Wasserstein distance [24]. T-SVDNet [30] incorporates Tensor Singular Value Decomposition (T-SVD) into the network's training pipeline. MKT [31] utilizes image-level and instance-level attention to promote positive cross-domain transfer and suppress negative transfer. **Intermediate domain generation** entails the creation of an adapted domain for each source, crafted to closely resemble the target domain. Following this, task models are then trained on these specifically adapted domains. For example, [27] adopt CoGAN [82], and [83] adopt CycleGAN [84] to construct an intermediate domain. Different from these methods, [37] use a variational autoencoder to learn a unified latent space that jointly aligns data from all source and target domains. **Task classifier refinement** addresses classification gaps that remain after feature and pixel alignment due to domain boundaries and class imbalance. Methods include pseudo-label training [33,40,85] and decision boundary refinement [9,29,41,86–88]. Advanced category-level alignment techniques utilize class-specific discriminators [38], MMD discrepancy measures [89,90], and prototypes [42,91,92].

These MSDA methods consider the domain gaps between source domains and leverage samples from multiple source domains, achieving superior performance compared to merely combining these source domains. However, these multi-source methods only consider samples from a single modality. In the case of multi-modal samples, performance may be impacted by the existence of modal gaps between modalities.

2.5. Sample selection by pseudo labels

Pseudo labels refer to the use of predicted labels during training which are treated as correct labels. The generation of pseudo labels for the target domain is a simple but effective method for DA to learn the feature representations of the target domain. In addition, these pseudo labels can serve as criteria for selecting samples for subsequent training, a process known as Sample Selection by Pseudo Labels. Typically, these methods can be divided into two categories: selecting samples in the source domain(s) and selecting samples in the target domain.

Selecting samples in source domain(s). PCA-SS [93] selects a subset of labeled data from the source domain. It ensures that the instance distribution of the source domain closely aligns with that of the target domain. [94] propose a landmark selection algorithm that reweights samples. CMSS [11] learns a dynamic curriculum for source samples. **Selecting samples in the target domain**. [95] jointly optimize representation, cross-domain transformation, and target label inference in an end-to-end manner. [96] employ an asymmetric use of three networks. Specifically, two networks are employed to label the target samples, while the third network is trained using the pseudo-labeled samples to yield target-discriminative representations. [97] iteratively select sets of pseudo-labeled target samples based on the image classifier and the domain classifier. [98] develop a pseudo-labeling curriculum using a density-based clustering algorithm. [99] propose a selective pseudo-labeling strategy based on structured prediction.

Our method selects high-quality target domain samples using pseudo labels. These selected samples are then used in self-learning and for alignment with the samples in source domains. Thus, our strategy differs from those methods that focus on selecting samples from the source domain. Additionally, these methods for selecting target domain samples primarily utilize filtered pseudo labels to adjust task loss or alignment loss. They do not utilize the pseudo labels to select the target domain samples that need to participate in alignment with the source domains. Therefore, our method is distinct from these approaches.

3. M2CAN

3.1. Problem setup

We consider the MSMMDA setting under the *covariate shift* assumption [100]. Assume that we have N source domains, denoted as $\mathcal{S} = \{S_i\}_{i=1}^{N}$, each containing labeled training data, and a target domain $\mathcal{T}$ consisting of only unlabeled training data from multiple modalities. Each source domain S_i contains a set of examples drawn from a joint distribution $p^{(S_i)}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M, \mathbf{y})$ on the input space $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_M$ and the output space $\mathcal{Y}$, where M represents the total number of modalities. Similarly, the target domain $\mathcal{T}$ consists of examples derived from a joint distribution $p^{(\mathcal{T})}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M, \mathbf{y})$, where both S_i and $\mathcal{T}$ share the same input space and output space. However, the label y for each sample in domain $\mathcal{T}$ remains unknown. Notably, there is a difference between the distributions $p^{(\mathcal{T})}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M, \mathbf{y})$ and $p^{(S_i)}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M, \mathbf{y})$. Also, there is a considerable difference between the source domain distributions $p^{(S_i)}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M, \mathbf{y})$ and $p^{(S_j)}(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M, \mathbf{y})$. This is why we need to introduce the multi-source setting into the multi-modal data. As a result, our objective is to learn a classifier $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_M \to \mathcal{Y}$ using labeled source samples from $\{S_i\}_{i=1}^{N}$, which can be transferred to the target domain $\mathcal{T}$, where only unlabeled data is available.

3.2. Overview

The proposed Multi-source Multi-modal Contrastive Adversarial Network (M2CAN) bridges the modal gaps and domain gaps of the MSMMDA task by executing both feature-level and label space-level alignments between the source and target domains. The framework is shown in Fig. 2. We use the pre-trained encoders to transform each modality from different domains into a semantic-preserving latent continuous feature space, and employ task classifiers for training the final classification model based on the aligned multi-modal features. M2CAN contains four main components:

Cross-modal contrastive feature alignment (CMCFA) aims to bridge the modal gap between multiple modalities in MSMMDA, aligning the encoded representations between different modalities by pair within each domain. However, due to the mismatch issue among multiple modalities, we further dynamically adjust the CMCFA based on the prediction differences between the individual headers of each modality.

Cross-domain contrastive feature alignment (CDCFA) aims to bridge the domain gaps in individual modalities across multiple domains in MSMMDA, aligning the encoded representations across different domains for each modality. Considering that the domain gap exists in each modality, the discrepancy between domains is reduced for every modality through contrastive learning.

Cross-domain adversarial feature alignment (CDAFA) aims to bridge the domain gaps between the fused multi-modal representations of multiple domains in MSMMDA. A fused multi-modal feature space $\mathcal{X}_{mm}$ is constructed by f_{mm}, which learns both semantic-preserving and semantic-relevant projection $f_{mm} : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_M \to \mathcal{X}_{mm}$. Adversarial learning is then employed to align the fused multi-modal features from different domains.

Uncertainty-aware classifier refinement (UACR) aims to bridge the domain gaps between the label distributions of multiple domains in MSMMDA, selecting target samples with highly confident pseudo labels to refine the task classifiers. The differences in predictions of target domain samples across multiple task classifiers, along with the aggregated output of these classifiers, are used to select pseudo labels. Only the selected target samples are involved in the task loss calculation to prevent negative optimization. Moreover, the selected target domain samples then participate in the alignment with the source domain samples. By employing UACR, M2CAN can discern the label distribution of samples in the target domain and adapt to the target domain on the label space level.

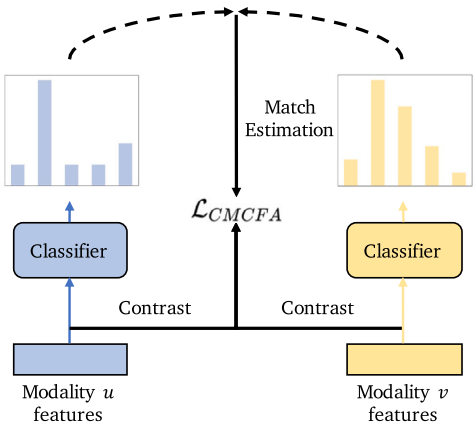


Fig. 3. Illustration of Cross-Modal Contrastive Feature Alignment (CMCFA). To avoid the impact of multi-modal mismatch during model training, the alignment loss is dynamically adjusted based on the predicted differences between the individual headers of each modality.

3.3. Cross-modal contrastive feature alignment

Motivation. Simply extracting features for each modality using separate encoders does not take the potential discrepancies between features in different modalities into account. In practice, certain modalities might encompass irrelevant or even misleading information. Moreover, features extracted without proper alignment come from disparate and potentially uncorrelated feature spaces, which could undermine the classification network's pattern recognition capability. Therefore, alignment between features across multiple modalities is necessary. For this purpose, we propose Cross-modal Contrastive Feature Alignment (CMCFA) to perform pairwise multi-modal alignment within each domain, depicted in Fig. 3. Firstly, we perform contrastive learning between paired modalities. Secondly, we propose dynamically adjusting contrastive learning based on the degree of matching between modalities to avoid forced alignment of modalities with mismatched information, reducing negative optimization.

Method. By applying data augmentation to each modality, we construct positive and negative sample pairs across different modalities within each domain. Assume that we have a batch of original features X_u and X_v for modality u and v, respectively. After data augmentation, the corresponding batches of original features are X'_u and X'_v. Note that the batch contains several samples from various source domains. Moreover, when pseudo labels in the target domain are generated, the batch also contains selected target domain samples with pseudo labels. The cross-modal contrastive loss can be constructed as follows [101]:

$$\mathcal{L}^{uv}_{CMCFA} = -\frac{1}{n} \cdot \mathbb{1}^T \cdot \log \left[\frac{e^{\mathbb{I} \circ \mathbb{T}} + e^{\mathbb{I} \circ \mathbb{T}'} + e^{\mathbb{I}' \circ \mathbb{T}} + e^{\mathbb{I}' \circ \mathbb{T}'}}{\mathbb{1}^T \cdot \left(e^{\mathbb{I} \cdot \mathbb{T}^T} + e^{\mathbb{I} \cdot \mathbb{T}'^T} + e^{\mathbb{I}' \cdot \mathbb{T}^T} + e^{\mathbb{I}' \cdot \mathbb{T}'^T} \right) \cdot \mathbb{1}} \right], \tag{1}$$

where $\mathbb{I} = X_u, \mathbb{I}' = X'_u, \mathbb{T} = X_v, \mathbb{T}' = X'_v$, $\circ$ represents the Hadamard product, and $\cdot$ is the dot product.

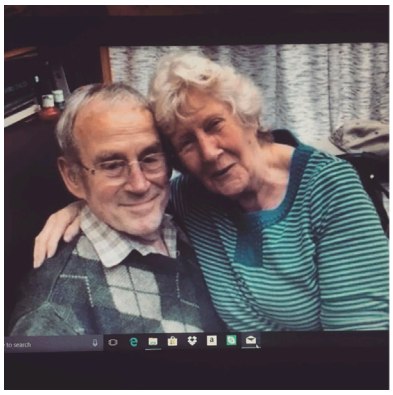
Note that there exists a mismatch issue between multiple modalities. Taking text and image modality as an example, as shown in Fig. 4(a), both the text and image convey the commentator's neutral sentiment, suggesting a match between the two modalities. On the other hand, in Fig. 4(b), while the text communicates the commentator's negative sentiment, the image portrays a positive sentiment, indicating a mismatch between these two modalities.

As a result, we introduce a match estimation mechanism for multiple modalities to dynamically adjust the cross-modal contrastive loss. We add distinct modality-related headers, using the output difference between them as the match estimation score for the modality pair. In detail, inspired by [102], we use KL-divergence between the predictions

Though I am often in the depths of misery, there is still calmness pure harmony and music inside me. I see paintings and drawings in the poorest cottages, in the dirtiest corners. And my mind is driven towards these things with an irresistible momentum. -Vincent van Gogh

Sentiment: neutral
Estimated Match Score = 0.9170

(a) Modalities matched example



Granny's laptop background caught me off guard and made the sadness of it all hit 6 billion times harder. Just heart broken.

Sentiment: negative
Estimated Match Score = 0.3080

(b) Modalities mismatched example

Fig. 4. Illustration of the modality mismatch issue in the TumEmo dataset [44]. The match score can be leveraged to measure the consistency between different modalities. A higher match score indicates a greater alignment in the meaning expressed across modalities, signifying better modality matching. (a) Example of matched text-image pair. This photo shows a painting by Vincent Van Gogh. Both the image and accompanying text convey the neutral sentiment of the commentator. (b) Example of mismatched text-image pair. Analyzing the image in isolation might lead one to believe that the uploader expresses a positive sentiment. Contrarily, the accompanying text reveals a negative sentiment.

of two classifiers as our measure of variance. Consider modalities u and v as an example:

$$Var^{uv} = \mathrm{KL}(F_u(X_u|\theta_u), F_v(X_v|\theta_v)), \tag{2}$$

where F_u and F_v represent classifiers for modalities u and v, respectively. KL represents KL-divergence operator. θ_u and θ_v are parameters for the classifiers of modalities u and v, respectively. If two classifiers of modalities u and v provide different predictions, the computed variance will yield a higher value, which reflects a lower match between the two modalities. We believe that during model training, it is advantageous to assign higher losses to samples exhibiting consistency across modalities and lower losses to samples showing inconsistency. This is premised on the notion that directing the model's focus towards consistent samples facilitates quicker convergence. Therefore, we define the match score for modalities u and v as $ExpVar^{uv} = \exp\{-Var^{uv}\}$.

By leveraging the estimated match score, we weight the cross-modal contrastive loss $\mathcal{L}^{uv}_{CMCFA}$ to derive an effective loss for optimization. However, when directly optimizing $\mathcal{L}^{uv}_{CMCFA} \cdot ExpVar^{uv}$, there is a risk that the model might only focus on minimizing the $ExpVar^{uv}$. By merely driving $ExpVar^{uv}$ towards zero, the overall loss could be minimized, which is not our intended outcome. Therefore, in order to avoid this issue, we introduce Var^{uv} to it as a regularization term. This ensures that the model does not overlook the importance of the base loss. Formally, our refined loss function is:

$$\mathcal{L}_{CMCFA} = \mathbb{E}[\sum_{u,v}(\mathcal{L}^{uv}_{CMCFA} \cdot \exp\{-Var^{uv}\} + Var^{uv})]. \tag{3}$$

By dynamically minimizing the distance between features of multiple modalities with matched modalities and maximizing the distance between features with mismatched modalities before and after data augmentation, CMCFA is able to force the encoders to extract closer features from semantically similar samples and farther apart features from semantically different samples.

3.4. Cross-domain contrastive feature alignment

Motivation. Training models across multiple domains within the same modality without accounting for domain gaps can lead to suboptimal performance. The distinct characteristics of each domain (*e.g.,* T4SA's social comments [45] and Yelp's food comments [46]) may result in divergent feature distributions, preventing the model from effectively generalizing across domains. These domain discrepancies can hinder the model's ability to recognize patterns consistently, leading to degraded classification accuracy. Therefore, aligning the feature distributions across domains within the same modality is critical to bridging these gaps. To this end, we propose Cross-domain Contrastive Feature Alignment (CDCFA), depicted in Fig. 5, to reduce the distribution distance of different domains within the same modality in the feature space. CDCFA utilizes contrastive learning to minimize the domain discrepancies, ensuring the model captures more unified and robust features across domains, thus improving generalization.

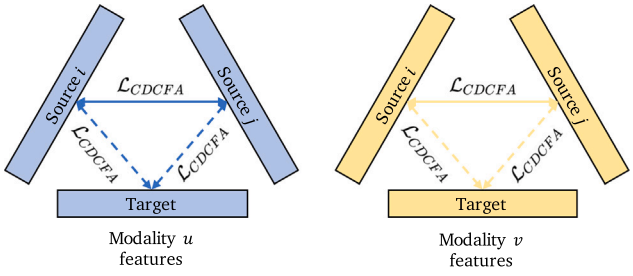


Fig. 5. Illustration of Cross-Domain Contrastive Feature Alignment (CDCFA). We minimize the distance between the feature distributions of different domains of the same modality to achieve the reduction of multiple domain gaps.

Method. To quantify the similarity between two distributions, we introduce MMD [62], which is defined as:

$$\mathcal{D}_{\mathcal{H}}(s_1, s_2) \triangleq \sup_{f_h \sim \mathcal{H}} \left(\mathbb{E}\left[f_h\left(\mathcal{I}^{s_1}\right)\right] - \mathbb{E}\left[f_h\left(\mathcal{I}^{s_2}\right)\right]\right)_{\mathcal{H}}, \tag{4}$$

where $\mathcal{I}^s \in X^s_u \cup X^{s\prime}_u$. Both X^s_u and $X^{s\prime}_u$ are all original feature batches of modality u in domain s without and with data augmentation, respectively. s_1 and s_2 are two domains. $\mathcal{H}$ denotes a class of functions and f_h is one of them. Formally, MMD measures the difference between two distributions based on their mean representations in the reproducing kernel Hilbert space (RKHS) [69].

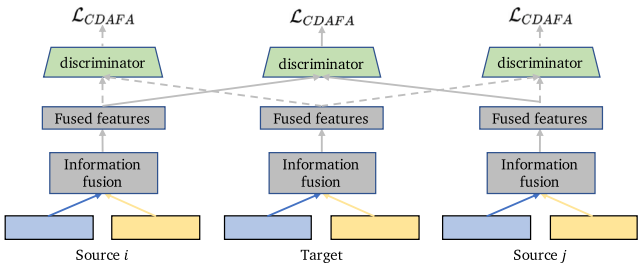


Fig. 6. Illustration of Cross-Domain Adversarial Feature Alignment (CDAFA). We align the fused multi-modal features between different domains in an adversarial manner to reduce the gap between multiple domains.

In practice, the squared value of MMD is estimated with the empirical kernel mean representations:

$$\hat{D}^{mmd} = \frac{1}{n_{s_1}^2}\sum_{i=1}^{n_{s_1}}\sum_{j=1}^{n_{s_1}} k\left(\mathcal{I}_i^{s_1}, \mathcal{I}_j^{s_1}\right) + \frac{1}{n_{s_2}^2}\sum_{i=1}^{n_{s_2}}\sum_{j=1}^{n_{s_2}} k\left(\mathcal{I}_i^{s_2}, \mathcal{I}_j^{s_2}\right) - \frac{2}{n_{s_1} n_{s_2}}\sum_{i=1}^{n_{s_1}}\sum_{j=1}^{n_{s_2}} k\left(\mathcal{I}_i^{s_1}, \mathcal{I}_j^{s_2}\right), \tag{5}$$

where $\mathcal{I}_i^s \in \mathcal{I}^s$. n_{s_1} and n_{s_2} denote the batch sizes of domain s_1 and s_2, respectively. k denotes a kernel function. The third term is adopted, while the first two terms are ignored in Eq. (5), as CDCFA mainly focuses on reducing the gaps between the two domains.

Due to the existence of multiple modalities, the domain gap between the two paired domains is decomposed into M parts, where each part represents the domain gap for a specific modality. The corresponding MMD for each modality is minimized as follows:

$$\mathcal{L}_{CDCFA}^u = \sum_{s_1,s_2}\sum_{\mathcal{I}^{s_1},\mathcal{I}^{s_2}}\left(-\frac{2}{n_{s_1} n_{s_2}}\sum_{i=1}^{n_{s_1}}\sum_{j=1}^{n_{s_2}} k\left(\mathcal{I}_i^{s_1}, \mathcal{I}_j^{s_2}\right)\right), \tag{6}$$

where $s_1 \in Dom, s_2 \in Dom \setminus s_1$. If the target domain pseudo labels are not generated, $Dom = \{S_1, S_2, \ldots, S_N\}$, otherwise $Dom = \{S_1, S_2, \ldots, S_N, \mathcal{T}\}$. The linear kernel is chosen as k for efficiency. Therefore, the above cross-domain contrastive loss for modality u can be simplified as below:

$$\mathcal{L}_{CDCFA}^u = \sum_{s_1,s_2}\sum_{\mathcal{I}^{s_1},\mathcal{I}^{s_2}} -\frac{2}{n_{s_1} n_{s_2}} \cdot \mathbb{1}^T \cdot \mathcal{I}^{s_1} \cdot \mathcal{I}^{s_2 T} \cdot \mathbb{1}. \tag{7}$$

Finally, the CDCFA of all modalities is calculated as follows:

$$\mathcal{L}_{CDCFA} = \sum_{u=1}^{M} \mathcal{L}_{CDCFA}^u. \tag{8}$$

3.5. Cross-domain adversarial feature alignment

Motivation. Merely reducing the gap between different modalities within the same domain or between different domains within the same modality is insufficient because the final task model relies on the fused multi-modal representations for prediction. Without aligning these multi-modal features across domains, the model may still struggle with inconsistencies in fused representations, leading to degraded performance. The discrepancies between multi-modal representations from different domains can introduce noise or irrelevant information, further complicating the model's ability to generalize. To address this, we propose Cross-domain Adversarial Feature Alignment (CDAFA), as illustrated in Fig. 6, which aims to bridge the domain gap between multi-modal representations across multiple domains. CDAFA employs an adversarial learning approach to ensure better alignment of fused representations, ultimately enhancing the model's ability to perform robust multi-modal predictions across domains.

Method. To better perform information fusion on the features of all modalities and produce a multi-modal feature space $\mathcal{X}_{mm}$ that includes enough task-related information, a multi-modal projection $f_{mm} : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_M \to \mathcal{X}_{mm}$ is constructed, where $\mathcal{X}_i$ is the feature space of the modality i.

To perform adversarial alignment, a set of domain classifiers are introduced as discriminators, which are used to distinguish features from each domain pair. Treating the aforementioned f_{mm} as a feature extractor, a conditional domain adversarial loss [7] is constructed to force the feature extractor to generate multi-modal features that are indistinguishable across domains. This results in the following cross-domain adversarial loss:

$$\mathcal{L}_{CDAFA} = \sum_{s_i}\sum_{s_j}(\mathbb{E}_{x_m^i \sim s_i} \log[D_{ij}(T(f_m^i, g_m^i))] + \mathbb{E}_{x_n^j \sim s_j} \log[1 - D_{ij}(T(f_n^j, g_n^j))]), \tag{9}$$

where $s_i \in Dom, s_j \in Dom \setminus s_i$. If the target domain pseudo labels are not generated, $Dom = \{S_1, S_2, \ldots, S_N\}$, otherwise $Dom = \{S_1, S_2, \ldots, S_N, \mathcal{T}\}$. x_m^i and x_n^j denote the mth sample in domain s_i and nth sample in domain s_j. D_{ij} denotes the discriminator between domain s_i and s_j. T is a map operation which converts fused features $f_m^i \in \mathcal{X}_{mm}$ and predicted logits g_m^i to class-conditional features. In this work, MultiLinearMap [7] is adopted as T.

Multiple modalities may not match each other, as explained in Section 3.3. Therefore, multi-modal fused features are noisy. To mitigate this, environment label smoothing [8] is introduced to the training procedure of the domain discriminators. It encourages the discriminators to output soft probability, which thus mitigates the over-confidence of the discriminator and alleviates the impact of the noisy multi-modal features. Therefore, the $\mathcal{L}_{CDAFA}$ is modified to a soft one:

$$\mathcal{L}_{CDAFA} = \sum_{s_i}\sum_{s_j}(\mathbb{E}_{x_m^i \sim s_i} W_m^{ij} \log[\alpha + D_{ij}(T(f_m^i, g_m^i))] + \mathbb{E}_{x_n^j \sim s_j} W_n^{ij} \log[1 - \alpha - D_{ij}(T(f_n^j, g_n^j))]), \tag{10}$$

where α is the label smoothing factor. In this work, α is set to 0.8. Loss weights are calculated as follows:

$$w_m^{ij} = 1 + exp\{-entropy(g_m^i)\}, \tag{11}$$

$$w_n^{ij} = 1 + exp\{-entropy(g_n^j)\}, \tag{12}$$

$$W_m^{ij} = \frac{(n_{s_i} + n_{s_j}) w_m^{ij}}{\sum_{m=1}^{n_{s_i}} w_m^{ij} + \sum_{n=1}^{n_{s_j}} w_n^{ij}}, \tag{13}$$

$$W_n^{ij} = \frac{(n_{s_i} + n_{s_j}) w_n^{ij}}{\sum_{m=1}^{n_{s_j}} w_m^{ij} + \sum_{n=1}^{n_{s_j}} w_n^{ij}}, \tag{14}$$

where the entropy function is defined as $entropy(x) = -x \cdot \log(x)$. n_{s_i} and n_{s_j} represent the batch sizes of domain s_i and s_j, respectively.

3.6. Uncertainty-aware classifier refinement

Motivation. A huge domain gap exists in multi-modal datasets. For instance, in the sentiment analysis task, the T4SA [45] dataset is derived from Twitter comments, which includes diverse tweets with informal language and varied image content. In contrast, the Yelp [46] dataset primarily focuses on structured food reviews, featuring vastly different image and comment styles, as demonstrated in Figs. 8(b) and 8(c). Confronted with such a substantial domain gap, directly aligning the source and target domains using traditional methods may lead to ineffective model training due to the stark differences in both content and style. To address this challenge, we propose Uncertainty-aware Classifier Refinement (UACR), as depicted in Fig. 7, which gradually generates high-quality pseudo-labels in the target domain for alignment with the source domain, achieving label space-level alignment. UACR not only accounts for content domain shifts but also ensures semantic consistency across label spaces by incorporating self-learning of target pseudo-labels, ultimately improving cross-domain adaptation.

Method. In the beginning, due to the lack of a well-trained model, only the source domains are aligned to obtain a preliminary model.

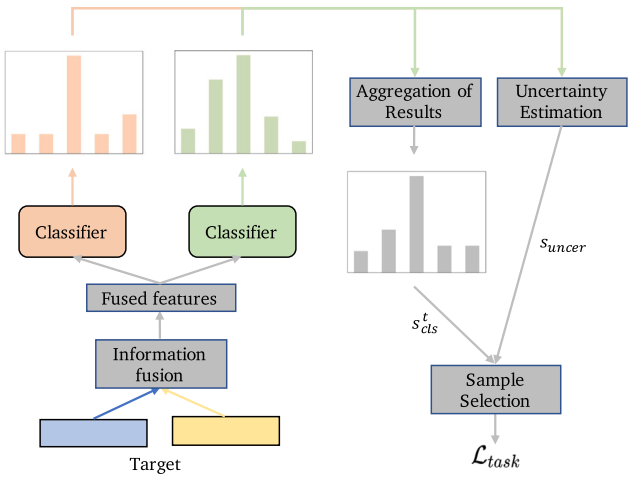


Fig. 7. Uncertainty-aware classifier refinement (UACR) on the label space level. The high-quality samples from the target domain are selected based on the classification score s^t_{cls} and uncertainty score s_{uncer}. These chosen samples will participate in task loss calculation and subsequent alignments with the source domain samples.

Afterward, the preliminary model is adopted to generate pseudo labels for the target domain to perform self-learning. During the generation of pseudo labels, especially in the early stage of model training, it is necessary to select pseudo labels due to the presence of low-quality labels. The uncertainty and confidence of multiple classification heads in predicting target domain samples are utilized to select pseudo labels. This is because the evaluation of uncertainty can identify unreliable predictions, while the evaluation of confidence ensures the accuracy of the predicted categories. This strategy can effectively improve the quality of the pseudo labels, thereby enhancing the performance of the model in the target domain.

Firstly, the uncertainty score is calculated by assessing the differences between the classification outputs for each sample in the target domain:

$$s_{uncer} = exp\{-Var_{ps}\}, \tag{15}$$

$$Var_{ps} = \sum_{i=1}^{N}\sum_{j=i+1}^{N}(\mathbb{E}[\mathrm{KL}(F^i_{cls}(f^t|\theta_i), F^j_{cls}(f^t|\theta_j))] + \mathbb{E}[\mathrm{KL}(F^j_{cls}(f^t|\theta_j), F^i_{cls}(f^t|\theta_i))]). \tag{16}$$

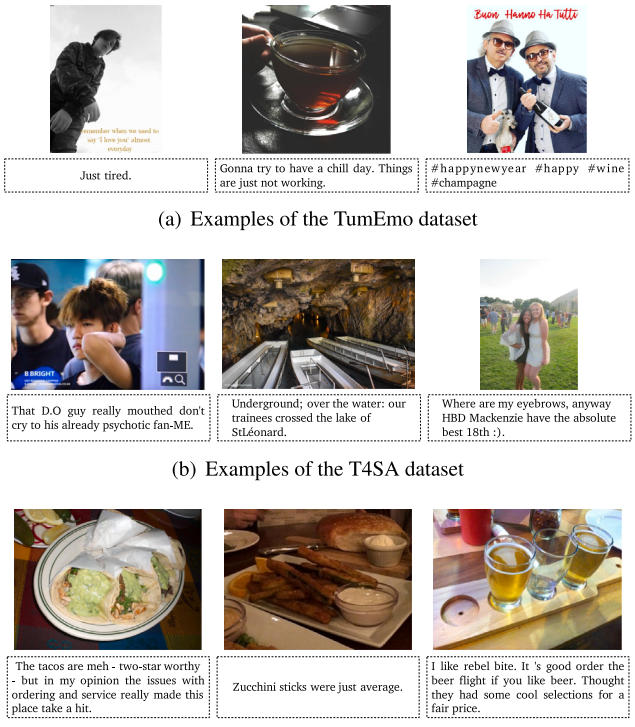
Here, F^i_{cls} denotes the ith classification head, and θ_i represents its parameters. A classification header is assigned for each source domain, resulting in a total of N classification headers. f^t stands for the fused multi-modal features of the input samples in the target domain. Secondly, the outputs of the N classification headers are aggregated as:

$$s^t_{cls} = \frac{\sum_{i=1}^{N} F^i_{cls}(f^t|\theta_i)}{N}. \tag{17}$$

Finally, the uncertainty score and the aggregated classification score s^t_{cls} are utilized as follows:

$$score = s_{uncer} \cdot s^t_{cls}. \tag{18}$$

$score$ is considered as the pseudo label confidence score for the sample. The samples are then sorted according to $score$ and those samples with the highest scores from each class are selected. This can ensure that the selected samples are reliable and roughly balanced in categories to avoid some categories being overlooked. The samples with pseudo labels have two purposes: (1) calculating task loss for self-learning, and (2) participating in alignment with the source domain samples. Through these two purposes, the model's perception of the distribution of target domain labels can be enhanced, enabling the model to better adapt to the target domain. At the same time, it can also alleviate the negative impact of low-quality target domain samples on the model alignment process.



(a) Examples of the TumEmo dataset

(b) Examples of the T4SA dataset

(c) Examples of the Yelp dataset

Fig. 8. Examples of three domains in the sentiment analysis task. Each sub-graph shows negative, neutral, and positive sentiments from left to right.

Algorithm 1 Learning procedure of M2CAN

Input: model M_{mm}, domain discriminators D, model optimizer opt_M, discriminators optimizer opt_D, total epochs E, mini batch B, pseudo label update rate r, and the b-th batch of samples $data_b$ from $Dom = \{S_1, S_2, ..., S_N\}$.

Initialize M_{mm} with pre-trained parameters, and initialize D randomly;
for e=1 to E **do**
  for b=1 to B **do**
    send $(x, y) \in data_b$ to model M_{mm};
    calculate $\mathcal{L}_{CMCFA}$ using Eq. (3);
    calculate $\mathcal{L}_{CDCFA}$ using Eq. (8);
    calculate $\mathcal{L}_{CDAFA}$ using Eq. (10);
    calculate $\mathcal{L}_{task}$ using Eq. (19);
    if target domain samples $\in data_b$ **then**
      calculate $\mathcal{L}_{mcc}$ using Eq. (20);
    end if
    backward;
    opt_M step;
    opt_D step;
  end for
  generate pseudo labels for the samples in the target domain with a ratio of $max\{\frac{r \cdot e}{E}, 1\}$ using Eq. (18);
  $Dom = Dom \cup \mathcal{T}$;
end for
return the adapted model M_{mm}.

3.7. M2CAN learning

We focus on MSMMDA for classification tasks. The task loss $\mathcal{L}_{task}$ adopts the standard cross entropy loss CE. The total task loss is formulated as:

$$\mathcal{L}_{task} = \sum_{i=1}^{N} \mathrm{CE}(F_{cls}^{i}(f_{mm}(X|\theta_i)), y) + \sum_{j=1}^{M} \mathrm{CE}(F_j(X_j|\theta_j), y), \tag{19}$$

where $X = \{X_1, \ldots, X_M\}$, X_j is the features of the jth modality for the sample $x \in (\bigcup_{i=1}^{N} S_i) \bigcup \mathcal{T}$, and y is the corresponding label of x. For the target domain sample, y is the pseudo label obtained in Section 3.6.

Further, in order to reduce the prediction confusion between correct and fuzzy categories on target domain samples, MCC [103] is introduced for additional label space level alignment. The aggregated output of the total N classification heads F_{cls}^{i} on the multi-modal features f^t of the target domain sample is obtained using Eq. (17). Then, s_{cls}^{t} is used as the input to the MCC, the following loss is obtained:

$$\mathcal{L}_{mcc} = \mathrm{MCC}(s_{cls}^{t}). \tag{20}$$

As a result, The overall objective function of M2CAN is formulated as:

$$\begin{aligned} \mathcal{L}_{M2CAN} = \alpha \cdot \mathcal{L}_{CMCFA} + \beta \cdot \mathcal{L}_{CDCFA} + \\ \gamma \cdot (\mathcal{L}_{CMAFA} + \mathcal{L}_{mcc}) + \mathcal{L}_{task}. \end{aligned} \tag{21}$$

The details of the M2CAN learning procedure are outlined in Algorithm 1.

In the inference stage, we perform information fusion and use the integrated average prediction s_{cls}^{t} as the final prediction result.

4. Experiments

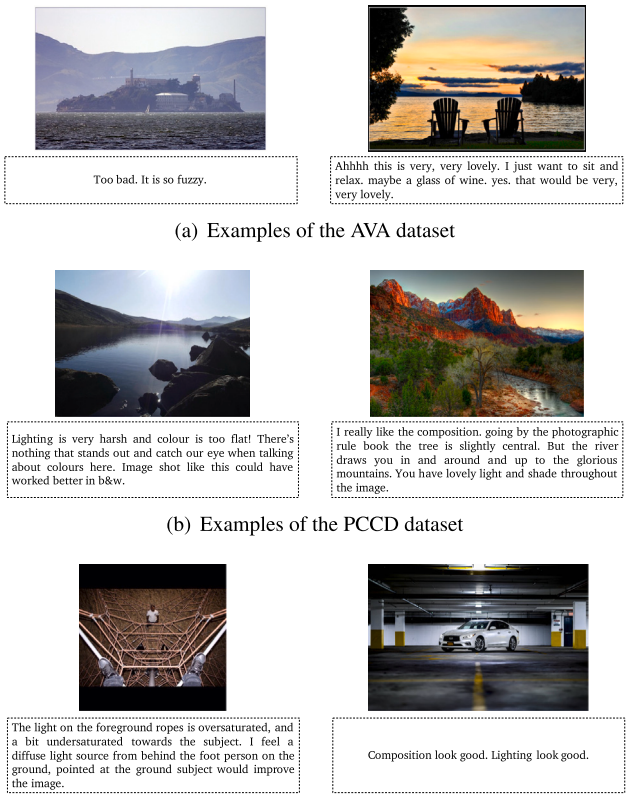
4.1. Experimental settings

4.1.1. Datasets

Due to the lack of a benchmark specifically designed for MSMMDA, we evaluate our approach using two combined datasets. The first task is sentiment analysis which consists of three public datasets on visual-textual modalities: TumEmo [44], T4SA [45], and Yelp [46]. Examples from these domains for the sentiment analysis task can be viewed in Fig. 8. The second task is aesthetics assessment which consists of three public visual-textual datasets: AVA [47], PCCD [48], and RPCD [49]. Examples of three domains in the aesthetics assessment task are shown in Fig. 9. We treat the three datasets as different domains because they follow distinct distributions. For our experiments, we establish an MSMMDA setting by taking each domain as the target and the rest as sources, totally including six scenes: →TumEmo (→TE), →T4SA (→T), →Yelp (→Y), →AVA (→A), →PCCD (→P), →RPCD (→R).

TumEmo [44] is a large-scale text-image emotion dataset, labeled by various emotions, with 195,265 instances from Tumblr. In TumEmo, there are seven emotions: Angry, Bored, Fear, Sad, Love, Calm, and Happy. In this paper, we treat Angry, Bored, Fear, and Sad as negative sentiments; Calm as a neutral sentiment; Love and Happy as positive sentiments. **T4SA** [45] consists of 470,586 user-generated tweets with images collected from the Twitter platform. Each tweet includes one textual review and several accompanying images, which is labeled with one of three sentiment types: negative, neutral, or positive. **Yelp** [46] contains customer-generated reviews of food services, such as restaurants, cafeterias, and dessert shops. In total, it has 44,305 reviews, including 244,569 images. Each review has a textual comment, a minimum of three images, and a sentiment polarity score ranging from 1 to 5. We consider scores of 1 and 2 as negative sentiment, a score of 3 as neutral sentiment, and scores of 4 and 5 as positive sentiment.

To balance the amount of samples in different domains, we randomly select 15,000 samples for training and 1500 samples for testing from each domain's training and testing set, respectively. The label distribution of the selected samples is shown in Fig. 10(a). Although the sample sizes of the three datasets remain the same across the three major classes, their respective sub-class distributions are different. This may potentially lead to domain gaps in the label space. The selection details and selected samples can be found in the source code we provide.



(a) Examples of the AVA dataset

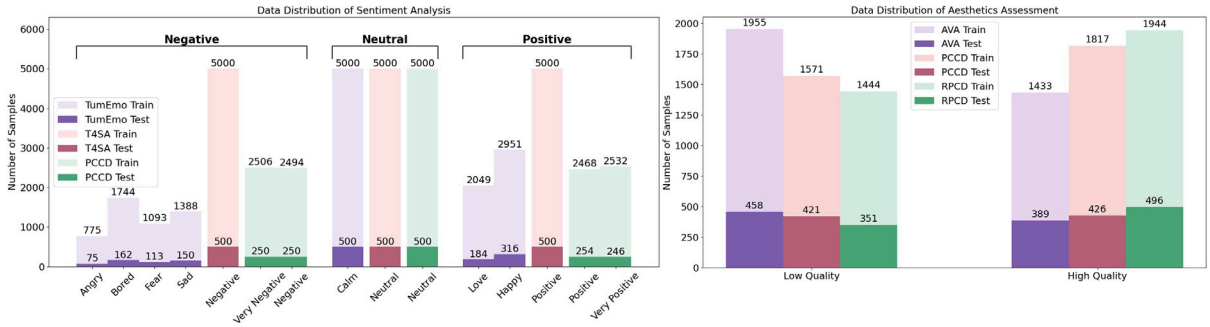
(b) Examples of the PCCD dataset

(c) Examples of the RPCD dataset

Fig. 9. Examples of three domains in the aesthetics assessment task. Each sub-graph shows low and high qualities from left to right.

AVA [47] contains more than 255,000 user-rated images. Each image is accompanied by an average of 200 ratings, ranging from 1 to 10, and also has user comments. The dataset is split into 235,000 training images and 20,000 testing images, ensuring no overlap. **PCCD** [48] is based on a professional photo critique website that provides experienced photographer reviews. A total of 4235 photos are showcased along with expert comments on seven key aspects: general impression, composition and perspective, color and lighting, subject of photo, depth of field, focus, use of camera, exposure, and speed. We consider the combination of these seven aspects as the comment for each image. **RPCD** [49] is a collection of 73,965 high-resolution images paired with photo critiques from Reddit communities. Following [49], we retain the samples with the same prediction of the two models [104,105] and discard those with different predictions.

The aesthetics assessment is formulated as a binary classification problem. In the AVA domain, images with a mean rating above 5.5 are treated as high-quality images, while the remaining images are classified as low-quality to ensure a balanced dataset. The PCCD domain uses a rating threshold of 8.0 to determine high quality. For the RPCD domain, the binary labels are provided by the two aforementioned models, and do not require a rating threshold. Then, we randomly select 3388 samples for training and 847 samples for testing from each domain's training and testing set, respectively. The label distribution of the selected samples is shown in Fig. 10(b). The label distributions of these three datasets are different, with AVA having more samples with low-quality, while PCCD and RPCD have more samples with high-quality. This shows the differences in label distribution between these



(a) Data distribution of sentiment analysis (b) Data distribution of aesthetics assessment

Fig. 10. The distribution of each of the datasets in the sentiment analysis and aesthetics assessment tasks.

three domains. The selection details and selected samples can be found in the source code we provide.

4.1.2. Baselines

We compare M2CAN with the following baselines: (1) **Source-only**, directly training on the source domains and testing on the target domain, which includes two settings: single-best, the best test accuracy on the target among all source domains individually; source-combined, the target accuracy of the model trained on the combined source domain.

(2) **Single-source DA methods**, including UDA methods CDAN [7], MCC [103], SDAT [106], and ELS [8], and MMDA methods xMUDA [2] and DsCML [3], all trained with both single-best and source-combined settings. CDAN [7] enhances discriminability and transferability through multilinear conditioning between feature and classifier predictions, and entropy conditioning to control prediction uncertainty. MCC [103] addresses class confusion, a common issue in domain adaptation, by reducing misclassifications between correct and ambiguous classes, resulting in improved transfer performance. SDAT [106] improves domain adversarial methods by enhancing smoothness in task loss minimization and stabilizing training, while smooth minima in adversarial loss are shown to hinder generalization. ELS [8] addresses the training instability of Domain Adversarial Training (DAT), reducing overconfidence in the domain discriminator and mitigating noisy environment labels, thereby improving stability, local convergence, and robustness. xMUDA [2] is proposed for cross-modal domain adaptive 3D semantic segmentation, where 2D images and 3D point clouds mutually learn through mimicking to handle domain shift, preventing the stronger modality from inheriting false predictions from the weaker one. DsCML [3] further enhances cross-modal domain adaptive 3D semantic segmentation by improving multi-modal interaction without losing 2D features. Additionally, Cross Modal Adversarial Learning (CMAL) is introduced to boost inter-domain complementarity between 2D and 3D data.

(3) **MSDA methods**, including MDAN [81], M^{3}SDA [9], and T-SVDNet [30]. MDAN [81] optimizes task-adaptive generalization bounds to learn feature representations that are invariant to multiple domain shifts while remaining discriminative for the learning task. M^{3}SDA [9] transfers knowledge from multiple labeled source domains to an unlabeled target domain by dynamically aligning their feature distribution moments. T-SVDNet [30] incorporates Tensor Singular Value Decomposition (T-SVD) and an uncertainty-aware weighting strategy to capture domain correlations and reduce negative transfer from noisy data.

(4) **Oracle**. Additionally, we report the results from an oracle setting, where the model is both trained and tested on the target domain. This can be considered an upper bound for domain adaptation performance.

4.1.3. Evaluation metrics

We employ average classification accuracy (Avg.) of all domains and classification accuracy (Acc), along with precision (P), recall (R), and F1-score (F1) of each domain, to evaluate the performance of both sentiment analysis and aesthetics assessment tasks. We use macro-averaging for these metrics, which better reflects the overall performance across all classes, especially in scenarios with class imbalance. This allows us to capture the performance of both sentiment analysis and aesthetics assessment more comprehensively. A higher classification accuracy, precision, recall, and F1-score directly correspond to better performance.

4.1.4. Implementation details

For the feature extractor of each modality, since only the visual and textual modalities are involved in our conducted experiments, we adopt two different settings for the image and text encoders. One uses ResNet50 [107] pre-trained on ImageNet as the image encoder and a 12-layer "bert-base-uncased" version BERT [108] for the text encoder. The other uses the image and text encoder of powerful Long-CLIP [109]. Due to the long length of the text in sentiment analysis and aesthetics assessment tasks, using the text encoder of CLIP [110] can result in truncating many texts, as it limits the text tokens to 77. Therefore, we adopt the text and image encoder of Long-CLIP with longer text token limits, up to 248. We utilize MLB [111] as the information fusion module f_{mm}. We use a fully connected layer to implement the discriminators, the modal headers, and the task classifiers.

In order to ensure that the loss terms have the same order of magnitude, the loss weight hyper-parameters α, β, and γ are set to 0.5, 0.2, and 0.05 by default, respectively. The pseudo label update rate r is set to 3 for all experiments. We conduct ablation studies on the values of these loss weights and find that they have negligible impact on the performance of the M2CAN.

At the initial training stage (also called the warm-up stage), we only align the source domain samples, without the target domain samples participating in the alignment. After generating pseudo labels for the target domain, the filtered target domain samples participate in alignment with the source domain samples. For all experiments, we perform 1 epoch for the former stage, and perform 9 epochs for the latter stage.

We use Adam [112] as the optimizer with a batch size of 8. The learning rate is 2e-5 for feature extractors, and 5e-4 for the rest by default. All experiments are implemented in PyTorch and conducted on a machine with a single NVIDIA RTX 3090 with 24 GB memory.

4.2. Comparison with the state-of-the-art

The performance comparisons based on the feature extractor of ResNet50+BERT between the proposed M2CAN and the other methods for visual-textual sentiment analysis and aesthetics assessment, including Source-only, UDA, MMDA, MSDA, and Oracle are presented

Table 1
Comparison with the state-of-the-art DA methods based on ResNet50+BERT for sentiment analysis, measured by average accuracy (Avg.) (%), accuracy (Acc) (%), macro precision (P) (%), macro recall (R) (%), and macro F1 (%). The best result is emphasized in **bold** and the second best method is emphasized in underline.

Standard	Method	Detail	Avg.	→TumEmo (→TE)				→T4SA (→T)				→Yelp (→Y)			
				Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Source-only	Single-best	–	58.0	57.3	59.4	57.3	57.0	61.1	61.6	61.1	56.7	55.5	37.3	55.5	44.6
	Combined		56.6	56.4	58.1	56.4	56.3	58.3	59.8	58.3	58.3	55.0	52.6	55.0	48.3
Single-best DA	CDAN(NeurIPS2018)	CDAN+ELS	62.7	60.9	60.4	60.9	60.5	68.5	74.9	68.5	68.9	58.7	57.5	58.7	56.8
	MCC(ECCV2020)	CDAN+MCC+ELS	61.9	61.6	60.4	61.6	60.1	67.2	67.7	67.2	67.4	56.9	56.8	56.9	56.7
	SDAT(ICML2022)	CDAN+SDAT+ELS	62.7	59.6	60.5	59.6	60.0	68.5	68.6	68.5	67.9	59.9	59.9	59.9	59.8
	ELS(ICLR2023)	CDAN+MCC+SDAT+ELS	62.3	57.5	60.0	57.5	57.4	74.1	74.7	74.1	73.9	55.3	55.1	55.3	54.6
	xMUDA (CVPR2020)	Text-only	58.3	57.8	58.1	57.8	57.3	60.2	58.9	60.2	54.4	56.9	54.3	56.9	48.5
		Image-only	34.9	33.8	41.9	33.8	20.3	35.8	35.7	35.8	35.6	35.0	36.0	35.0	26.0
		Fusion	58.8	57.9	58.2	57.9	57.8	61.9	61.2	61.9	58.4	56.5	55.0	56.5	49.1
	DsCML (ICCV2021)	Text-only	61.6	59.5	59.8	59.5	58.8	69.1	74.4	69.1	69.3	56.1	38.9	56.1	45.3
		Image-only	36.4	37.3	37.2	37.3	34.6	33.9	33.9	33.9	33.8	37.9	38.0	37.9	36.9
		Fusion	62.0	60.2	60.5	60.2	59.5	69.6	75.8	69.6	70.0	56.1	39.0	56.1	45.3
Source-combined DA	CDAN(NeurIPS2018)	CDAN+ELS	58.9	57.9	57.4	57.9	57.6	63.0	68.7	63.0	62.0	55.8	55.1	55.8	55.1
	MCC(ECCV2020)	CDAN+MCC+ELS	62.7	57.3	56.7	57.3	55.6	75.1	78.1	75.1	75.3	55.7	55.3	55.7	55.5
	SDAT(ICML2022)	CDAN+SDAT+ELS	62.2	57.9	57.0	57.9	56.9	69.8	70.6	69.8	69.9	58.9	60.0	58.9	59.2
	ELS(ICLR2023)	CDAN+MCC+SDAT+ELS	67.9	62.3	62.8	62.3	62.3	83.4	83.6	83.4	83.5	57.9	57.7	57.9	57.2
	xMUDA (CVPR2020)	Text-only	59.6	59.1	59.5	59.1	58.9	64.1	64.3	64.1	59.1	55.7	51.6	55.7	45.0
		Image-only	36.8	34.0	37.5	34.0	26.5	39.7	39.6	39.7	39.3	36.8	39.8	36.8	28.6
		Fusion	59.5	57.4	58.4	57.4	57.6	64.3	64.5	64.3	59.5	56.7	53.3	56.7	48.1
	DsCML (ICCV2021)	Text-only	58.8	58.3	58.1	58.3	57.6	62.7	64.0	62.7	61.1	55.5	49.1	55.5	45.9
		Image-only	37.9	40.7	41.0	40.7	40.2	36.9	36.8	36.9	36.6	36.1	36.3	36.1	35.9
		Fusion	58.9	58.7	58.8	58.7	58.1	63.0	64.6	63.0	62.5	55.1	48.3	55.1	43.6
MSDA	MDAN (NeurIPS2018)	–	58.8	59.1	60.1	59.1	59.2	61.9	67.8	61.9	62.2	55.5	53.1	55.5	52.8
	M^3SDA (ICCV2019)		60.4	58.0	56.7	58.0	56.9	67.1	69.9	67.1	67.1	56.1	54.7	56.1	53.6
	T-SVDNet (ICCV2021)		59.1	58.2	59.1	58.2	58.0	61.5	63.7	61.5	53.9	57.7	54.8	57.7	53.9
MSMMDA	M2CAN (Ours)	–	**70.7**	**63.8**	**63.2**	**63.8**	**63.4**	**84.7**	**84.8**	**84.7**	**84.7**	**63.7**	**63.7**	**63.7**	**63.7**
Oracle	Oracle	–	85.1	85.4	86.0	85.4	85.5	95.1	95.1	95.1	95.1	74.7	74.6	74.7	74.7

in Table 1 and Table 2, respectively. From the results, we have the following observations:

(1) Without alleviating the domain shift between the source and target domains, both source-only settings, *i.e.*, single-best and source-combined, obtain low classification accuracy. Specifically, their average accuracy is almost 30% and 15% lower than the oracle setting in both tasks, respectively. Note that under the source-only setting, having more source domain training data does not necessarily guarantee better performance in the target domain, *e.g.*, 58.0% vs. 56.6% average accuracy in the sentiment analysis task in Table 1. As a result, the gap between source domains is also worth considering.

(2) When directly applying single-source DA methods to the MSMMDA task, most of them, including UDA and MMDA in single-best and source-combined settings, outperform the source-only setting. Since multi-modal data vary a lot across domains, the extracted features related to tasks differ as well. Therefore, these single-source methods can help bridge the domain gap to improve the results. When fusing text and image modalities, the performance of MMDA methods might not be better than that of using a single modality. For instance, both xMUDA and DsCML are late-fusion methods. If the alignment of a single modality fails to achieve good consistency, the predictions from the two modalities will diverge largely when they are combined, leading to a decline in performance. For example, after fusing multi-modal features, the performance of xMUDA is reduced by 0.4% compared to non-fusion in the single-best DA setting of →RPCD on aesthetics assessment in Table 2.

(3) When comparing the performances of the source-combined and single-best settings of single-source DA methods, it becomes apparent that naively applying single-source domain adaptation approaches, including UDA and MMDA methods, to a combined dataset from different sources can lead to sub-optimal results, *e.g.*, the average accuracy of 62.0% vs. 58.9% on DsCML in Table 1. This observation underscores the motivation for our research on MSMMDA.

(4) Due to the consideration of domain gaps in the fused features between multiple source domain samples, MSDA methods achieve better results than Source-only settings on both sentiment analysis and aesthetics assessment in Tables 1 and 2. However, their performance is sub-optimal as they ignore the modal gaps across multiple modalities.

(5) The proposed M2CAN outperforms other methods in all adaptation settings, achieving an average accuracy of 70.7% in sentiment analysis (Table 1) and 74.7% in aesthetics assessment (Table 2). When compared to the best results from the Source-only, Single-best DA, Source-combined DA, and MSDA settings, M2CAN exhibits average accuracy gains of 12.7%, 8.0%, 2.8%, and 10.3% for sentiment analysis, and 8.0%, 2.1%, 2.8%, and 4.0% for aesthetics assessment, respectively. These results demonstrate that the proposed M2CAN can achieve better performance than the state-of-the-art DA methods including UDA, MMDA, and MSDA methods. The superior performance of M2CAN benefits from the joint feature-level and label space-level alignment to reduce the modal gaps and domain gaps in MSMMDA. Notably, the proposed M2CAN method achieves similar values across accuracy, precision, recall, and F1, highlighting its balanced and stable performance. This advantage ensures that the model not only performs well in terms of overall classification accuracy but also maintains consistent performance in precision and recall, indicating reliable and comprehensive results across all metrics.

The experimental results of Long-CLIP as a feature extractor on sentiment analysis and aesthetics assessment are shown in Table 3 and Table 4, respectively. The observation is consistent with that for ResNet50+BERT. Our M2CAN is 12.2%, 3.5%, 4.8%, and 6.1% for sentiment analysis, and 6.7%, 2.1%, 4.0%, and 5.3% for aesthetics assessment higher than other methods on average accuracy in Source-only, Single-best DA, Source-combined DA, and MSDA settings, respectively.

Table 2
Comparison with the state-of-the-art DA methods based on ResNet50+BERT for aesthetics assessment, measured by average accuracy (Avg.) (%), accuracy (Acc) (%), macro precision (P) (%), macro recall (R) (%), and macro F1 (%). The best result is emphasized in **bold** and the second best method is emphasized in underline.

Standard	Method	Detail	Avg.	→AVA (→A)				→PCCD (→P)				→RPCD (→R)			
				Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Source-only	Single-best	–	66.3	68.0	69.1	66.6	66.3	64.7	65.3	64.4	64.1	66.2	68.9	68.6	66.2
	Combined		66.7	70.5	74.3	71.9	70.1	66.1	67.4	66.0	65.4	63.5	72.3	68.0	62.7
Single-best DA	CDAN(NeurIPS2018)	CDAN+ELS	71.1	73.3	76.9	74.7	73.0	68.4	69.7	68.3	67.7	71.6	70.7	70.8	70.7
	MCC(ECCV2020)	CDAN+MCC+ELS	72.6	76.0	76.6	76.6	76.0	69.2	70.7	69.1	68.6	72.7	72.7	73.4	72.5
	SDAT(ICML2022)	CDAN+SDAT+ELS	70.9	77.9	77.9	78.1	77.9	68.2	69.0	68.2	67.9	66.5	67.4	66.8	66.2
	ELS(ICLR2023)	CDAN+MCC+SDAT+ELS	70.8	77.1	77.6	77.6	77.1	68.7	68.9	68.7	68.6	66.6	70.2	69.4	66.5
	xMUDA (CVPR2020)	Text-only	72.6	75.3	76.5	76.1	75.3	69.3	71.5	69.4	68.6	73.1	73.7	74.4	73.1
		Image-only	54.4	54.2	60.4	50.2	35.6	50.5	63.5	50.8	35.6	58.6	29.3	50.0	36.9
		Fusion	72.1	74.0	77.1	72.5	72.3	69.5	**71.8**	69.7	69.1	72.7	73.0	73.6	72.6
	DsCML (ICCV2021)	Text-only	71.8	76.5	76.8	76.9	76.5	66.9	70.3	67.1	65.6	72.1	71.6	70.1	70.4
		Image-only	54.5	53.7	51.3	50.5	42.7	51.2	52.1	50.3	36.5	58.6	29.3	50.0	36.9
		Fusion	71.1	77.0	77.2	77.3	77.0	66.5	69.8	66.7	65.3	69.7	69.5	66.9	67.1
Source-combined DA	CDAN(NeurIPS2018)	CDAN+ELS	69.3	75.7	76.3	76.2	75.7	67.5	68.1	67.5	67.2	64.8	66.0	66.3	64.8
	MCC(ECCV2020)	CDAN+MCC+ELS	71.9	77.3	77.7	76.7	76.9	67.8	68.9	67.7	67.2	70.7	73.1	73.0	70.7
	SDAT(ICML2022)	CDAN+SDAT+ELS	69.4	76.0	77.5	76.9	76.0	68.5	68.8	68.4	68.3	63.6	65.9	65.8	63.6
	ELS(ICLR2023)	CDAN+MCC+SDAT+ELS	70.7	70.4	73.5	71.7	70.0	68.8	69.3	68.9	68.7	73.0	**75.3**	75.3	73.0
	xMUDA (CVPR2020)	Text-only	67.2	71.2	76.5	76.5	70.6	67.5	69.2	67.4	66.8	62.8	61.8	62.0	61.9
		Image-only	53.9	54.0	47.0	49.9	35.5	50.3	25.2	50.0	33.5	57.3	54.2	53.4	52.4
		Fusion	67.7	72.9	76.1	74.2	72.6	67.4	69.1	67.3	66.6	62.8	61.9	62.1	62.0
	DsCML (ICCV2021)	Text-only	66.7	71.9	76.2	73.4	71.5	67.3	68.3	67.2	66.8	60.8	70.8	65.7	59.6
		Image-only	52.3	54.6	54.2	51.2	42.4	50.5	58.4	50.8	36.3	51.8	53.7	53.3	50.9
		Fusion	66.5	72.4	76.5	73.9	72.0	67.7	68.5	67.4	67.0	59.4	70.7	64.6	57.8
MSDA	MDAN (NeurIPS2018)	–	69.8	72.9	75.8	74.1	72.6	68.5	68.5	68.5	68.4	68.1	72.9	71.4	68.0
	M^3SDA (ICCV2019)		69.8	74.9	77.3	76.0	74.7	68.0	69.5	67.9	67.3	66.5	65.3	64.6	64.8
	T-SVDNet (ICCV2021)		70.7	75.3	76.9	76.2	75.3	68.2	68.4	68.2	68.2	68.7	73.7	72.0	68.6
MSMMDA	M2CAN (Ours)	–	**74.7**	**79.9**	**79.8**	**80.0**	**79.9**	**69.8**	69.8	**69.8**	**69.8**	**74.5**	74.7	**75.4**	**74.4**
Oracle	Oracle	–	79.9	81.0	80.3	79.7	79.9	77.6	77.7	77.6	77.6	81.1	80.7	80.1	80.4

Table 3
Comparison with the state-of-the-art DA methods based on Long-CLIP for sentiment analysis, measured by average accuracy (Avg.) (%), accuracy (Acc) (%), macro precision (P) (%), macro recall (R) (%), and macro F1 (%). The best result is emphasized in **bold** and the second best method is emphasized in underline.

Standard	Method	Detail	Avg.	→TumEmo (→TE)				→T4SA (→T)				→Yelp (→Y)			
				Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Source-only	Single-best	–	55.3	55.3	56.1	55.3	55.1	58.2	59.1	58.2	55.8	52.5	51.7	52.5	51.6
	Combined		54.3	54.6	56.0	54.6	54.4	60.2	61.1	60.2	53.3	48.1	47.7	48.1	47.7
Single-best DA	CDAN(NeurIPS2018)	CDAN+ELS	60.4	55.4	54.9	55.4	54.9	71.4	72.2	71.4	71.6	54.4	53.1	54.4	53.5
	MCC(ECCV2020)	CDAN+MCC+ELS	62.0	56.5	55.5	56.5	55.8	74.7	75.4	74.7	74.8	54.9	53.5	54.9	53.5
	SDAT(ICML2022)	CDAN+SDAT+ELS	63.1	55.8	55.2	55.8	55.4	77.9	78.3	77.9	77.8	55.5	54.3	55.5	54.6
	ELS(ICLR2023)	CDAN+MCC+SDAT+ELS	64.0	57.0	56.3	57.0	56.4	79.7	79.6	79.7	79.4	55.3	54.1	55.3	54.4
	xMUDA (CVPR2020)	Text-only	57.2	57.4	57.5	57.4	56.6	60.7	60.7	60.7	54.4	53.5	53.9	53.5	53.7
		Image-only	44.6	51.9	52.3	51.9	51.3	42.9	43.3	42.9	40.7	38.9	39.3	38.9	38.1
		Fusion	57.1	57.5	57.7	57.5	56.7	60.5	60.6	60.5	57.5	53.2	53.3	53.2	53.0
	DsCML (ICCV2021)	Text-only	59.4	55.5	55.3	55.5	54.9	70.1	69.6	70.1	69.6	52.5	54.1	52.5	53.0
		Image-only	40.4	41.1	41.1	41.1	41.0	40.7	43.2	40.7	37.4	39.4	39.8	39.4	38.4
		Fusion	59.3	55.8	55.6	55.8	55.2	69.4	68.7	69.4	68.6	52.8	54.6	52.8	53.4
Source-combined DA	CDAN(NeurIPS2018)	CDAN+ELS	57.7	54.1	52.4	54.1	52.6	62.9	61.9	62.9	60.9	56.1	55.3	56.1	55.6
	MCC(ECCV2020)	CDAN+MCC+ELS	62.7	57.6	57.0	57.6	57.1	75.8	75.8	75.8	75.7	54.7	54.1	54.7	54.4
	SDAT(ICML2022)	CDAN+SDAT+ELS	60.7	57.4	56.1	57.4	56.3	70.7	70.4	70.7	70.3	54.1	53.8	54.1	53.9
	ELS(ICLR2023)	CDAN+MCC+SDAT+ELS	61.7	59.5	58.8	59.5	58.8	71.1	70.7	71.1	70.4	54.6	55.2	54.6	54.9
	xMUDA (CVPR2020)	Text-only	57.8	57.3	57.2	57.3	57.1	60.8	60.3	60.8	57.3	55.2	51.9	55.2	46.0
		Image-only	44.2	51.5	51.5	51.5	51.2	43.1	43.5	43.1	41.4	38.1	42.9	38.1	32.5
		Fusion	57.8	58.4	58.1	58.4	58.1	60.0	59.3	60.0	56.2	55.1	51.3	55.1	45.7
	DsCML (ICCV2021)	Text-only	56.4	56.1	54.8	56.1	54.0	59.5	58.5	59.5	58.6	53.6	55.6	53.6	47.7
		Image-only	39.4	40.9	42.1	40.9	40.4	40.9	41.0	40.9	40.8	36.5	35.7	36.5	30.6
		Fusion	56.4	55.9	54.3	55.9	53.7	60.1	59.5	60.1	59.6	53.3	54.8	53.3	47.2
MSDA	MDAN (NeurIPS2018)	–	56.5	54.5	55.1	54.5	54.5	62.2	61.7	62.2	60.1	52.9	48.9	52.9	45.6
	M^3SDA (ICCV2019)		61.4	57.3	57.9	57.3	57.3	73.1	73.1	73.1	73.1	53.8	54.3	53.8	54.0
	T-SVDNet (ICCV2021)		56.6	55.5	55.7	55.5	55.5	61.1	61.4	61.1	55.1	53.1	54.2	53.1	53.5
MSMMDA	M2CAN (Ours)	–	**67.5**	**61.8**	**61.7**	**61.8**	**61.7**	**80.7**	**80.7**	**80.7**	**80.7**	**60.1**	**59.8**	**60.1**	**60.0**
Oracle	Oracle	–	85.7	90.3	90.4	90.3	90.3	94.3	94.3	94.3	94.3	72.5	72.5	72.5	72.5

Table 4
Comparison with the state-of-the-art DA methods based on Long-CLIP for aesthetics assessment, measured by average accuracy (Avg.) (%), accuracy (Acc) (%), macro precision (P) (%), macro recall (R) (%), and macro F1 (%). The best result is emphasized in **bold** and the second best method is emphasized in underline.

Standard	Method	Detail	Avg.	→AVA (→A)				→PCCD (→P)				→RPCD (→R)			
				Acc	P	R	F1	Acc	P	R	F1	Acc	P	R	F1
Source-only	Single-best	–	66.5	70.7	70.9	71.0	70.7	67.5	67.5	67.5	67.5	61.2	60.8	61.1	60.7
	Combined		66.1	73.8	74.8	74.5	73.8	67.1	67.2	67.1	66.8	57.5	62.0	60.8	57.1
Single-best DA	CDAN(NeurIPS2018)	CDAN+ELS	70.1	74.0	75.4	74.9	74.0	69.0	69.8	69.0	68.7	67.2	66.3	66.5	66.4
	MCC(ECCV2020)	CDAN+MCC+ELS	70.2	77.5	77.3	77.4	77.3	67.9	67.9	67.9	67.9	65.2	65.3	65.8	65.0
	SDAT(ICML2022)	CDAN+SDAT+ELS	67.5	69.3	71.9	70.5	69.1	67.8	69.0	67.8	67.3	65.3	64.3	64.4	64.3
	ELS(ICLR2023)	CDAN+MCC+SDAT+ELS	65.9	69.0	75.3	70.8	68.0	67.1	68.2	67.1	66.6	61.6	66.3	64.9	61.4
	xMUDA (CVPR2020)	Text-only	70.5	73.9	74.7	74.5	73.9	69.0	69.8	69.0	68.6	68.7	69.0	69.6	68.6
		Image-only	63.2	69.2	69.0	69.1	69.0	62.8	63.0	62.8	62.7	57.6	54.6	53.7	52.7
		Fusion	69.5	75.1	75.9	75.6	75.0	67.9	70.9	68.0	66.8	65.5	65.3	61.1	60.3
	DsCML (ICCV2021)	Text-only	71.1	75.4	75.3	75.5	75.4	67.8	68.0	67.8	67.7	70.1	69.2	69.1	69.1
		Image-only	55.0	51.8	53.6	53.2	50.8	54.7	55.1	54.7	53.9	58.6	54.5	51.0	42.3
		Fusion	70.4	75.2	75.0	75.2	75.0	67.1	67.3	67.1	67.0	69.0	67.9	67.3	67.5
Source-combined DA	CDAN(NeurIPS2018)	CDAN+ELS	67.7	72.9	76.2	74.2	72.6	68.9	68.9	68.9	68.9	61.2	69.0	65.5	60.4
	MCC(ECCV2020)	CDAN+MCC+ELS	69.2	77.1	78.0	77.8	77.1	68.7	69.0	68.8	68.6	61.8	66.4	65.1	61.5
	SDAT(ICML2022)	CDAN+SDAT+ELS	65.9	69.2	71.2	70.5	69.0	**69.5**	69.7	69.5	**69.5**	59.0	68.2	63.8	57.8
	ELS(ICLR2023)	CDAN+MCC+SDAT+ELS	65.6	68.5	69.4	68.9	68.5	68.6	68.6	68.6	68.6	59.6	58.0	57.7	57.8
	xMUDA (CVPR2020)	Text-only	64.7	72.7	75.1	73.9	72.6	66.5	66.8	66.5	66.3	55.0	67.7	60.8	52.4
		Image-only	63.5	70.4	70.4	70.6	70.4	62.1	63.7	62.2	61.1	57.9	55.7	55.3	55.2
		Fusion	66.9	73.9	76.4	75.1	73.7	67.5	68.1	67.6	67.3	59.4	57.9	57.8	57.9
	DsCML (ICCV2021)	Text-only	68.4	76.3	77.4	77.0	76.3	69.1	69.2	69.2	69.2	59.9	70.9	65.0	58.4
		Image-only	53.7	53.3	54.0	53.9	53.2	54.7	55.2	54.8	53.7	53.0	50.5	50.5	50.3
		Fusion	68.0	74.4	76.3	75.3	74.2	69.0	68.6	68.6	68.6	60.6	71.3	65.6	59.2
MSDA	MDAN (NeurIPS2018)	–	66.9	73.6	75.9	74.7	73.4	69.1	69.2	69.0	69.0	58.1	65.9	62.6	57.0
	M^3SDA (ICCV2019)		67.5	73.1	77.2	74.6	72.7	67.2	68.7	67.3	66.6	62.1	64.6	64.4	62.1
	T-SVDNet (ICCV2021)		67.9	74.7	75.5	75.3	74.7	66.8	66.9	66.8	66.8	62.3	65.3	64.8	62.3
MSMMDA	M2CAN (Ours)	–	**73.2**	**77.7**	**78.1**	**78.2**	**77.9**	**69.5**	**70.0**	**69.6**	69.4	**72.3**	**73.6**	**74.0**	**72.2**
Oracle	Oracle	–	79.9	83.7	83.6	83.7	83.6	75.7	75.8	75.7	75.7	80.4	79.8	80.0	79.9

Table 5
Ablation study on main components in M2CAN, measured by accuracy (%). We conduct experiments on both sentiment analysis and aesthetics assessment tasks. The best result is emphasized in **bold**.

Module	→TE	→T	→Y	Avg.	→A	→P	→R	Avg.
w/o $\mathcal{L}_{CMCFA}$	63.0	80.5	58.3	67.3	78.3	67.8	71.0	72.4
w/ $\mathcal{L}_{CMCFA}$, w/o Dynamic	62.8	82.1	58.8	67.9	76.7	69.0	73.6	73.1
w/o $\mathcal{L}_{CDCFA}$	62.3	82.0	59.9	68.1	78.4	69.2	73.4	73.7
w/o $\mathcal{L}_{CDAFA}$	62.3	84.5	63.0	69.9	78.3	67.8	74.4	73.5
w/o $\mathcal{L}_{MCC}$	63.4	83.9	59.9	69.1	78.5	68.1	71.2	72.6
w/o UACR	52.8	66.9	47.9	55.9	66.7	65.3	61.4	64.5
M2CAN	**63.8**	**84.7**	**63.7**	**70.7**	**79.9**	**69.8**	**74.5**	**74.7**

Table 6
Ablation study on alternative components in M2CAN, measured by accuracy (%). We conduct experiments on both sentiment analysis and aesthetics assessment tasks. The best result is emphasized in **bold**.

Type	Detail	→TE	→T	→Y	Avg.	→A	→P	→R	Avg.
Feature Fusion	concat instead	62.3	83.1	63.6	69.7	79.0	69.0	69.7	72.6
	add instead	63.2	83.4	58.4	68.3	79.1	69.4	67.7	72.1
Pseudo Label Selection	classification score	62.4	**84.7**	59.4	68.8	78.5	67.5	71.7	72.6
	uncertainty score	62.4	78.7	59.7	66.9	77.7	67.5	69.2	71.5
Warm-up Method	source-combined	63.2	79.3	59.8	67.4	78.3	67.9	71.0	72.4
Full	M2CAN	**63.8**	**84.7**	**63.7**	**70.7**	**79.9**	**69.8**	**74.5**	**74.7**

4.3. Ablation study

We conduct a series of ablation experiments based on ResNet50+BERT on both the sentiment analysis and aesthetics assessment tasks to demonstrate the effectiveness of the different components within M2CAN. Our ablation study is organized as follows: (1) analysis of the effectiveness of individual components; (2) sensitivity analysis of hyperparameters.

Analysis of the effectiveness of individual components. The results of our ablation study on the components of the proposed M2CAN are presented in Tables 5 and 6.

(1) The proposed three feature-level alignments, *i.e.*, CMCFA, CDCFA, and CDAFA are all effective. When we omit any one of these three alignment terms, there is a drop in accuracy. CMCFA aims to perform contrastive alignment between different modalities within each domain to reduce the multi-modal gap. After removing it, the gap between multi-modalities is not well alleviated, resulting in an average accuracy decrease of 3.4% and 2.3% in sentiment analysis and aesthetics assessment, respectively. Especially in sentiment analysis, there are many cases of modal mismatch. CDCFA aims to perform contrastive alignment between different domains of each modality to reduce the multi-domain gap. When it is removed, the average accuracy decreases by 2.6% and 1.0% on two tasks, respectively. This is

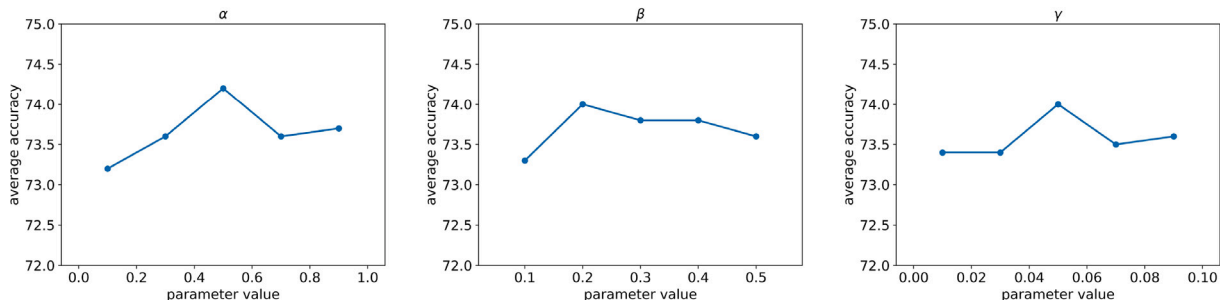


Fig. 11. Ablation study on the hyperparameters in M2CAN on the aesthetics assessment task, measured by average accuracy (%), including loss weights α, β, and γ.

due to the gap between different domains within the same modality, such as the different comment styles in sentiment analysis, which is not well considered. CDAFA aims to perform adversarial alignment on fused multi-modal representations between multiple domains to alleviate the domain gap in multi-modal features. When it is excluded, the average accuracy decreases by 0.8% and 1.2%, respectively. This is because the final task classification is performed on multi-modal features, and ignoring domain gaps on these features can result in sub-optimal performance.

(2) We observe that the dynamic $\mathcal{L}_{CMCFA}$ is highly effective. If dynamic $\mathcal{L}_{CMCFA}$ is removed and only ordinary $\mathcal{L}_{CMCFA}$ is used, the average accuracy will decrease by 2.8% in sentiment analysis and 1.6% in aesthetics assessment. Because, in multi-modal settings, modality mismatch is a pervasive challenge, and in multi-source settings, this issue arises not just within a single domain but also across domains. Therefore, the proposed dynamic $\mathcal{L}_{CMCFA}$ adaptively adjusts the cross-modal contrastive feature alignment loss based on the differences predicted by individual headers of each modality, alleviating the interference of modality mismatch.

(3) After removing MCC [103], the average accuracy decreases by 1.6% and 2.1%, respectively. This is because MCC can alleviate class confusion and help the model better identify ambiguous samples.

(4) The ablation study demonstrates that UACR is effective, which utilizes both classification score and uncertainty score to select target samples to perform self-learning and join the alignments with source samples. If all target domain samples are directly aligned with the source domain samples in the early stage of model training, the samples with huge differences between the target domain and the source domains will affect the domain adaptation process of the model. Therefore, the performance decreases substantially.

(5) When we replace the multi-modal fusion module MLB [111] with concatenation or addition, the performance decreases, indicating that better multi-modal fusion modules are beneficial for the MSMMDA task. However, the performance degradation is not noticeable, which also indicates that our M2CAN exhibits robustness to different fusion modules. Using classification score or uncertainty score alone to select pseudo labels leads to a decrease in performance, which indicates that the strategy of integrating the two in Eq. (18) is effective. When we change the warm-up method before generating pseudo labels for the target domain from aligned between source domains in M2CAN to source combined without alignment, the model performance also decreases. This is because there are domain gaps between source domains.

Sensitivity analysis of hyperparameters. We conduct an ablation study on the aesthetics assessment task, focusing on the loss weights α, β, and γ. The experimental results are depicted in Fig. 11. As observed, M2CAN remains relatively robust to changes in hyperparameters and most of the time does not particularly affect the results, with an accuracy difference limited to 1%.

4.4. Visualization

We conduct case studies on six datasets based on ResNet50 + BERT, including source-only method, UDA method CDAN [7]+ELS [8], MMDA method DsCML [3], MSDA method MDAN [81], and the proposed M2CAN, as shown in Fig. 12. From the case study, these phenomena can be observed: (1) The huge gap between multiple multi-modal domains can lead to a lack of good generalization of the model. For example, in sentiment analysis tasks, neutrality is easily misjudged. In TumEmo [44], neutrality is often explicitly expressed, in T4SA [45], neutrality is expressed in a plain manner, while in Yelp [46], neutrality is usually caused by mixed reviews. Our method, benefiting from the joint effect of four feature-level and label space-level alignments, reduces the multi-modal gap and multi-domain gap in the MSMMDA task, achieves better generalization, and can more accurately identify difficult-to-distinguish examples. (2) Modal mismatch is common in the MSMMDA task. For example, in the first sample of → T4SA, the image expresses positive sentiment, while the text expresses negative sentiment. In the second sample from → PCCD, the image is of high quality, but the evaluation tends to be negative. The proposed dynamically adjusted CMCFA effectively alleviates this issue by allowing the model to first learn samples with consistent modalities to achieve initial convergence, and then learn samples with inconsistent modalities.

We further visualize the features of source and target samples before and after adaptation using different DA methods based on ResNet50+BERT. Using t-SNE [113] to reduce the dimension of the samples' multi-modal fused features, we plot the learned features on a 2-dimensional plane, as depicted in Fig. 13. From the visualizations, it is evident that, before adaptation (Source-only), the features of source and target domains can be discriminated because of the existence of domain gap (the distribution areas of red, blue, and green points are easily distinguished in Fig. 13(a)). After domain adaptation by UDA in Fig. 13(b), MMDA in Fig. 13(c) and MSDA in Fig. 13(d), the gaps between multiple source domains and target domains are alleviated. However, these methods still cannot achieve satisfactory results due to the lack of comprehensive consideration of the multiple modal gaps and domain gaps in the MSMMDA tasks. M2CAN in Fig. 13(e) effectively aligns the multi-modal fusion feature space between multiple source and target domains, reducing the modal and domain gaps between them. Meanwhile, it also maintains the separability between classes, demonstrating the superiority of M2CAN over the MSMMDA task.

4.5. Limitation discussion

Because of the absence of datasets collected for the MSMMDA task, we only verify the effectiveness of the proposed M2CAN on two combined datasets with three different domains for visual-textual classification tasks. But in reality, M2CAN can be easily applied to other tasks. Meanwhile, M2CAN can be viewed as a strong baseline for future research on MSMMDA. Some parts of our proposed framework can be replaced with more powerful modules, such as task-related backbones and losses. It is easy to extend MSMMDA to other tasks such as object detection and semantic segmentation, if there are relevant datasets available.

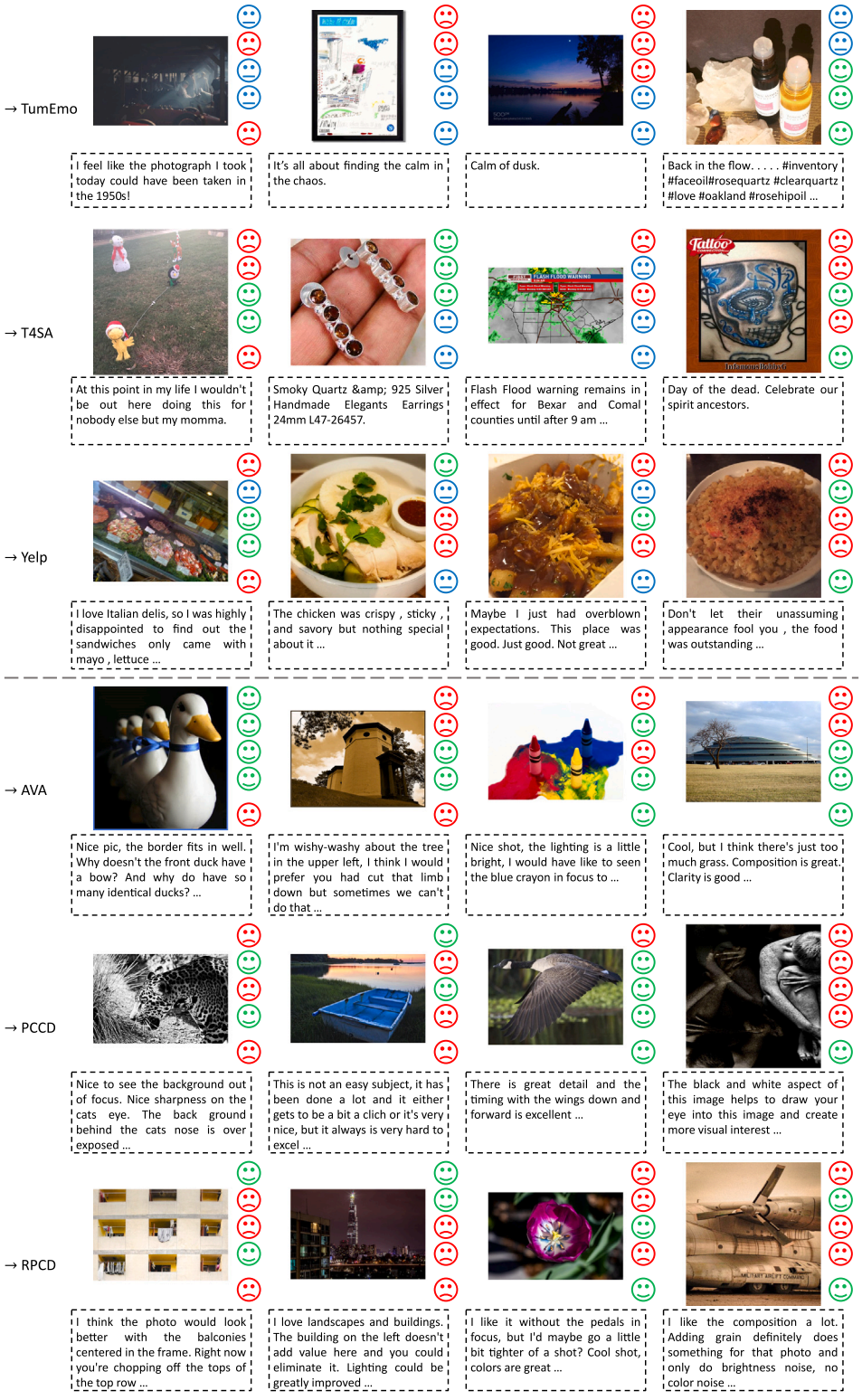
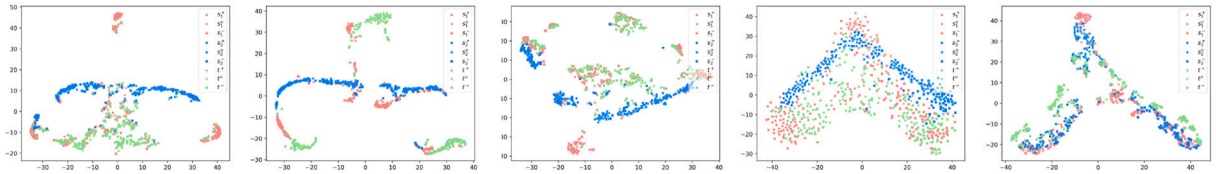


Fig. 12. Sample analysis. Next to each sample, from top to bottom, are the predicted results of Source-only, CDAN [7]+ELS [8] (UDA), DsCML [3] (MMDA), MDAN [81] (MSDA), and M2CAN (Ours)/Ground Truth using ResNet50+BERT as the feature extractor. Part of the text in the sample is omitted.

5. Conclusion

In this paper, we study a novel and practical domain adaptation setting: multi-source multi-modal domain adaptation (MSMMDA). To address the modal gaps and domain gaps in the MSMMDA task, we propose a Multi-source Multi-modal Contrastive Adversarial Network (M2CAN) to align the multiple source and target domains on both feature level and label space level. M2CAN learns domain-invariant multi-modal representations by three different feature-level alignment strategies: cross-modal contrastive feature alignment (CMCFA) within each domain, cross-domain contrastive feature alignment (CDCFA) for each modality, and cross-domain adversarial feature alignment (CDAFA) on the fused multi-modal representations. Further, M2CAN conducts



(a) Source-only (b) CDAN+ELS (UDA) (c) DsCML (MMDA) (d) MDAN (MSDA) (e) M2CAN (Ours)

Fig. 13. t-SNE visualization of DA methods on →T4SA settings in the sentiment analysis task. Red and blue respectively represent two source domain features and green represents target domain features. Dots, crosses, and squares represent positive, neutral, and negative categories, respectively. For example, s_1^+ represents the positive samples in source domain S_1.

a label space-level alignment, uncertainty-aware classifier refinement (UACR), which generates and selects pseudo labels in the target domain to perform self-learning and participate in the alignments with source domains. After such alignment, both the source and target domains are mapped into a shared multi-modal representation space and the trained task classifiers can be better adapted to the target domain. Extensive experiments on two tasks, *i.e.*, sentiment analysis and aesthetics assessment, demonstrate the superiority of M2CAN over the previous state-of-the-art DA methods.

CRediT authorship contribution statement

Sicheng Zhao: Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jing Jiang:** Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation. **Wenbo Tang:** Writing – review & editing, Software, Funding acquisition, Formal analysis, Data curation. **Jiankun Zhu:** Writing – review & editing, Validation, Software, Data curation. **Hui Chen:** Writing – review & editing, Validation, Resources, Conceptualization. **Pengfei Xu:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Conceptualization. **Björn W. Schuller:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization. **Jianhua Tao:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Hongxun Yao:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Guiguang Ding:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by CCF-DiDi GAIA Collaborative Research Funds, China and the National Natural Science Foundation of China (Nos. U21B2010, 62441202).

Data availability

Data will be made available on request.

References

[1] Y. Wang, Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion, ACM Trans. Multimed. Comput. Commun. Appl. 17 (1s) (2021) 1–25.

[2] M. Jaritz, T.-H. Vu, R.d. Charette, E. Wirbel, P. Pérez, Xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 12605–12614.

[3] D. Peng, Y. Lei, W. Li, P. Zhang, Y. Guo, Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation, in: IEEE International Conference on Computer Vision, 2021, pp. 7108–7117.

[4] S. Zhao, G. Jia, J. Yang, G. Ding, K. Keutzer, Emotion recognition from multiple modalities: Fundamentals and methodologies, IEEE Signal Process. Mag. 38 (6) (2021) 59–73.

[5] S. Swetha, M.N. Rizve, N. Shvetsova, H. Kuehne, M. Shah, Preserving modality structure improves multi-modal learning, in: IEEE International Conference on Computer Vision, 2023, pp. 21993–22003.

[6] R. Shao, T. Wu, Z. Liu, Detecting and grounding multi-modal media manipulation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 6904–6913.

[7] M. Long, Z. Cao, J. Wang, M.I. Jordan, Conditional adversarial domain adaptation, in: Advances in Neural Information Processing Systems, 2018, pp. 1640–1650.

[8] Y. Zhang, J. Liang, Z. Zhang, L. Wang, R. Jin, T. Tan, et al., Free lunch for domain adversarial training: Environment label smoothing, in: International Conference on Learning Representations, 2023.

[9] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, B. Wang, Moment matching for multi-source domain adaptation, in: IEEE International Conference on Computer Vision, 2019, pp. 1406–1415.

[10] S. Zhao, X. Yue, S. Zhang, B. Li, H. Zhao, B. Wu, R. Krishna, J.E. Gonzalez, A.L. Sangiovanni-Vincentelli, S.A. Seshia, et al., A review of single-source deep unsupervised visual domain adaptation, IEEE Trans. Neural Netw. Learn. Syst. 33 (2) (2022) 473–493.

[11] L. Yang, Y. Balaji, S.-N. Lim, A. Shrivastava, Curriculum manager for source selection in multi-source domain adaptation, in: European Conference on Computer Vision, 2020, pp. 608–624.

[12] T. Xu, W. Chen, W. Pichao, F. Wang, H. Li, R. Jin, CDTrans: Cross-domain transformer for unsupervised domain adaptation, in: International Conference on Learning Representations, 2021.

[13] J. Zhu, H. Bai, L. Wang, Patch-mix transformer for unsupervised domain adaptation: A game perspective, in: IEEE Conference on Computer Vision and Pattern Recognition, 2023, pp. 3561–3571.

[14] M. Long, H. Zhu, J. Wang, M.I. Jordan, Deep transfer learning with joint adaptation networks, in: International Conference on Machine Learning, 2017, pp. 2208–2217.

[15] Y.-W. Luo, C.-X. Ren, Conditional bures metric for domain adaptation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 13989–13998.

[16] Z. Pei, Z. Cao, M. Long, J. Wang, Multi-adversarial domain adaptation, in: AAAI Conference on Artificial Intelligence, 2018, pp. 3934–3941.

[17] Y. Du, Z. Tan, Q. Chen, X. Zhang, Y. Yao, C. Wang, Dual adversarial domain adaptation, 2020, arXiv:2001.00153.

[18] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, T. Darrell, Cycada: Cycle-consistent adversarial domain adaptation, in: International Conference on Machine Learning, Pmlr, 2018, pp. 1989–1998.

[19] T. Kim, M. Jeong, S. Kim, S. Choi, C. Kim, Diversify and match: A domain adaptive representation learning paradigm for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12456–12465.

[20] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, D. Erhan, Domain separation networks, in: Advances in Neural Information Processing Systems, 2016, pp. 343–351.

[21] X. Chen, H. Li, C. Zhou, X. Liu, D. Wu, G. Dudek, Fido: Ubiquitous fine-grained wifi-based localization for unlabelled users via domain adaptation, in: The Web Conference, 2020, pp. 23–33.

[22] Y. Wang, S. Qiu, D. Li, C. Du, B.-L. Lu, H. He, Multi-modal domain adaptation variational autoencoder for eeg-based emotion recognition, IEEE J. Automat. Sin. 9 (9) (2022) 1612–1626.

[23] S. Hu, F. Bonardi, S. Bouchafa, D. Sidibé, Multi-modal unsupervised domain adaptation for semantic image segmentation, Pattern Recognit. 137 (2023) 109299.

[24] S. Zhao, G. Wang, S. Zhang, Y. Gu, Y. Li, Z. Song, P. Xu, R. Hu, H. Chai, K. Keutzer, Multi-source distilling domain adaptation, in: AAAI Conference on Artificial Intelligence, 2020, pp. 12975–12983.

[25] R. Xu, Z. Chen, W. Zuo, J. Yan, L. Lin, Deep cocktail network: Multi-source unsupervised domain adaptation with category shift, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[26] S. Rakshit, B. Banerjee, G. Roig, S. Chaudhuri, Unsupervised multi-source domain adaptation driven by deep adversarial ensemble learning, Pattern Recogn. (2019) 485–498.

[27] P. Russo, T. Tommasi, B. Caputo, Towards multi-source adaptive semantic segmentation, in: International Conference on Image Analysis and Processing, 2019, pp. 292–301.

[28] D.J. Shah, Multi-Source Domain Adaptation with Mixture of Experts (Ph.D. thesis), Massachusetts Institute of Technology, 2019.

[29] X. Yao, S. Zhao, P. Xu, J. Yang, Multi-source domain adaptation for object detection, in: IEEE/CVF International Conference on Computer Vision, 2021, pp. 3273–3282.

[30] R. Li, X. Jia, J. He, S. Chen, Q. Hu, T-svdnet: Exploring high-order prototypical correlations for multi-source domain adaptation, in: IEEE International Conference on Computer Vision, 2021, pp. 9991–10000.

[31] D. Zhang, M. Ye, Y. Liu, L. Xiong, L. Zhou, Multi-source unsupervised domain adaptation for object detection, Inf. Fusion 78 (2022) 138–148.

[32] T. Gao, J. Yang, Q. Tang, A multi-source domain information fusion network for rotating machinery fault diagnosis under variable operating conditions, Inf. Fusion 106 (2024) 102278.

[33] J. Jiang, S. Zhao, J. Zhu, W. Tang, Z. Xu, J. Yang, P. Xu, H. Yao, Multi-source domain adaptation for panoramic semantic segmentation, 2024, arXiv:2408.16469.

[34] S. Zhao, H. Yao, C. Lin, Y. Gao, G. Ding, Multi-source-free domain adaptive object detection, Int. J. Comput. Vis. (2024).

[35] H. Guo, R. Pasunuru, M. Bansal, Multi-source domain adaptation for text classification via distancenet-bandits, in: AAAI Conference on Artificial Intelligence, 2020, pp. 7830–7838.

[36] Z. Chen, P. Wei, J. Zhuang, G. Li, L. Lin, Deep CockTail networks: A universal framework for visual multi-source domain adaptation, Int. J. Comput. Vis. 129 (8) (2021) 2328–2351.

[37] C. Lin, S. Zhao, L. Meng, T.-S. Chua, Multi-source domain adaptation for visual sentiment classification, in: AAAI Conference on Artificial Intelligence, 2020, pp. 2661–2668.

[38] S. Zhao, B. Li, P. Xu, X. Yue, G. Ding, K. Keutzer, MADAN: Multi-source adversarial domain aggregation network for domain adaptation, Int. J. Comput. Vis. 129 (8) (2021) 2399–2424.

[39] S. Zhao, Y. Xiao, J. Guo, X. Yue, J. Yang, R. Krishna, P. Xu, K. Keutzer, Curriculum cyclegan for textual sentiment domain adaptation with multiple sources, in: The Web Conference, 2021, pp. 541–552.

[40] N. Venkat, J.N. Kundu, D. Singh, A. Revanur, et al., Your classifier can secretly suffice multi-source domain adaptation, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 4647–4659.

[41] K. Li, J. Lu, H. Zuo, G. Zhang, Dynamic classifier alignment for unsupervised multi-source domain adaptation, IEEE Trans. Knowl. Data Eng. 35 (5) (2022) 4727–4740.

[42] A. Belal, A. Meethal, F.P. Romero, M. Pedersoli, E. Granger, Multi-source domain adaptation for object detection with prototype-based mean teacher, in: IEEE Winter Conference on Applications of Computer Vision, 2024, pp. 1277–1286.

[43] X. Ma, T. Zhang, C. Xu, Deep multi-modality adversarial networks for unsupervised domain adaptation, IEEE Trans. Multimed. 21 (9) (2019) 2419–2431.

[44] X. Yang, S. Feng, D. Wang, Y. Zhang, Image-text multimodal emotion classification via multi-view attentional network, IEEE Trans. Multimed. 23 (2020) 4014–4026.

[45] L. Vadicamo, F. Carrara, A. Cimino, S. Cresci, F. Dell'Orletta, F. Falchi, M. Tesconi, Cross-media learning for image sentiment analysis in the wild, in: IEEE International Conference on Computer Vision Workshops, 2017, pp. 308–317.

[46] Q.-T. Truong, H.W. Lauw, Vistanet: Visual aspect attention network for multimodal sentiment analysis, in: AAAI Conference on Artificial Intelligence, 2019, pp. 305–312.

[47] Y. Zhou, X. Lu, J. Zhang, J.Z. Wang, Joint image and text representation for aesthetics analysis, in: ACM International Conference on Multimedia, 2016, pp. 262–266.

[48] K.-Y. Chang, K.-H. Lu, C.-S. Chen, Aesthetic critiques generation for photos, in: IEEE International Conference on Computer Vision, 2017, pp. 3514–3523.

[49] D. Vera Nieto, L. Celona, C. Fernandez Labrador, Understanding aesthetics with language: A photo critique dataset for aesthetic assessment, in: Advances in Neural Information Processing Systems, 2022, pp. 34148–34161.

[50] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443.

[51] S.K. D'mello, J. Kory, A review and meta-analysis of multimodal affect detection systems, ACM Comput. Surv. 47 (3) (2015) 1–36.

[52] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, in: Asian Conference on Computer Vision, 2017, pp. 213–228.

[53] Y. Sun, W. Zuo, M. Liu, RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes, IEEE Robot. Autom. Lett. 4 (3) (2019) 2576–2583.

[54] H. Sun, J. Liu, Y.-W. Chen, L. Lin, Modality-invariant temporal representation learning for multimodal sentiment classification, Inf. Fusion 91 (2023) 504–514.

[55] Y. Oh, S. Kim, Multi-modal lifelog data fusion for improved human activity recognition: A hybrid approach, Inf. Fusion 110 (2024) 102464.

[56] S. Xiong, G. Zhang, V. Batra, L. Xi, L. Shi, L. Liu, TRIMOON: Two-round inconsistency-based multi-modal fusion network for fake news detection, Inf. Fusion 93 (2023) 150–158.

[57] Y. Lin, D. Guo, Y. Wu, L. Li, E.Q. Wu, W. Ge, Fuel consumption prediction for pre-departure flights using attention-based multi-modal fusion, Inf. Fusion 101 (2024) 101983.

[58] C. Lu, J. Yin, H. Yang, S. Sun, Enhancing multi-modal fusion in visual dialog via sample debiasing and feature interaction, Inf. Fusion 107 (2024) 102302.

[59] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: International Conference on Machine Learning, 2015, pp. 97–105.

[60] D. Mekhazni, A. Bhuiyan, G. Ekladious, E. Granger, Unsupervised domain adaptation in the dissimilarity space for person re-identification, in: European Conference on Computer Vision, 2020, pp. 159–174.

[61] H. Huang, Q. Liu, Domain structure-based transfer learning for cross-domain word representation, Inf. Fusion 76 (2021) 145–156.

[62] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, A. Smola, A kernel method for the two-sample-problem, in: Advances in Neural Information Processing Systems, 2006, pp. 513–520.

[63] B. Sun, J. Feng, K. Saenko, Return of frustratingly easy domain adaptation, in: AAAI Conference on Artificial Intelligence, 2016, pp. 2058–2065.

[64] B. Sun, K. Saenko, Deep coral: Correlation alignment for deep domain adaptation, in: European Conference on Computer Vision, 2016, pp. 443–450.

[65] Y. Wang, W. Li, D. Dai, L. Van Gool, Deep domain adaptation by geodesic distance minimization, in: IEEE International Conference on Computer Vision Workshops, 2017, pp. 2651–2657.

[66] P. Morerio, V. Murino, Correlation alignment by riemannian metric for domain adaptation, 2017, arXiv:1705.08180.

[67] Y. Zhang, N. Wang, S. Cai, L. Song, Unsupervised domain adaptation by mapped correlation alignment, IEEE Access 6 (2018) 44698–44706.

[68] C. Chen, Z. Fu, Z. Chen, S. Jin, Z. Cheng, X. Jin, X.-S. Hua, Homm: Higher-order moment matching for unsupervised domain adaptation, in: AAAI Conference on Artificial Intelligence, 2020, pp. 3422–3429.

[69] G. Kang, L. Jiang, Y. Yang, A.G. Hauptmann, Contrastive adaptation network for unsupervised domain adaptation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4893–4902.

[70] C. Li, N. Bian, Z. Zhao, H. Wang, B.W. Schuller, Multi-view domain-adaptive representation learning for EEG-based emotion recognition, Inf. Fusion 104 (2024) 102156.

[71] A. Qayyum, I. Razzak, M. Mazher, X. Lu, S.A. Niederer, Unsupervised unpaired multiple fusion adaptation aided with self-attention generative adversarial network for scar tissues segmentation framework, Inf. Fusion 106 (2024) 102226.

[72] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2107–2116.

[73] G. Kang, L. Zheng, Y. Yan, Y. Yang, Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization, in: European Conference on Computer Vision, 2018, pp. 401–416.

[74] M. Ghifary, W.B. Kleijn, M. Zhang, D. Balduzzi, W. Li, Deep reconstruction-classification networks for unsupervised domain adaptation, in: European Conference on Computer Vision, 2016, pp. 597–613.

[75] P. Singhal, R. Walambe, S. Ramanna, K. Kotecha, Domain adaptation: Challenges, methods, datasets, and applications, IEEE Access 11 (2023) 6973–7020.

[76] H. Zhang, S. Qian, Q. Fang, C. Xu, Multimodal disentangled domain adaption for social media event rumor detection, IEEE Trans. Multimed. 23 (2020) 4441–4454.

[77] H. Li, Y. Kim, C.-H. Kuo, S. Narayanan, Acted vs. improvised: Domain adaptation for elicitation approaches in audio-visual emotion recognition, in: International Conference on Spoken Language Processing, 2021.

[78] F. Qi, X. Yang, C. Xu, A unified framework for multimodal domain adaptation, in: ACM International Conference on Multimedia, 2018, pp. 429–437.

[79] J. Munro, D. Damen, Multi-modal domain adaptation for fine-grained action recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 122–132.

[80] S. Zhao, H. Chen, H. Hu, P. Xu, G. Ding, More is better: Deep domain adaptation with multiple sources, in: International Joint Conference on Artificial Intelligence, 2024, pp. 8354–8362.

[81] H. Zhao, S. Zhang, G. Wu, J.M. Moura, J.P. Costeira, G.J. Gordon, Adversarial multiple source domain adaptation, in: Advances in Neural Information Processing Systems, 2018, pp. 8568–8579.
[82] M.-Y. Liu, O. Tuzel, Coupled generative adversarial networks, in: Advances in Neural Information Processing Systems, 2016, pp. 469–477.
[83] S. Zhao, B. Li, X. Yue, Y. Gu, P. Xu, R. Hu, H. Chai, K. Keutzer, Multi-source domain adaptation for semantic segmentation, in: Advances in Neural Information Processing Systems, 2019, pp. 7285–7298.
[84] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[85] J. He, X. Jia, S. Chen, J. Liu, Multi-source domain adaptation with collaborative learning for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 11008–11017.
[86] Y. Zhu, F. Zhuang, D. Wang, Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources, in: AAAI Conference on Artificial Intelligence, 2019, pp. 5989–5996.
[87] V.-A. Nguyen, T. Nguyen, T. Le, Q.H. Tran, D. Phung, Stem: An approach to multi-source domain adaptation with guarantees, in: IEEE International Conference on Computer Vision, 2021, pp. 9352–9363.
[88] P. Karisani, Multiple-source domain adaptation via coordinated domain encoders and paired classifiers, in: AAAI Conference on Artificial Intelligence, 2022, pp. 7087–7095.
[89] K. Li, J. Lu, H. Zuo, G. Zhang, Multi-source contribution learning for domain adaptation, IEEE Trans. Neural Netw. Learn. Syst. 33 (10) (2021) 5293–5307.
[90] Z. Wang, C. Zhou, B. Du, F. He, Self-paced supervision for multi-source domain adaptation., in: International Joint Conference on Artificial Intelligence, 2022, pp. 3551–3557.
[91] L. Zhou, M. Ye, D. Zhang, C. Zhu, L. Ji, Prototype-based multisource domain adaptation, IEEE Trans. Neural Netw. Learn. Syst. 33 (10) (2021) 5308–5320.
[92] Y.-H. Liu, C.-X. Ren, A two-way alignment approach for unsupervised multi-source domain adaptation, Pattern Recognit. 124 (2022) 108430.
[93] R. Xia, C. Zong, X. Hu, E. Cambria, Feature ensemble plus sample selection: domain adaptation for sentiment classification, IEEE Intell. Syst. 28 (3) (2013) 10–18.
[94] J. Li, K. Lu, Z. Huang, L. Zhu, H.T. Shen, Transfer independently together: A generalized framework for domain adaptation, IEEE Trans. Cybern. 49 (6) (2018) 2144–2155.
[95] O. Sener, H.O. Song, A. Saxena, S. Savarese, Learning transferrable representations for unsupervised domain adaptation, in: Advances in Neural Information Processing Systems, 2016, pp. 2110–2118.
[96] K. Saito, Y. Ushiku, T. Harada, Asymmetric tri-training for unsupervised domain adaptation, in: International Conference on Machine Learning, 2017, pp. 2988–2997.
[97] W. Zhang, W. Ouyang, W. Li, D. Xu, Collaborative and adversarial network for unsupervised domain adaptation, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3801–3809.
[98] J. Choi, M. Jeong, T. Kim, C. Kim, Pseudo-labeling curriculum for unsupervised domain adaptation, in: British Machine Vision Conference, 2019, p. 67.
[99] Q. Wang, T. Breckon, Unsupervised domain adaptation via structured prediction based selective pseudo-labeling, in: AAAI Conference on Artificial Intelligence, 2020, pp. 6243–6250.
[100] V.M. Patel, R. Gopalan, R. Li, R. Chellappa, Visual domain adaptation: A survey of recent advances, IEEE Signal Process. Mag. 32 (3) (2015) 53–69.
[101] A.v.d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, 2018, arXiv:1807.03748.
[102] Z. Zheng, Y. Yang, Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation, Int. J. Comput. Vis. 129 (4) (2021) 1106–1120.
[103] Y. Jin, X. Wang, M. Long, J. Wang, Minimum class confusion for versatile domain adaptation, in: European Conference on Computer Vision, 2020, pp. 464–480.
[104] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019, arXiv:1907.11692.
[105] D. Loureiro, F. Barbieri, L. Neves, L.E. Anke, J. Camacho-Collados, Timelms: Diachronic language models from twitter, in: Annual Meeting of the Association for Computational Linguistics, 2022, pp. 251–260.
[106] H. Rangwani, S.K. Aithal, M. Mishra, A. Jain, V.B. Radhakrishnan, A closer look at smoothness in domain adversarial training, in: International Conference on Machine Learning, 2022, pp. 18378–18399.
[107] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[108] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171–4186.
[109] B. Zhang, P. Zhang, X. Dong, Y. Zang, J. Wang, Long-clip: Unlocking the long-text capability of clip, 2024, arXiv:2403.15378.
[110] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[111] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, B.-T. Zhang, Hadamard product for low-rank bilinear pooling, in: International Conference on Learning Representations, 2016.
[112] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: International Conference on Learning Representations, 2015.
[113] L.v.d. Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008) 2579–2605.