Learning Disentangled Representation for Multimodal Cross-Domain Sentiment Analysis

Yuhao Zhang, Ying Zhang, Wenya Guo, Xiangrui Cai, and Xiaojie Yuan

***Abstract*—Multimodal cross-domain sentiment analysis aims at transferring domain-invariant sentiment information across datasets to address the insufficiency of labeled data. Existing adaptation methods achieve well performance by remitting the discrepancies in characteristics of multiple modalities. However, the expressive styles of different datasets also contain domain-specific information, which hinders the adaptation performance. In this article, we propose a disentangled sentiment representation adversarial network (DiSRAN) to reduce the domain shift of expressive styles for multimodal cross-domain sentiment analysis. Specifically, we first align the multiple modalities and obtain the joint representation through a cross-modality attention layer. Then, we disentangle sentiment information from the multimodal joint representation that contains domain-specific expressive style by adversarial training. The obtained sentiment representation is domain-invariant, which can better facilitate the sentiment information transfer between different domains. Experimental results on two multimodal cross-domain sentiment analysis tasks demonstrate that the proposed method performs favorably against state-of-the-art approaches.**

***Index Terms*—Adversarial learning, disentangled representation learning, domain adaptation, multimodal sentiment analysis.**

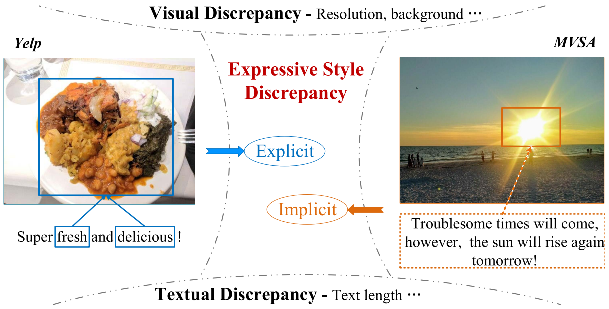
Visual Discrepancy - Resolution, background ···
Yelp
MVSA
Expressive Style Discrepancy
Explicit
Implicit
Super fresh and delicious !
Troublesome times will come, however, the sun will rise again tomorrow!
Textual Discrepancy - Text length ···

Fig. 1. Example of discrepancies (i.e., domain shifts) between the instances from *Yelp* (left) and *MVSA* (right) datasets. Besides the discrepancies in image background, resolution, and text length, the expressive styles are also discrepant across different datasets. The instance from the *Yelp* dataset explicitly conveys the positive sentiment toward the food in the image by the two words fresh and delicious. The instance from the *MVSA* dataset expresses the same sentiment with implicit literary knowledge.

I. Introduction

Sentiment analysis has broad applications in building recommendation systems, modeling customer preferences, and monitoring user behaviors [1]–[4]. Driven by these promising applications, extensive studies have been devoted to this task. Sentiment analysis can be formulated as a classification task that classifies the input into a sentiment category (e.g., positive, negative, and neutral). Traditional methods commonly focus on analyzing the sentiment of pure textual documents [5], [6]. With the development of smartphones and mobile Internet, nowadays, documents consist of multimodal data that contain more knowledge and richer underlying sentiment [7], [8]. Previous works focused on exploring the semantic interactions between modalities through machine learning methods [9]–[11]. Recent approaches have improved multimodal sentiment analysis performance with deep learning technologies [7], [12], [13].

The impressive performances of these deep learning methods rely on massive labeled training data. However, in many real-world applications, manual annotation is labor-intensive and time-consuming [14], making it difficult to obtain large-scale annotated sentiment datasets. The lack of sufficient labeled data has given rise to the task of cross-domain sentiment analysis [15], which aims to transfer domain-invariant sentiment information from the labeled datasets (the source domain) to the unlabeled one (the target domain). One straightforward solution is to train the model on the source domain and then directly apply the learned model to the target domain during testing. However, this direct transfer scheme typically results in performance degradation on the target domain because of the discrepancies between different datasets (i.e., the domain shifts) [14], [16]. Zhao *et al.* [17] proposed a weakly supervised deep embedding (WDE) framework for product review sentiment analysis. The proposed framework leverages the vast amount of review ratings as weak labels to pretrain an embedding layer. Then, a classifier is added on top of the embedding layer and fine-tuned with labeled review

Manuscript received February 7, 2021; revised July 10, 2021 and November 19, 2021; accepted January 20, 2022. This work was supported in part by the Chinese Scientific and Technical Innovation Project 2030 under Grant 2018AAA0102100, in part by the National Natural Science Foundation of China (NSFC)-Xinjiang Joint Fund under Grant U1903128, in part by the NSFC General Technology Joint Fund for Basic Research under Grant U1936206, and in part by the Fundamental Research Funds for the Central Universities, Nankai University, under Grant 63211128. *(Corresponding author: Ying Zhang.)*

Yuhao Zhang, Xiangrui Cai, and Xiaojie Yuan are with the Tianjin Key Laboratory of Network and Data Security Technology, College of Cyber Science, Nankai University, Tianjin 300350, China (e-mail: zhangyuhao@dbis.nankai.edu.cn; caixr@nankai.edu.cn; yuanxj@nankai.edu.cn).

Ying Zhang and Wenya Guo are with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: yingzhang@nankai.edu.cn; guowenya@dbis.nankai.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3147546.

Digital Object Identifier 10.1109/TNNLS.2022.3147546

sentences. Although WDE has been proven to be effective for product review sentiment analysis, the proposed idea of "weak pretraining + supervised fine-tuning" is not feasible for cross-domain sentiment analysis without available weak labels. Therefore, the imperative need for knowledge transfer across domains has given rise to domain adaptation studies.

Extensive domain adaptation studies have achieved great success in computer vision [18]–[20] and natural language processing [21]–[23]. However, knowledge transfer across multimodal datasets is more complicated due to the discrepancies brought by multiply modalities [24]. While "multimodal" could refer to various combinations of multiple modalities, we focus on the scenario that only contains images and text. Beyond the discrepancies in image background and resolution, and text length that are studied in the existing multimodal domain adaptation methods [24], [25], the style to express sentiment is also discrepant across datasets. As illustrated in Fig. 1, in the review from the *Yelp* dataset, the positive sentiment about the food in the image is explicitly expressed by the two words fresh and delicious. In contrast, the tweet from the multiview sentiment analysis (*MVSA*) dataset implicitly conveys the positive feeling with intrinsic literary knowledge. It can be observed that the samples from different domains express the same sentiment in different styles. This domain-specific expressive style should be removed when the sentiment information is transferred across different domains.

In this article, we propose a disentangled sentiment representation adversarial network (DiSRAN) to reduce the domain shift of expressive styles for multimodal cross-domain sentiment analysis. Specifically, first, for semantic alignment across different modalities, we employ a cross-modality attention layer in our feature embedding module. The attention layer reweights the single-modality features based on the knowledge from the other modality to obtain an aligned multimodal joint representation with rich multimodal semantic interactions. Second, we employ a sentiment embedding module to remit the discrepancy of expressive styles between different datasets. The sentiment embedding module disentangles sentiment information from the joint representation and learns a disentangled sentiment representation without domain-specific style information via adversarial training. Moreover, a reconstruction loss is applied to constrain the learning process of the sentiment embedding module and encourage the disentangled representation to carry as much useful information as possible. We evaluate the proposed method on *Yelp* and *MVSA* datasets, and experimental results show that our approach performs favorably against the state-of-the-art methods.

Our main contributions are summarized as follows.

1) We address the task of multimodal cross-domain sentiment analysis by a DiSRAN that remits the discrepancy of expressive styles between different datasets.
2) To transfer invariant sentiment information across different domains, we design a sentiment embedding module to disentangle the sentiment representation that does not contain style information.
3) Experimental results demonstrate that the proposed model outperforms the baseline methods on both multimodal single- and cross-domain sentiment classification tasks.

II. Related Work

A. Sentiment Analysis

Sentiment analysis is a rapidly developing task with promising real-world applications. The traditional formulation of sentiment analysis is text classification [5], which classifies the input document into a corresponding sentiment category [6]. Extensive lexicon- [26]–[28] and machine learning-based [29]–[31] textual sentiment classification methods have been well studied. Recently, deep learning approaches [1], [4], [32], [33] have significantly improved sentiment classification precision using abstract features. Specifically, Zhu *et al.* [33] introduced a kernel optimization function system called SentiVec for word embedding. With the integrated statistical information, sentiment similarity, and sentiment polarity scores in the embedded sentiment word vectors, SentiVec achieves better performances on the sentiment analysis task than the state-of-the-art methods, such as word2vec [34] and GloVe [35]. Luo *et al.* [4] proposed an interpretable framework FISHQA to hierarchically model the documents with the word, sentence, and document representations in multiple granularities. FISHQA significantly outperforms the compared methods in polarity identifications and produces meaningful evidence for the prediction results of financial documents.

The immense advancements in the sentiment analysis have encouraged researchers to explore more varied tasks in this domain. Extensive recent works [36], [37] have made great progress in aspect-based sentiment analysis, which is a fundamental task of sentiment analysis aiming to identify the sentiment polarity of a specific aspect in the context. A more challenging variation of the sentiment analysis task is visual sentiment classification because images are more abstract and subjective [38]. Previous works [39]–[41] used machine learning algorithms to extract low-level features from the input images and generated a predicted sentiment category through a classifier. Motivated by the effectiveness of deep neural networks on extracting abstract visual features, recent sentiment analysis studies spare no effort to explore deep visual representation learning [42], [43].

The boom in mobile Internet and social media has given rise to multimodal sentiment analysis [7], [8], [44]. Early works on this task are mainly based on the exploration of correlations between modalities. Chen *et al.* learned a visual–textual sentiment analysis model through a multimodal hypergraph [10]. Rasiwasia *et al.* [9] applied the canonical correlation analysis (CCA) to learn the correlations between visual and textual features. Li *et al.* [11] proposed a multimodal correlation model (MCM) that fully utilizes the hierarchical correlations between images and text for sentiment analysis.

Deep learning-based methods primarily follow the scheme that first extracts multimodal features and then conducts fusion algorithms for sentiment classification [45]. Zadeh *et al.* proposed an end-to-end multimodal sentiment classification method, termed the tensor fusion network (TFN), to dynamically model the intramodality and intermodality interactions in online videos. Since TFN is tailored for time series data of spoken words, gestures, and voices in videos, it is less effective in our image–text scenario. Therefore, extensive works have been devoted to the multimodal alignment of images and texts [7], [13]. Motivated by the intrinsic characteristics of the image–text sentiment correlations, Truong *et al.* developed a visual aspect attention network (VistaNet) that incorporated images as attention for a source of alignment with textual contents [7]. Gui *et al.* studied the problem of detecting depression in social media. They proposed a multiagent reinforcement learning method to incorporate textual and visual information [13].

B. Domain Adaptation

Over the past decades, domain adaptation has been widely explored to address the lack of labeled data [14]. Early methods are commonly divided into instance- and feature-based domain adaptations [16]. The first group of methods reduces the domain gap by reweighting training samples from the labeled source domain [46], [47]. The second group of methods learns shared feature spaces to minimize the discrepancy between different domains [48]–[51].

Recent studies resolve the domain shift by exploring embedding distribution matching. The maximum mean discrepancy (MMD) [52] is a widely used similarity measurement of different distributions. Tzeng *et al.* [53] proposed a domain adaptation network by minimizing the MMD distance of the source and target representations. Long *et al.* [54] employed a modified version of MMD with multiple kernels to learn transferable features. Lately, inspired by generative adversarial networks (GANs) [55], a variety of new network structures with embedded domain classifiers have been proposed [56], [57]. This group of methods plays a minimax game that the domain classifier is trained to distinguish between samples drawn from source and target domain datasets, while the feature encoder tries to deceive the classifier. Domain-invariant representations could be extracted through an adversarial training process, without access to target labels.

Despite the great achievement of domain adaptation, the aforementioned methods are primarily based on single-modality data. In the recent few years, several studies tend to focus on multimodal domain adaptation. Qi *et al.* proposed the multimodal domain adaptation neural network (MDANN) [24] to achieve state-of-the-art performance in the cross-domain audio–visual data emotion recognition task. Different from multimodal domain adaptation of videos in [24], Ma *et al.* introduced the multimodality adversarial network (MMAN) [25] to address the domain adaptation problem existing in the image–text-based data. The proposed MMAN yielded the best performance in the multimodal social event recognition task.

III. Methodology

A. Formulation and Notation

In this section, we formally define the multimodal cross-domain sentiment classification (MCDSC) task. We assume that there are two domains: D_s denotes the labeled source domain, and D_t denotes the unlabeled target domain. The two domains have different feature spaces, X_s and X_t, but share the same label space Y. That is to say, D_s and D_t have similar but different distributions that are shifted from each other by some domain shifts. The ultimate goal of the MCDSC task is to learn a sentiment classification function for D_t based on labeled source domain instances and unlabeled target domain instances. In other words, the learned model transfers knowledge from source to target domain via unsupervised domain adaptation. In this study, we use $\{X_s, Y\} = \{\boldsymbol{x}_i^s, \boldsymbol{y}_i^s\}_{i=1}^{N_s}$ and $X_t = \{\boldsymbol{x}_j^t\}_{j=1}^{N_t}$ to denote training samples from D_s and D_t, where N_s and N_t are the numbers of training samples for each domain, respectively. Every instance $\boldsymbol{x}$ consists of an image I and a piece of text T, i.e., $\boldsymbol{x} = \langle I, T\rangle$. $Y = \{1, 2, \ldots, K\}$ is a finite set that contains K sentiment categories.

B. Overview of DiSRAN

As shown in Fig. 2, our method consists of a feature embedding module and a sentiment embedding module. Inspired by [7], we apply the cross-modality attention in the feature embedding module to model a semantic-aligned joint representation $\boldsymbol{z}$. After that, we design the sentiment embedding module to remove the style information containing $\boldsymbol{z}$ and obtain a sentiment representation $\boldsymbol{s}$ without style information. During training, all the data from the source and target domains are simultaneously fed into the network for domain-adaptive learning with the domain classifiers. Annotations are only available for instances drawn from the source domain. The model is optimized according to the loss function denoted as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\text{domain}} + \gamma\mathcal{L}_{\text{rec}} \tag{1}$$

where θ is the parameter of our model, and β and γ are the hyperparameters for a tradeoff among losses.

C. Feature Embedding Module

In this section, we introduce the feature embedding module F in detail. F consists of a visual encoder, a textual encoder, and a cross-modality attention layer. It transforms the multimodal input $\boldsymbol{x}$ into the joint representation $\boldsymbol{z} = F(\boldsymbol{x}; \theta_f)$, where θ_f denotes the parameters of F.

1) Visual Encoder: For the input image I, we employ VGG-16 [58] to obtain the visual representation $\boldsymbol{v} = \text{VGG}(I)$, where $\boldsymbol{v} \in \mathbb{R}^{D_R}$ is drawn from the outputs of the last fully connected layer (FC7).

2) Textual Encoder: We employ NLTK [59] to tokenize the input text. Its vector representation $\text{Seq} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_L\}$ is initialized with pretrained word embedding from GloVe [35], where $\boldsymbol{e}_k \in \mathbb{R}^{D_e}$, D_e is the dimension of the embeddings, and L is the maximum length of the textual sequence. We feed the embedded sequence into the bidirectional recurrent neural network (BiRNN) with LSTM cells to obtain a hidden state vector $\boldsymbol{h} = \text{BiRNN}(\text{Seq})$, where $\boldsymbol{h} \in \mathbb{R}^{D_R}$. Since some words in a sequence are relatively meaningful [7], we use self-attention to emphasize those informative features in $\boldsymbol{h}$. The attended hidden vector is denoted as $\widetilde{\boldsymbol{h}}$.

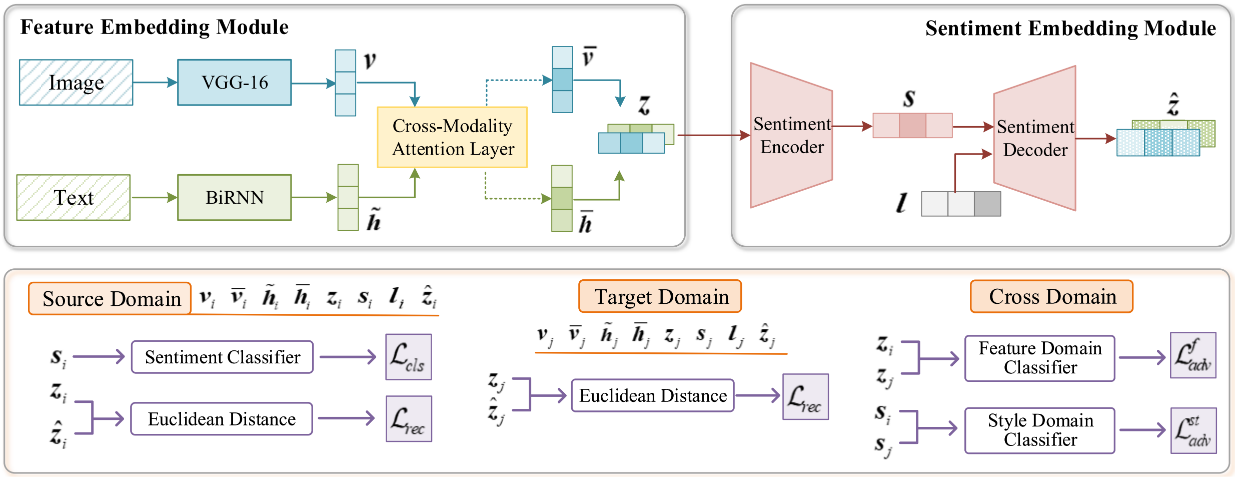

Fig. 2. Framework of our proposed DiSRAN. The feature embedding module consists of a visual encoder (i.e., the VGG-16), a textual encoder (i.e., the BiRNN), and a cross-modality attention layer. After feature embedding, the multimodal joint representation $\boldsymbol{z}$ is fed forward into the sentiment embedding module. The sentiment encoder learns a disentangled sentiment representation $\boldsymbol{s}$, and the sentiment decoder outputs a reconstructed joint representation $\hat{\boldsymbol{z}}$. During training, both labeled source domain instances and unlabeled target domain instances are simultaneously fed into the network for domain adaptive learning. Both the feature embedding module and the sentiment embedding module share weights for the two domains.

3) Cross-Modality Attention Layer: With the extracted features $\boldsymbol{v}$ and $\tilde{\boldsymbol{h}}$, we employ the cross-modality attention layer to obtain multimodal joint representations. Truong and Lauw [7] designed the visual aspect attention to highlight the semantic-related parts of a document based on the complementary visual knowledge and improved the sentiment classification precision with the attended document representation. Our cross-modality attention layer extends this idea to a bidirectional version for modeling semantic-aligned joint representations. The feature vector of each modality is reweighted by the attention weights, which are conditioned on the information from the other modality.

To obtain a reweighted visual representation $\bar{\boldsymbol{v}}$, we learn the text-guided attention weights $\boldsymbol{\alpha}^{\tilde{h},v}$ with respect to the textual representation $\tilde{\boldsymbol{h}}$

$$\boldsymbol{\alpha}^{\tilde{h},v} = \mathrm{Softmax}\Big(\sigma\Big(\mathrm{Conv}\big(\tilde{\boldsymbol{h}}, \boldsymbol{v}\big) + \boldsymbol{v}\Big)\Big) \tag{2}$$

where $\sigma(\cdot)$ denotes the ReLU activation function and $\mathrm{Softmax}(\cdot)$ normalizes the attention weights to 1. Compared with the visual aspect attention, we replace the elementwise multiplication with a 1-D convolutional layer $\mathrm{Conv}(\cdot)$ with the kernel size of 1 to extract multimodal features from stacked $\tilde{h}$ and v. Due to the expanded receptive field, our modified version can capture cross-dimension interactions between the visual and textual feature vectors. We compute the image-guided attention weights $\boldsymbol{\alpha}^{v,\tilde{h}} = \mathrm{Softmax}(\sigma(\mathrm{Conv}(\tilde{\boldsymbol{h}}, \boldsymbol{v}) + \tilde{\boldsymbol{h}}))$ in a similar way. Finally, the reweighted visual and textual representation $\bar{\boldsymbol{v}} = \{\boldsymbol{\alpha}_p^{\tilde{h},v} * \boldsymbol{v}_p\}_{p=1}^{D_R}$ and $\bar{\boldsymbol{h}} = \{\boldsymbol{\alpha}_q^{v,\tilde{h}} * \tilde{\boldsymbol{h}}_q\}_{q=1}^{D_E}$ are obtained via cross-modality attention. With $\bar{\boldsymbol{v}}$ and $\bar{\boldsymbol{h}}$, we obtain the joint representation $\boldsymbol{z}$ through concatenation.

D. Sentiment Embedding Module

In this section, we introduce the sentiment embedding module in detail. Inspired by the success in text style transfer [60], we design an encoder and a decoder, termed E and D, to disentangle the sentiment information from the joint representation. The embedded sentiment representation $\boldsymbol{s} = E(\boldsymbol{z}; \theta_E)$ only contains sentiment information. To achieve that goal, we employ C_{st} and C_{se} to constrain the learning process.

The style domain classifier C_{st} is designed to facilitate separating sentiment and style information. It tries to detect the intrinsic expressive style containing in $\boldsymbol{s}$. On the contrary, E manages to make the classifier unable to identify the style of the sentiment representation. E and C_{st} simultaneously learn under an adversarial object $\mathcal{L}_{\mathrm{adv}}^{\mathrm{st}}$.

The sentiment classifier C_{se} ensures that the learned sentiment representation is informative and discriminative for sentiment classification. Given on $\boldsymbol{s}$, C_{se} generates the predicted sentiment category. As demonstrated in the following equation, the training of sentiment classifier is formulated as minimizing a cross-entropy loss based on source domain instances and ground-truth labels

$$\mathcal{L}_{\mathrm{cls}}(\theta_f, \theta_E, \theta_y) = -\sum_{i}^{N_s} \boldsymbol{y}_i^s \log C_{\mathrm{se}}\big(\boldsymbol{s}_i^s; \theta_y\big) \tag{3}$$

where θ_E and θ_y represent the parameters of E and C_{se}, respectively.

Due to the lack of label observations of the target domain during training, the embedded sentiment representations of the target domain instances remain unconstrained. Inspired by [61], we employ the sentiment decoder D that helps to constrain the learning process. Based on the sentiment representation $\boldsymbol{s}$ and the embedded style label $\boldsymbol{l}$, the sentiment decoder generates a reconstructed joint representation $\hat{\boldsymbol{z}} = D(\boldsymbol{s}, \boldsymbol{l}; \theta_D)$. D encourages the sentiment encoder E to keep as much sentiment information as possible by minimizing a reconstruction loss $\mathcal{L}_{\mathrm{rec}}$

$$\mathcal{L}_{\mathrm{rec}}(\theta_E, \theta_D) = \|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_2 \tag{4}$$

where $\hat{z}$ is the reconstructed joint representation, and θ_D is the parameter of the sentiment decoder.

E. Adversarial Domain Adaptation

Now, we introduce how to learn domain-invariant representations via adversarial training. As discussed in the previous section, domain discrepancies between multimodal sentiment datasets are commonly brought by the visual component, the textual component, and the expressive style. Therefore, we employ an adversarial learning method with two domain classifiers to address this problem.

To reduce multimodal discrepancies in the joint representation $\boldsymbol{z}$, we design the feature domain classifier C_f to encourage F to learn invariant features. Given $\boldsymbol{z}$, C_f tries to identify whether a training sample is drawn from the source or the target domain. On the contrary, F manages to fool C_f by extracting features without domain-specific information. F and C_f are simultaneously optimized according to the following equation:

$$\mathcal{L}^{f}_{\text{adv}}\left(\theta_f, \theta^{f}_{C}\right) = -\sum_{i}^{N_s}\sum_{j}^{N_t} \boldsymbol{d}^{s} \log C_f\left(\boldsymbol{z}^{s}_{i}; \theta^{f}_{C}\right) + \left(1 - \boldsymbol{d}^{t}\right) \log\left(1 - C_f\left(\boldsymbol{z}^{t}_{j}; \theta^{f}_{C}\right)\right) \quad (5)$$

where $\boldsymbol{d}^{s} = 0$ and $\boldsymbol{d}^{t} = 1$ denote the source and target domains, and θ^{f}_{C} is the parameter of C_f.

To remit the domain discrepancy of expressive styles, we employ the sentiment embedding module to obtain the disentangled sentiment representation that does not contain domain-specific style information. As introduced in the last section, the style domain classifier C_{st} and the sentiment embedding module E are trained to minimize a negative log probability of the style labels $\mathcal{L}^{\text{st}}_{\text{adv}}$, which is demonstrated in the following equation:

$$\mathcal{L}^{\text{st}}_{\text{adv}}\left(\theta_E, \theta^{\text{st}}_{C}\right) = -\sum^{N} \log p\left(\boldsymbol{l} | C_{\text{st}}\left(\boldsymbol{s}; \theta^{\text{st}}_{C}\right)\right) \quad (6)$$

where $N = N_s + N_t$, $\boldsymbol{l} = \{\boldsymbol{l}^{s}, \boldsymbol{l}^{t}\}$, and $\boldsymbol{s} = \{\boldsymbol{s}^{s}, \boldsymbol{s}^{t}\}$ denote the embedded expressive style vectors and sentiment representations of the two domains, and θ^{st}_{C} is the parameter of C_{st}.

Therefore, the loss function for domain adaptive learning is formulated as

$$\mathcal{L}_{\text{domain}} = \mathcal{L}^{f}_{\text{adv}} + \mathcal{L}^{\text{st}}_{\text{adv}}. \quad (7)$$

Based on the idea of adversarial training, we are seeking the parameters by optimizing the following equation:

$$\begin{aligned} \hat{\theta}_f, \hat{\theta}_E, \hat{\theta}_y &= \arg\min \mathcal{L}\left(\theta_f, \theta_E, \theta_y, \hat{\theta}^{\text{st}}_{C}, \hat{\theta}^{f}_{C}\right) \\ \hat{\theta}^{\text{st}}_{C}, \hat{\theta}^{f}_{C} &= \arg\max \mathcal{L}\left(\hat{\theta}_f, \hat{\theta}_E, \hat{\theta}_y, \theta^{\text{st}}_{C}, \theta^{f}_{C}\right). \end{aligned} \quad (8)$$

This optimization is implemented by a gradient reversal layer (GRL), as introduced in [62].

TABLE I

DATA STATISTICS. *Avg. #W* AND *Avg. #SW* DENOTE THE AVERAGE NUMBERS OF WORDS AND SENTIMENTAL WORDS CONTAINED IN EACH INSTANCE. THE *#Pos. inst*, *#Neu. inst*, AND *#Neg. inst* DENOTE THE NUMBERS OF LABELED POSITIVE, NEUTRAL, AND NEGATIVE INSTANCES, RESPECTIVELY

Statistics	*MVSA*	*Yelp*
Avg. #W.	36.9	12.7
Avg. #SW.	0.08	0.29
#Pos. inst.	10,195	21,261
#Neu. inst.	6,131	21,261
#Neg. inst.	1,633	21,261

IV. EXPERIMENTS

In this section, we report the experimental validation of the proposed method. To demonstrate the effectiveness of our model, we conduct experiments on multimodal single- and cross-domain sentiment classification tasks. The former task is designed to testify to the improvement of our method on multimodal sentiment classification, and the latter task is to evaluate the transferability of DiSRAN across different datasets. Experimental results show that our approach outperforms the baseline methods on both tasks according to the classification metrics.

A. Experimental Setup

1) Datasets: Two commonly used multimodal sentiment datasets, namely, *MVSA* and *Yelp*, are employed in our experiments. Table I illustrates the statistics of *MVSA* and *Yelp* datasets, where *#Pos. inst.*, *#Neu. inst.*, and *#Neg. inst.* denote the numbers of labeled positive, neutral, and negative instances, respectively. For all the two datasets, we split 80% of the data for training, 5% for validation, and 15% for testing.

The *MVSA* dataset was originally introduced in [63], including image–text pairs collected from Twitter, and three sentiment labels, i.e., positive, negative, and neutral. Each instance in the *MVSA* dataset has a visual sentiment label and a textual sentiment label. For the purpose of simplicity, we select the instances that have the same visual and textual sentiment labels in our experiments.

We adapt the *Yelp* dataset [7] to our MCDSC task. We select three categories of instances from the original *Yelp* dataset and reassign them with positive, negative, and neutral sentiment labels for label space consistency. Each multimodal review has an image and several sentences as the visual and textual inputs to the learning model.

Here, we discuss the distinction of the sentiment categories between datasets. Based on the distinction of sentiment categories and the statistics of the two datasets, we better demonstrate the discrepancy of expressive styles.

As previously mentioned, the samples in the *MVSA* dataset are tweets collected from social networks, while the instances in the *Yelp* dataset are online reviews of food, drinks, restaurants, and so on posted on *yelp.com*.[1] According to the theory

[1] https://www.yelp.com/dataset

TABLE II

EXPERIMENTAL RESULTS OF PRECISION, RECALL, AND F1-SCORE IN THE MULTIMODAL SINGLE-DOMAIN SENTIMENT CLASSIFICATION TASK

Methods	*MVSA*			*Yelp*		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
VisualSent	0.388	0.431	0.318	0.349	0.365	0.241
TextualSent	0.396	0.415	0.349	0.327	0.373	0.263
EarlyFusionSent	0.401	0.383	0.364	0.397	0.361	0.324
LateFusionSent	0.407	0.428	0.327	0.416	0.415	0.395
TFN	0.433	0.451	0.417	0.456	0.448	0.439
VistaNet	0.417	0.485	0.376	0.622	0.629	0.615
DiSRAN_A^{-}	0.467	0.506	0.381	0.695	0.685	0.672
DiSRAN_SE^{-}	0.415	0.378	0.214	0.444	0.435	0.423
DiSRAN	**0.476**	**0.532**	**0.466**	**0.698**	**0.699**	**0.691**

proposed in [64], categories of the *MVSA* indicate the sentiments possessed by individuals, which are neuropsychic dispositions toward a certain object or situation. Those categories of the *Yelp* dataset, however, are the polarities of opinions about a specific topic (e.g., food, drinks, and restaurants). Due to the widely interchangeable use of "sentiment" and "opinions" [64], for narrative convenience, we use "sentiment" to describe the categories in different datasets.

Because opinions are descriptive and assessing expressions of a fact of the matter [64], the expressive style of the *Yelp* dataset is therefore relatively explicit. On the contrary, sentiments conveyed by users in social networks determine an intangible and implicit expressive style of the *MVSA* dataset. Besides, based on the sentiment vocabulary (1915 positive and 2291 negative words) included in the Harvard General Inquiry dictionary [26], we obtained the statistics on the sentiment words of the *MVSA* and *Yelp* datasets. As shown in Table I, in terms of the average number of words (*Avg. #W*) and sentimental words (*Avg. #SW*), instances from the *Yelp* dataset have shorter textual content with more frequent sentimental vocabulary. Hence, the *Yelp* dataset does have a relatively explicit style of expressing sentiments compared to the *MVSA* dataset. In other words, there is a discrepancy in expressive styles between datasets. With the *MVSA* and *Yelp* datasets, we form two cross-domain sentiment classification tasks: *MVSA*→*Yelp* and *Yelp*→*MVSA*.

2) Implementation Details: To preprocess the textual inputs, we use NLTK [59] for stemming and tokenizing. We employ the pretrained GloVe [35] to initialize our 300-D embedding layer. The output dimension of the BiRNN is 4096. For the visual component, we use the pretrained VGG-16 model for feature extraction and fix its parameters during training. We draw the 4096-D output from the FC7 layer as our visual representation. The cross-modality attention matrices share the same dimensionality with the visual and textual representations. We apply the gradient descent optimization for updating the parameters of our model in the training process and tune the hyperparameters on the validation set. The proposed framework is implemented using PyTorch, and all of our approaches are trained on an NVIDIA GeForce RTX 2080Ti. All the results are collected from the experiments conducted on the testing set, which has empty intersections with the training and the validation sets.

3) Metrics: In this article, we employ precision, recall, and F1-score to evaluate the performance of the proposed model and the baseline methods. The macroaverage is applied to the three metrics to remit the impact of the imbalance of the label space.

B. Baselines

We compare our method with two categories of approaches: advanced sentiment classification deep neural networks and multimodal domain adaptation models. For the first category, we employ deep logistic regression models (VisualSent, TextualSent, EarlyFusionSent, and LateFusionSent), TFN [12], and VistaNet [7]; for the second category, we employ MDANN [24] and MMAN [25]. Some details for those baseline methods are given as follows.

1) *VisualSent:* A logistic regression model built on the top of a pretrained VGG-16 network.
2) *TextualSent:* A logistic regression model built on the top of a BiRNN with LSTM cells.
3) *EarlyFusionSent:* We integrate features extracted by VGG-16 and BiRNN through a concatenation operation [65] and feed the joint representation into a logistic regression model for sentiment classification.
4) *LateFusionSent:* We combine and average the results of VisualSent and TextualSent to generate the final predicted sentiment category [65].
5) *TFN:* It employs the tensor fusion layer to combine the visual and textual features, which are derived from VGG and LSTM, for sentiment classification.
6) *VistaNet:* The method applies visual aspect attention to model the interactions between the encoded visual and textual features. Because each instance in our dataset contains only one image, the reweighting layer of image-specific document representation has been removed.
7) *MMAN:* This network applies a stacked attention module to obtain multimodal representations and reduces the domain shift through adversarial training. The multichannel constraint is employed to capture fine-grained

TABLE III

EXPERIMENTAL RESULTS OF PRECISION, RECALL, AND F1-SCORE IN THE MULTIMODAL CROSS-DOMAIN SENTIMENT CLASSIFICATION TASK

Methods	*MVSA→Yelp*			*Yelp→MVSA*		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
VisualSent	0.221	0.332	0.258	0.213	0.311	0.222
TextualSent	0.278	0.415	0.324	0.221	0.342	0.248
EarlyFusionSent	0.132	0.333	0.168	0.191	0.341	0.189
LateFusionSent	0.262	0.361	0.276	0.203	0.331	0.245
TFN	0.330	0.331	0.311	0.322	0.316	0.164
VistaNet	0.403	0.413	0.331	0.404	0.403	0.351
MMAN	0.429	0.453	0.401	0.431	0.414	0.360
MDANN	0.469	0.463	0.439	0.401	0.403	0.369
DiSRAN_A$^-$	0.372	0.375	0.349	0.367	0.439	0.347
DiSRAN_SE$^-$	0.343	0.338	0.295	0.232	0.303	0.173
DiSRAN	**0.538**	**0.510**	**0.492**	**0.467**	**0.456**	**0.424**

TABLE IV

PRECISION (%) OF THREE MULTIMODAL DOMAIN ADAPTATION METHODS TRAINED WITH DIFFERENT PROPORTIONS OF SOURCE AND TARGET DOMAIN INSTANCES. THE PROPORTION R RANGES FROM 0.2 TO 5.0. THE BEST RESULT OF EACH LINE IS IN BOLD. THE DOWN ARROW (↓) DEMONSTRATES THE PRECISION DEGRADATION COMPARED TO THE BEST PERFORMANCE IN THE CORRESPONDING LINE

Task	Method	0.2	0.5	0.8	1.0	1.25	2.0	5.0
MVSA→Yelp	MMAN	23.1$\downarrow_{19.8}$	24.2$\downarrow_{18.7}$	38.5$\downarrow_{4.4}$	**42.9**	39.6$\downarrow_{3.3}$	27.6$\downarrow_{15.3}$	24.9$\downarrow_{18.0}$
	MDANN	29.6$\downarrow_{17.3}$	37.8$\downarrow_{9.1}$	41.1$\downarrow_{5.8}$	**46.9**	42.0$\downarrow_{4.9}$	40.8$\downarrow_{6.1}$	30.9$\downarrow_{16.0}$
	DiSRAN	46.2$\downarrow_{7.6}$	48.9$\downarrow_{4.9}$	50.3$\downarrow_{3.5}$	**53.8**	52.9$\downarrow_{0.9}$	51.8$\downarrow_{2.0}$	46.0$\downarrow_{7.8}$
Yelp→MVSA	MMAN	22.6$\downarrow_{20.5}$	26.0$\downarrow_{17.1}$	39.1$\downarrow_{4.0}$	**43.1**	42.0$\downarrow_{1.1}$	28.6$\downarrow_{14.5}$	26.9$\downarrow_{16.2}$
	MDANN	27.3$\downarrow_{12.8}$	34.7$\downarrow_{5.4}$	37.4$\downarrow_{2.7}$	**40.1**	37.5$\downarrow_{2.6}$	36.1$\downarrow_{4.0}$	30.1$\downarrow_{10.0}$
	DiSRAN	33.7$\downarrow_{13.0}$	40.8$\downarrow_{5.9}$	41.1$\downarrow_{5.6}$	**46.7**	43.3$\downarrow_{3.4}$	42.4$\downarrow_{4.3}$	40.2$\downarrow_{6.5}$

information and boost the performance on the target domain.

8) *MDANN:* It consists of covariant multimodal attention, fusion module, and hybrid domain constraints to obtain domain-invariant features. The model is learned under an adversarial objective.

For a fair comparison, all the baseline methods are implemented with the default settings provided in their published papers.

Besides, we implement several simplified versions of our method to testify the effectiveness of the proposed modules.

1) *DiSRAN_A$^-$:* A simplified implementation without self-attention and cross-modality attention layer.
2) *DiSRAN_SE$^-$:* A simplified implementation without the sentiment embedding module and the style domain classifier C_{st}. The predicted sentiment is generated based on the joint representations z.

C. Single-Domain Sentiment Classification Results

For the multimodal single-domain sentiment classification task, we apply the standard supervised training to all the methods. Training and testing samples are drawn from the same dataset. The domain classifiers of DiSRAN are removed to fit this task.

As the experimental results shown in Table II, DiSRAN outperforms the baseline methods on *MVSA* and *Yelp* datasets. The classification precisions of VisualSent and TextualSent are relatively low since only one modality of data has been used for sentiment prediction. EarlyFusionSent and LateFusionSent employ naive multimodal fusion approaches and slightly improve the performance. This improvement validates that integrated knowledge from multiple modalities facilitates precise sentiment classification. TFN and VistaNet, which apply advanced multimodal fusion methods, largely improve the classification precision. Compared with TFN and VistaNet, DiSRAN still makes a considerable improvement and achieves the best performances. Experimental results on the single-domain task verify the effectiveness of DiSRAN in sentiment classification.

In addition, DiSRAN increases the precision by about 0.6% and 15% on average in comparison to DiSRAN_A$^-$ and DiSRAN_SE$^-$. This observation verifies that the attention modules and the sentiment embedding module are indispensable roles in our proposed method.

D. Cross-Domain Sentiment Classification Results

In this section, we analyze the experimental results of DiSRAN and baseline methods in the MCDSC task. For the multimodal domain adaptation methods (i.e., DiSRAN, DiSRAN_A$^-$, DiSRAN_SE$^-$, MDANN, and MMAN), we follow the unsupervised domain adaptation protocol to sample labeled instances from the source domain and unlabeled

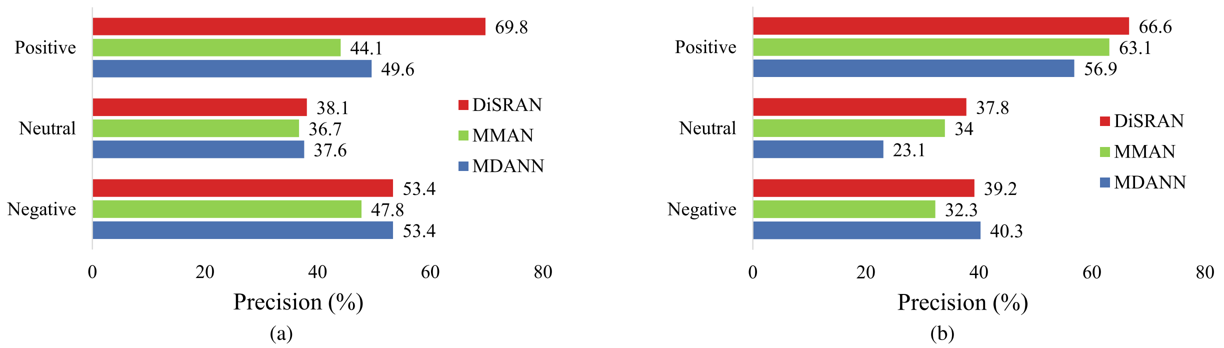

Fig. 3. Precision of each sentiment category on the multimodal cross-domain sentiment classification task. (a) *MVSA→Yelp*. (b) *Yelp→MVSA*.

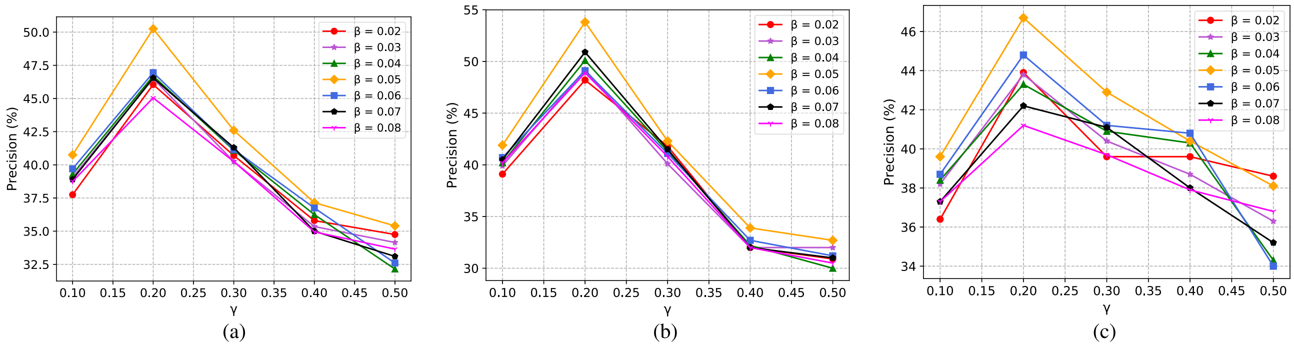

Fig. 4. Average precisions (%) of DiSRAN on multimodal cross-domain sentiment classification task when β and γ are set to different values. (a) Average. (b) *MVSA→Yelp*. (c) *Yelp→MVSA*.

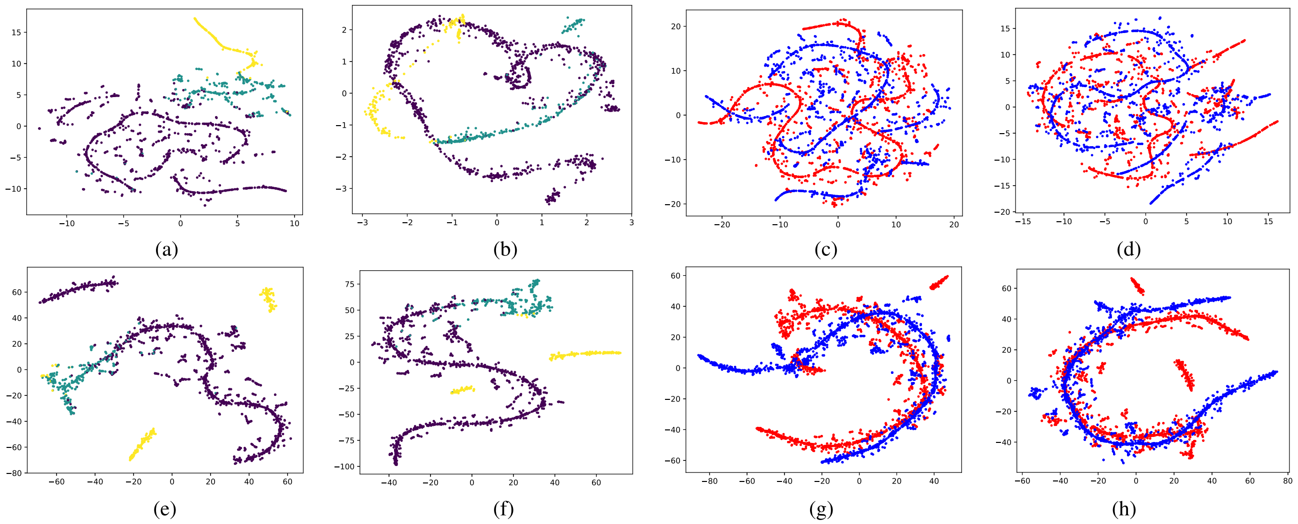

Fig. 5. Visualization of the learned sentiment representation. The left four figures show the results of DiSRAN and VistaNet on multimodal single-domain sentiment classification. The right four figures are the visualization results of our DiSRAN and MMAN on multimodal cross-domain sentiment classification. Blue and red points represent features from the source and target domains, respectively, and the points in other colors represent different sentiment categories. (a) VistaNet: *MVSA*. (b) VistaNet: *Yelp*. (c) MMAN: *MVSA→Yelp*. (d) MMAN: *Yelp→MVSA*. (e) DiSRAN: *MVSA*. (f) DiSRAN: *Yelp*. (g) DiSRAN: *MVSA→Yelp*. (h) DiSRAN: *Yelp→MVSA*.

instances from the target domain in the training process and evaluate them in the target domain. For the other methods, we apply the standard supervised training in the source domain and then directly transfer the learned model to the target domain.

Table III illustrates the experimental results of the MCDSC task. We can observe that the proposed DiSRAN model outperforms the baseline methods on both *MVSA→Yelp* and *Yelp→MVSA* tasks. Here, we present a detailed analysis of the results.

For those baseline methods, which take a direct transfer, they suffer a severe degradation of performance due to the domain shifts between the datasets. Surprisingly, EarlyFusionSent and LateFusionSent even have lower precisions than single-modality methods on target domains. This observation proves that domain discrepancies brought by multiple modalities exacerbate the degradation. One practical solution is to employ advanced multimodal fusion approaches. As shown, TFN and VistaNet achieve better adaptation performances due to the well-integrated multimodal features.

For those multimodal domain adaptation methods, they largely improve the performance by exploiting the target domain instances and learning domain-invariant features in an unsupervised manner. Specifically, MDANN and MMAN remit the discrepancies of visual and textual modalities between different domains through adversarial learning with the designed domain classifiers. These two methods, therefore, achieve higher performances than the approaches with direct transfer. However, we find that the different expressive styles closely correspond to the discrepancy between different domains. Due to the domain-specific style information in the learned representations, MDANN and MMAN yield lower performances than our proposed DiSRAN. To remit the discrepancy of expressive styles and learn domain-invariant features, DiSRAN optimizes the adversarial loss in (6) to enhance the joint distribution alignment of the source and target domains, ensuring that the embedded feature distributions over the two domains are more similar.

Besides, compared with DiSRAN_A$^-$ and DiSRAN_SE$^-$, the improvement of DiSRAN proves that the cross-modality attention layer and the sentiment embedding module play indispensable roles in the proposed model. They facilitate aligning multimodal semantic interactions and embedding disentangled sentiment representations without domain-specific style information, respectively. Based on the domain-invariant disentangled sentiment representations, DiSRAN outperforms the baseline methods on the MCDSC task.

Moreover, we present the classification precision of each sentiment category in Fig. 3. As shown, the Neutral category always has the lowest classification since this sentiment is relatively subtle to detect. For the *Yelp→MVSA* cross-domain sentiment classification, the precision of the Positive category on the target domain is significantly higher than the precisions of the other two categories. According to Table I, we can find that, in the *MVSA* dataset (i.e., the target domain dataset), there are more instances from the Positive category than those from the other two classes. That is to say, although the samples from the target domain have no label observations during the training process, the data distribution of the target domain still affects the adaptation performance.

For further analysis, we train the three multimodal domain adaptation methods with different proportions of source and target domain instances. We use $R = (N_s/N_t)$ to denote the proportion of training instances sampled from the two domains. When R ranges from 0.2 to 5.0, Table IV demonstrates the performances of the learned models on *MVSA→Yelp* and *Yelp→MVSA* tasks. As shown, all the models achieve their best performance when equal numbers of instances from the two domains are used during training. When R is relatively small, i.e., training with a small amount of source domain data, DiSRAN yields an inferior performance because of the insufficient sentiment information from the labeled source domain. On the contrary, when the source domain data becomes richer than 100%, the adaptation performance of the proposed model drops due to the lack of information about the target domain. Compared with baseline methods, DiSRAN consistently gets higher classification precisions and is less sensitive to the change of the proportion of training instances from different domains. This indicates that our model is more effective in the multimodal cross-domain sentiment analysis.

E. Discussions

1) Hyperparameter Analysis: In this subsection, we analyze the impact of the hyperparameters that are used to weight different loss functions, as shown in (1). We evaluate the performance of DiSRAN with different β and γ values on the MCDSC tasks to find the best hyperparameter setting. As demonstrated in Fig. 4, DiSRAN obtains the best performance when $\beta = 0.05$ and $\gamma = 0.2$. When β becomes rather small, the network cannot learn useful domain-invariant knowledge and remit the discrepancies between datasets. The domain shifts degrade the performance of the learned model on the target domain. When γ becomes rather small, the disentangled representation contains less sentiment information, which is not informative for sentiment classification. On the contrary, when the hyperparameters become rather big, the sentiment classifier remains underfitting. As a result, the classification precision on the target domain generally decreases. Based on experimental results, we set β and γ to 0.05 and 0.2, respectively, for promising transfer performance.

2) Visualization: To demonstrate the effectiveness of our method, we visualize the learned sentimental representation $\boldsymbol{s}$ via t-SNE [66]. In Fig. 5, we present the visualization results of multimodal single- and cross-domain sentiment classification tasks. As a comparison, we visualize the learned features of the best performing baseline methods (i.e., VistaNet and MMAN) on single- and cross-domain tasks. As shown, the DiSRAN clusters sentiment representations from different categories into separate regions, which are more discriminative for sentiment classification. Besides, DiSRAN has a better fusion of the sentiment representations that are learned from different domains. This demonstrates that our approach learns domain-invariant sentiment information that improves the performance of multimodal cross-domain sentiment analysis.

V. Conclusion

In this article, we propose a DiSRAN that remits the discrepancy of expressive styles between different datasets for multimodal the cross-domain sentiment analysis task. Specifically, we design a sentiment embedding module to disentangle sentiment information from the multimodal joint representation that contains domain-specific style information via adversarial training. Because the disentangled sentiment representation is domain-invariant, the proposed DiSRAN achieves better adaptation performance. Experimental results on the MCDSC task have proved the effectiveness of our method. Besides, recent studies have shown that pretrained textual representations of BERT and its variants help improve performances in many downstream tasks. However, since existing sentiment analysis baselines did not use BERT-based pretrained textual representations, we followed the traditional representations with GloVe to ensure that our approach yields convincing performance improvements. In future work, we will explore methods that use BERT for pretrained representations and extend the proposed method to multimodal sentiment datasets with more than two modalities.

References

[1] D. Tang, B. Qin, and T. Liu, "Learning semantic representations of users and products for document level sentiment classification," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2015, pp. 1014–1023.

[2] D. Tang, B. Qin, T. Liu, and Y. Yang, "User modeling with neural network for review rating prediction," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 1340–1346.

[3] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar, "Beyond binary labels: Political ideology prediction of Twitter users," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2017, pp. 729–740.

[4] L. Luo *et al.*, "Beyond polarity: Interpretable financial sentiment analysis with hierarchical query-driven attention," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 4244–4250.

[5] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.

[6] D. Tang, B. Qin, and T. Liu, "Deep learning for sentiment analysis: Successful approaches and future challenges," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 5, no. 6, pp. 292–303, Nov. 2015.

[7] Q.-T. Truong and H. W. Lauw, "Vistanet: Visual aspect attention network for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 305–312.

[8] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, no. 1, pp. 3–14, Sep. 2017.

[9] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 251–260.

[10] F. Chen, Y. Gao, D. Cao, and R. Ji, "Multimodal hypergraph learning for microblog sentiment prediction," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jun. 2015, pp. 1–6.

[11] L. Li, D. Cao, S. Li, and R. Ji, "Sentiment analysis of Chinese microblog based on multi-modal correlation model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 4798–4802.

[12] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.

[13] T. Gui *et al.*, "Cooperative multimodal approach to depression detection in Twitter," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 110–117.

[14] S. Zhao *et al.*, "A review of single-source deep unsupervised visual domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 473–493, Feb. 2022.

[15] Z. Li, Y. Wei, Y. Zhang, and Q. Yang, "Hierarchical attention transfer network for cross-domain sentiment classification," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 5852–5859.

[16] W. Mei and D. Weihong, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, Jul. 2018.

[17] W. Zhao *et al.*, "Weakly-supervised deep embedding for product review sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 185–197, Jan. 2017.

[18] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, "Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2272–2281.

[19] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7167–7176.

[20] Y.-C. Liu, Y.-Y. Yeh, T.-C. Fu, S.-D. Wang, W.-C. Chiu, and Y.-C.-F. Wang, "Detach and adapt: Learning cross-domain disentangled deep representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8867–8876.

[21] B. Heredia, T. M. Khoshgoftaar, J. Prusa, and M. Crawford, "Cross-domain sentiment analysis: An empirical investigation," in *Proc. IEEE 17th Int. Conf. Inf. Reuse Integr. (IRI)*, Jul. 2016, pp. 160–165.

[22] M. Peng, Q. Zhang, Y.-G. Jiang, and X. Huang, "Cross-domain sentiment classification with target domain specific information," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 2505–2513.

[23] J. Meng, Y. Long, Y. Yu, D. Zhao, and S. Liu, "Cross-domain text sentiment analysis based on CNN_FT method," *Information*, vol. 10, no. 5, p. 162, May 2019.

[24] F. Qi, X. Yang, and C. Xu, "A unified framework for multimodal domain adaptation," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 429–437.

[25] X. Ma, T. Zhang, and C. Xu, "Deep multi-modality adversarial networks for unsupervised domain adaptation," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2419–2431, Sep. 2019.

[26] L. Gatti, M. Guerini, and M. Turchi, "SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 409–421, Oct./Dec. 2016.

[27] H. Kanayama and T. Nasukawa, "Fully automatic lexicon expansion for domain-oriented sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2006, pp. 355–363.

[28] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, 2011.

[29] T. Mullen and N. Collier, "Sentiment analysis using support vector machines with diverse information sources," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2004, pp. 412–418.

[30] J. Bollen, H. Mao, and A. Pepe, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 5, no. 1, 2011, pp. 450–453.

[31] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2011, pp. 1031–1040.

[32] S. Lai, L. Xu, K. Liu, and J. Zhao, "Recurrent convolutional neural networks for text classification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 29, no. 1, 2015, pp. 2267–2273.

[33] L. Zhu, W. Li, Y. Shi, and K. Guo, "SentiVec: Learning sentiment-context vector via kernel optimization function for sentiment analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2561–2572, Jun. 2021.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent., (ICLR)*, Y. Bengio and Y. LeCun, Eds. Scottsdale, AZ, USA, 2013, pp. 1–12.

[35] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[36] A. Nazir, Y. Rao, L. Wu, and L. Sun, "Issues and challenges of aspect-based sentiment analysis: A comprehensive survey," *IEEE Trans. Affect. Comput.*, early access, Jan. 30, 2020, doi: 10.1109/TAFFC.2020.2970399.

[37] G. Yu *et al.*, "Making flexible use of subtasks: A multiplex interaction network for unified aspect-based sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics, (ACL-IJCNLP)*, 2021, pp. 2695–2705.

[38] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," *Knowl.-Based Syst.*, vol. 167, pp. 26–37, Mar. 2019.

[39] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 715–718.

[40] D. Borth, T. Chen, R. Ji, and S.-F. Chang, "SentiBank: Large-scale ontology and classifiers for detecting sentiment and emotions in visual content," in *Proc. 21st ACM Int. Conf. Multimedia (MM)*, 2013, pp. 459–460.

[41] Y. Yang *et al.*, "How do your friends on social media disclose your emotions?" in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, 2014, pp. 306–312.

[42] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 29, no. 1, 2015, pp. 1–8.

[43] Q. You, H. Jin, and J. Luo, "Visual sentiment analysis by attending on local image regions," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017, pp. 231–237.

[44] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proc. 13th Int. Conf. Multimodal Interface (ICMI)*, 2011, pp. 169–176.

[45] Z. Li, Y. Fan, B. Jiang, T. Lei, and W. Liu, "A survey on sentiment analysis and opinion mining for social multimedia," *Multimedia Tools Appl.*, vol. 78, no. 6, pp. 6939–6967, 2019.

[46] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.

[47] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Selective transfer machine for personalized facial action unit detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3515–3522.

[48] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2006, pp. 120–128.

[49] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2007, pp. 137–144.

[50] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Heterogeneous domain adaptation through progressive alignment," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1381–1391, May 2019.

[51] S. Pachori, A. Deshpande, and S. Raman, "Hashing in the zero shot framework with domain adaptation," *Neurocomputing*, vol. 275, pp. 2137–2149, Jan. 2018.

[52] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2007, pp. 513–520.

[53] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[54] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Int. Conf. Mach. Learn.*, 2015, pp. 97–105.

[55] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[56] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[57] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko, "Simultaneous deep transfer across domains and tasks," in *Proc. Int. Conf. Comput. Vis.*, 2015, pp. 4068–4076.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[59] S. Bird, "NLTK: The natural language toolkit," in *Proc. COLING/ACL Interact. Presentation Sessions*, 2006, pp. 214–217.

[60] Z. Fu, X. Tan, N. Peng, D. Zhao, and R. Yan, "Style transfer in text: Exploration and evaluation," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 663–670.

[61] T. Shen, T. Lei, R. Barzilay, and T. Jaakkola, "Style transfer from non-parallel text by cross-alignment," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6830–6841.

[62] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, 2017.

[63] T. Niu, S. Zhu, L. Pang, and A. El-Saddik, "Sentiment analysis on multi-view social data," in *Proc. Int. Conf. Multimedia Modeling*, 2016, pp. 15–27.

[64] M. D. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Trans. Affective Comput.*, vol. 5, no. 2, pp. 101–111, Apr./Jun. 2014.

[65] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[66] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Yuhao Zhang is currently pursuing the master's degree with the College of Cyber Science, Nankai University, Tianjin, China.

His research interests include multimodal analysis, sentiment analysis, and transfer learning.

Ying Zhang received the Ph.D. degree from Nankai University, Tianjin, China, in 2013.

From 2011 to 2013, she studied at Purdue University, West Lafayette, IN, USA. She is currently a Professor with the College of Computer Science, Nankai University. Her main research interests include sentiment analysis, multimodal data analysis, and information retrieval.

Wenya Guo is currently pursuing the Ph.D. degree with the College of Computer Science, Nankai University, Tianjin, China.

Her current research interests include multimodal sentiment analysis, visual question answering, and referring expression comprehension.

Xiangrui Cai received the Ph.D. degree from Nankai University, Tianjin, China, in 2018.

He is currently an Assistant Professor with the College of Cyber Science, Nankai University. His research interests include natural language processing and medical data analysis.

Xiaojie Yuan received the Ph.D. degree from Nankai University, Tianjin, China, in 2000.

She is currently a Professor with the College of Computer Science, Nankai University. Her main research interests include big data management and data mining.