Shared and Private Information Learning in Multimodal Sentiment Analysis with Deep Modal Alignment and Self-supervised Multi-Task Learning

Songning Lai, Xifeng Hu, Yulong Li, Zhaoxia Ren, Zhi Liu and Danmin Miao

***Abstract*—Designing an effective representation learning method for multimodal sentiment analysis tasks is an important research direction. A complete modal representation should contain both shared and private information, which is challenging to learn with uniform multimodal labels and a raw feature fusion approach. In our work, we designed a deep modal shared information learning module based on the covariance matrix to capture the shared information between modalities. Additionally, we used a label generation module based on a self-supervised learning strategy to capture the private information of the modalities. It is worth noting that our module is plug-and-play in multimodal tasks. By changing the parameterization, the module can adjust the information exchange relationship between the modes and learn the private or shared information between the specified modes. We also employed a multi-task learning strategy to help the model focus its attention on the modal differentiation training data. We provided a detailed formulation derivation and feasibility proof for the design of the deep modal shared information learning module. We conducted extensive experiments on three common multimodal sentiment analysis baseline datasets, and the experimental results validated the reliability of our model. Furthermore, we explored more combinatorial techniques for the use of the module. Most of the metrics of our approach on the three public datasets outperformed current state-of-the-art methods.**

***Index Terms*—Multimodal sentiment analysis, multi-task learning, modal alignment, self-supervised learning, Domain generalization.**

I. Introduction

MULTIMODAL sentiment analysis (MSA) [1], [2] is a rapidly growing field that utilizes information from various modalities to accomplish sentiment analysis. By incorporating multiple modalities, such as text, audio, and visual data, MSA can provide a more complete representation of human emotion. Unlike previous analyses that focus on a single modality [3], MSA recognizes the complementarity between different modalities and uses them jointly to improve accuracy. Studies have shown [4]–[7] that non-text modal information can effectively enhance the accuracy of sentiment analysis. The goal of MSA is to analyze sentiment using data from multiple modalities, as illustrated in Figure 1.

Multimodal sentiment analysis (MSA) has received significant attention, but it still faces many challenges [8], [9]. There are five fundamental challenges for multimodal tasks [10]: alignment, translation, representation, fusion, and co-learning. Among these, the representation of individual modalities and the overall representation challenge of multimodality are the most significant and meaningful challenges. In the past, previous work has not adequately addressed the issue of learning shared or private information between modalities. Instead, they typically fused the features of each modality without distinguishing between shared and private information. Future research could focus on developing methods to deliberately learn shared or private information between modalities to improve the accuracy of MSA.

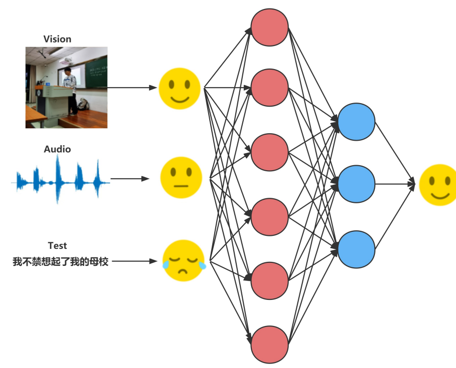

Fig. 1. Explanation diagram of multimodal sentiment analysis task.

In the field of multimodal sentiment analysis (MSA), distinguishing between shared and private information among modalities is crucial for improving accuracy. Previous work, such as [2], [11], has attempted to address this issue. In [12], a self-supervised multimodal Multi-Task learning strategy II was proposed, focusing on the backward-guided approach. We drew inspiration from this strategy and used it to automatically generate unimodal labels. This approach allowed us to focus on the private information among modalities, and update weights based on momentum for effective learning.

In the domain generalization area [13]–[15], inter-domain alignments [16] have been used to learn distributions outside of the domain. We applied this idea to multimodal tasks to learn shared information between modalities. To this end, we designed a deep inter-modal shared information learning module, which utilizes a deep inter-modal covariance matrix-based loss function.

Each modality has its own unique virtual label, and the

Songning Lai , Xifeng Hu and Zhi Liu are with the School of Information Science and Engineering, Shandong University, Qingdao, 266237, China.
Yulong Li and Danmin Miao are with Department of Military Medical Psychology, Air Force Medical University, Xi'an, 710032, China
Zhaoxia Ren is with Assets and Laboratory Management Department, Shandong University, Qingdao, 266237, China.

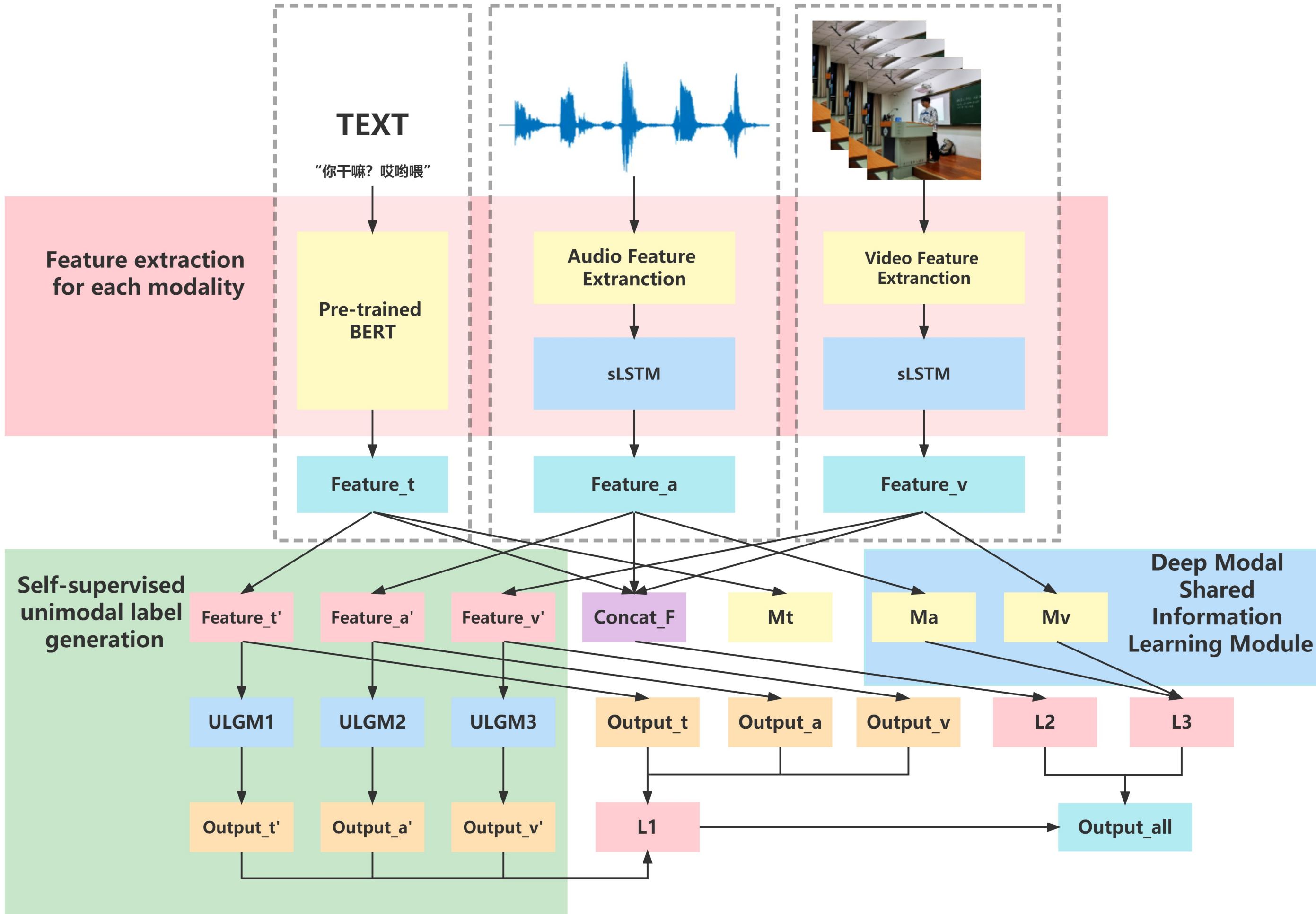

Fig. 2. Flowchart of the complete model architecture. It includes feature extraction modules for each modality, a self-supervised unimodal label generation module, a deep modal shared information learning module, and a multimodal sentiment analysis output module.

model should focus on specific differential information to learn shared and private information. Therefore, we merged multiple losses, including shared information loss, private information loss, multi-task loss, and task prediction loss. By doing so, we were able to effectively learn both shared and private information among modalities, such as sarcastic text and mannerisms, or visual micro-expressions distinct from audio expressions.

In conclusion, improving the accuracy of MSA requires a deep understanding of shared and private information among modalities. Our approach, utilizing a deep inter-modal shared information learning module and multiple loss functions, shows promise in effectively learning this information. However, more research is needed to further address the challenges in this area.

Our work presents innovative contributions that can be summarized as follows:

1. We propose a function based on the covariance matrix as a second-order statistic to measure the distribution of features between aligned and drawn-out modes.

2. We design a differentiable loss function to train the network to learn shared information between modalities.

3. We use a self-supervised learning strategy generation module to guide the multimodal task to focus on modality-specific private information.

4. We conduct comprehensive experiments on three benchmark datasets for multimodal sentiment analysis to validate the feasibility of our designed module. Our approach outperforms the current state-of-the-art methods.

II. Related Work

A. Multimodal Sentiment Analysis

With the rapid development of social networking platforms and video sites, the Internet is experiencing an explosion of information. poria et al [17] tried to explore multimodal sentiment analysis with three modalities: visual, audio and text. For text they use text2vec. for visual features they use CNN and SVM to work together to extract features. For audio features they use openSMILE to extract features [18]. Semantic Feature Fusion Neural Network (SFNN) [19] uses CNN and attention mechanism for the first time to get the emotional features of the image. Then the emotional features are mapped to the semantic feature level. Finally, the image features of visual information and semantic features are fused. The impact caused by the variability of heterogeneous data is effectively reduced. Modality-invariant and specific representation for multimodal emotion analysis (MISA) [20] serves as an independent framework to learn the modality-invariant and specific features by dividing the subspace of modality. Numerous studies have also shown that visual and audio features can reflect emotions well, but textual features are still important for entity and subjectivity recognition.

With the rise of social networking platforms and video sites, the internet has become a seemingly endless pool of information. In an effort to explore multimodal sentiment analysis, Poria et al. [18] utilized three modalities: visual, audio, and text. They utilized text2vec for text, CNN and SVM for visual features, and openSMILE for audio features. The Semantic Feature Fusion Neural Network (SFNN) [19] used CNN and attention mechanisms to extract emotional features from images, which were then mapped to semantic feature levels before being fused with visual information. The Modality-invariant and specific representation for multimodal emotion analysis (MISA) [20] was used to learn modality-invariant and specific features by dividing the subspace of modality. While visual and audio features have proven effective in reflecting emotions, textual features continue to be important for entity and subjectivity recognition.

Our work aims to enhance the ability to extract both private and shared information across modalities, and to fuse multimodal features in later stages. To achieve this, we utilize self-supervised and multi-task learning strategies, as well as a deep inter-modal shared information learning module. This module employs a deep inter-modal covariance matrix-based loss function to effectively learn shared and private information among modalities.

Our proposed method has been extensively evaluated on benchmark datasets, demonstrating its effectiveness in extracting and fusing multimodal features. However, further research is needed to address the challenges associated with this area. We believe that our work provides a valuable contribution to the field, and will inspire future research in multimodal sentiment analysis and related areas.

B. BERT

Bidirectional Encoder Representations from Transformer (BERT) [21]–[24] has become a game-changer in text analysis tasks. The development of pre-trained models has been a major breakthrough in the field of natural language processing, and BERT has shown exceptional accuracy in various text processing tasks. BERT addresses the limitations of text features by attempting to learn ”masked language models”. This involves learning a specific representation of the objective function to predict randomly selected and masked text, while taking contextual relationships into account.

BERT has two primary uses in multimodal sentiment analysis tasks. Firstly, it can extract features from text data using pre-trained BERT models. Secondly, it can fuse modal features of data across modalities using BERT.

In our work, we utilized open-source pre-trained BERT models to extract features from text data. Although we achieved promising results, there are still challenges in this area that require further research. We believe that our work provides a valuable contribution to the field and will inspire future research in multimodal sentiment analysis and related areas.

C. LSTM

Long Short Term (LSTM) [25]–[27] is a special Recurrent Neural Networks (RNN) network structure (with the addition of a cell state) that learns dependency information for long time periods. In many tasks concerning time series [28], [29], LSTM has achieved great success and has been widely promoted. Its design by can effectively avoids the problem of long-term dependencies, and memorizing long-term feature information is the default capability of LSTM.

Our work utilizes a one-way LSTM network for feature extraction of video and audio information. This approach captures the highly correlated features of emotions with time series information in these two modal data.

D. ULGM

The Unimodal Label Generation Module (ULGM) [12] is an automatic generation module used for creating unimodal labels in multimodal tasks. Initially developed for multimodal sentiment analysis by Wenmeng Yu et al., ULGM is based on two assumptions. Firstly, the distance of a modal label is positively correlated with the distance between the modal feature representation and the class center. Secondly, unimodal labels are highly correlated with multimodal labels. The non-parametric module is based on self-supervised learning and calculates the migration of unimodal labels compared to multimodal labels based on the relative distance from the unimodal feature representation to the class center of the multimodal. The ULGM module can effectively guide the subtask to focus on samples with large differences between modalities.

In our work, we utilize the ULGM module as a subtask for multimodal sentiment analysis. This allows us to capture differentiation information between modalities through the multimodal sentiment analysis task.

E. Domain generalization

Domain generalization (DG) [13], [14], [30] has emerged as a popular research direction in recent years. The goal is to develop models with strong generalization ability by training them on multiple datasets with different data distributions. DG models fall into three categories: data augmentation, replacement learning strategies, and learning to domain-invariant features. Transfer component analysis [31] aims to find a kernel function that minimizes the distribution differences between all data in the feature domain. Meanwhile, domain adversarial neural networks [32] leverage the GAN framework to identify a data’s source domain and obtain domain-invariant features. Our work builds on this idea of domain generalization and focuses on learning domain-invariant features. We introduce a deep modal shared information learning module based on the covariance matrix. This module helps our model learn shared information between different modalities.

F. Multi-Task learning

Multi-Task learning [33]–[35] is a machine learning subfield that leverages the similarities between different tasks to solve them simultaneously, improving the learning potential of each task. It falls under the umbrella of migration learning, which capitalizes on the domain-specific information implicit in multiple related tasks' training signals. Multi-Task learning uses shared parameters during the backward propagation process, allowing features to be used by other tasks. This results in the learning of features that can be applied to several different tasks, enhancing the generalization performance of multiple tasks. The shared parameters mainly include soft and hard sharing. Balancing the learning process of multiple tasks is a crucial issue that must be addressed. In the realm of multimodal sentiment analysis, Multi-Task learning [36]–[39] is widely used.

For our work, we adopt a hard sharing approach for subtasks to share parameters and utilize a weight adjustment strategy to balance each task's learning process.

III. Methods

The classical multimodal sentiment analysis model is used for the multimodal task. This model comprises a feature extraction module for each modality, a modal feature fusion module, and a result output module. We have designed a deep modal shared information learning module that optimizes the feature extraction. To further enhance the learning of private features, we have also incorporated the Unimodal Label Generation Module (ULGM) into the multimodal sentiment analysis task. Please refer to Figure 2 for a complete diagram of the model's generalization.

A. Feature Extraction for Text Modality

In text modality, pre-trained language models based on BERT show robust capabilities in text tasks. In this task, we use the pre-trained BERT model trained over large corpus data to extract text features of the sentence. The first word vector of the last layer is chosen as the textual feature of the whole sentence - Feature_t.

$$Feature\_t \ = \ BERT(Input\_t; \omega_t^{BERT}) \in R^{d_t} \quad (1)$$

B. Feature extraction for audio modality

In audio modality, we use a pre-trained toolkit [40], [41] to extract the initial feature $Input_a$ from the original audio sequence data. this initial feature is a feature matrix of$L_a \times d_a$. where L_a is the length of the audio sequence and d_a is the length of the extracted audio features at each moment. Then, $Input_a$ is fed into the unidirectional LSTM to learn the time series features in the audio modality. Finally, the hidden vector of the last layer of the unidirectional LSTM network is used as the feature representation of the whole audio modality.

$$Feature\_a \ = \ sLSTM(Input\_a; \omega_a^{sLSTM}) \in R^{d_a} \quad (2)$$

C. Feature extraction for visual modality

In the visual modality, we also use a pre-trained toolkit specifically for extracting visual initial features [40], [41]. The initial feature matrix $Input_v$ with visual sequences is extracted from the original video sequence data amount. this initial feature is a feature matrix of $L_v \times d_v$. Where L_v is the length of the video sequence and d_v is the length of the extracted image features at each moment. Then, $Input_v$ is fed into the unidirectional LSTM to learn the time-series features in the visual modality. Finally, the hidden vector of the last layer of the unidirectional LSTM network is used as the feature representation of the whole visual modality.

$$Feature\_v \ = \ sLSTM(Input\_v; \omega_v^{sLSTM}) \in R^{d_v} \quad (3)$$

D. Modal fusion

For the deep modal features ($Feature_t$, $Feature_a$, $Feature_v$) acquired by each modality, we concatenate them into a single sequence and project them into the same low-dimensional space (R^{d_m}).

$$Concat\_F \ = \ [Feature\_t; Feature\_a; Feature\_V] \quad (4)$$

$$Feature\_all^* = ReLU({\omega_{l1}^{all}}^T Concat\_F + b_{l1}^{all}) \quad (5)$$

where, $\omega_{l1}^{all} \in R^{(d_t+d_a+d_v)\times d_{all}}$.

E. Predictive Analysis

For the obtained $Feature_all^*$, the classification or regression prediction task of multimodal sentiment analysis is done by one linear layer.

$$y_{all}^{output} = {\omega_{l2}^{all}}^T Feature\_all^* + b_{l2}^{all} \quad (6)$$

where, $\omega_{l2}^{all} \in R^{d_{all}\times 1}$.

F. Feature Projection & Fusion with ULGM Module

The ULGM module is the main module for subtasking in multi-task learning, which can automatically generate unimodal labels.

For the deep modal features obtained in each modality are projected into the same feature space separately. The prediction task of multimodal sentiment analysis is accomplished by a linear layer to obtain y_i^{output}. Then the ULGM module is used to complete the classification or regression prediction task for unimodal sentiment analysis, and $y_i^{output\prime}$ is obtained.

$$Feature\_s^* = ReLU({\omega_{l1}^{s}}^T Feature\_s + b_{l1}^{s}) \quad (7)$$

$$y_s^{output} = {\omega_{l2}^{s}}^T Feature\_s^* + b_{l2}^{s} \quad (8)$$

$$y_s^{output\prime} = ULGM(y_{all}^{output}, Feature\_all^*, Feature\_s^*) \quad (9)$$

$$s \in \{\ t,\ a,\ v\} \tag{10}$$

The ULGM module calculates the offset (relative distance from the unimodal representation to the positive and negative centers) based on the relative distance from the unimodal special to the multimodal class center. A momentum-based update strategy is used to combine the newly generated unimodal labels with the historically generated unimodal labels. This strategy helps the self-supervised generated unimodal labels to gradually stabilize during the subtask training. For the detailed introduction of the ULGM module and the derivation of the formula, Wenmeng Yu et al [12] have provided a comprehensive explanation, which will not be introduced in this paper.

G. Deep Modal Shared Information Learning Module

The deep modal shared information learning module is a useful tool for extracting shared information from multiple modalities in a deep learning system. This module enables the extraction of shared information between any two specified modalities, such as audio and visual data. To optimize feature extraction, we experimented with various inter-modal and intra-modal combinations before selecting the audio-visual modality combination for this experiment. In this subsection, we provide a detailed explanation of this module with a focus on these two modalities.

For the audio modality, $Feature_a$ is used; for the visual modality, $Feature_v$. Since each training is based on a specific batch of data, the number of audio modality $Feature_a$ and $Feature_v$ is set here to be N_a and N_v ($N_a = N_v$ since it is the same batch of data). Each $Feature_a$ and $Feature_t$ is projected into the space of dimension d.

The single batch matrix of audio modality.

$$M_a = \{Feature\_a_i\}, i = 1, 2, \cdots, N_a \tag{11}$$

The single batch matrix of visual modality.

$$M_v = \{Feature\_v_i\}, i = 1, 2, \cdots, N_v \tag{12}$$

The covariance matrices of audio modality and visual modality are constructed separately. A function based on the covariance matrix [42] is used to reflect the shared information content between the modalities.

$$C_a = \frac{1}{N_a - 1}({M_a}^T M_a - \frac{1}{N_a}(1^T M_a)^T(1^T M_a)) \tag{13}$$

$$C_v = \frac{1}{N_v - 1}({M_v}^T M_v - \frac{1}{N_v}(1^T M_v)^T(1^T M_v)) \tag{14}$$

Constructing loss function to facilitate multimodal sentiment analysis models to focus on and learn shared information between modalities.

$$\theta_{share} = \frac{1}{4d^2}||C_a - C_v||_F^2 \tag{15}$$

The gradient can be calculated using the chain rule. It is known that θ_{share} is a differentiable function and can be back-propagated in the network.

$$\frac{\partial \theta_{share}}{\partial M_a^{ij}} = \frac{1}{d^2\left(N_{\ a} - 1\right)}(({M_a}^T - \frac{1}{N_a}(1^T M_a)^T(1^T))^T(C_a - C_v))^{ij} \tag{16}$$

$$\frac{\partial \theta_{share}}{\partial M_v^{ij}} = \frac{1}{d^2(N_{v} - 1)}(({M_v}^T - \frac{1}{N_v}(1^T M_v)^T(1^T))^T(C_a - C_v))^{ij} \tag{17}$$

The thought of the module is to match the inter-modal distribution by the second order statistics between the modes. The task of the whole module is to minimize the function.

$$\min_A ||C_a - C_v||_F^2 = \min_A ||A^T C_a A - C_v||_F^2 \tag{18}$$

The following is a brief proof that the function has an optimal solution.

Set ε^+ to be the Moore-Penrose pseudoinverse of ε, and R_{C_a} and R_{C_v} to be the ranks of C_a and C_v, respectively. a is a linear transformation of C_a. $A^T C_a A$ does not increase the rank of C_a. Therefore, $R'_{C_a} \leq R_{C_a}$ and the covariance matrices are all symmetric matrices. So SVD on the covariance matrices of the two modes yields:

$$C_a = U_a \varepsilon_a U_a^T \tag{19}$$

$$C_v = U_v \varepsilon_v U_v^T \tag{20}$$

when $R_{C_a} > R_{C_v}$, the optimal solution is $C'_a = C_v$. So the optimal solution of is:

$$C_a^{'} = U_v \varepsilon_v U_v^T = U_{v[1:R]} \varepsilon_{v[1:R]} U_{v[1:R]}^T \tag{21}$$

$$R\ =\ R_{C_v} \tag{22}$$

where $\varepsilon_{v[1:R]}, U_{v[1:R]}$ are the maximum singular value of v and the corresponding left singular vector, respectively.
when $R_{C_a} \leq R_{C_v}$, the optimal solution is:

$$C_a^{'} = U_v \varepsilon_v U_v^T = U_{v[1:R]} \varepsilon_{v[1:R]} U_{v[1:R]}^T \tag{23}$$

$$R\ =\ R_{C_a} \tag{24}$$

In summary, $R = min(R_{C_a}, R_{C_v})$.

$$C_a^{'} = U_v \varepsilon_v U_v^T = A^T C_a A \tag{25}$$

$$C_a = U_a \varepsilon_a U_a^T \tag{26}$$

Combining the above equations yields.

TABLE I
EXPERIMENTAL RESULTS FOR REGRESSION TASK AND CLASSIFICATION TASK ON MOSI AND MOSEI DATASETS. (1) INDICATES THAT THE RESULTS ARE FROM THE EXPERIMENTAL RESULTS OF HAZARIKA ET AL [20].

Model	**MOSI**				**MOSEI**				**Data Setting**
	MAE	**Corr**	**Acc-2**	**F1-Score**	**MAE**	**Corr**	**Acc-2**	**F1-Score**	
TFN(1)	0.901	0.698	-/80.8	-/80.7	0.593	0.7	-/82.5	-/82.1	Unaligned
LMF(1)	0.917	0.695	-/82.5	-/82.4	0.623	0.677	-/82.0	-/82.1	Unaligned
RAVEN(1)	0.915	0.691	78.0/-	76.6/-	0.614	0.662	79.1/-	79.5/-	Aligned
MFM(1)	0.877	0.706	-/81.7	-/81.6	0.568	0.717	-/84.4	-/84.3	Aligned
MulT(1)	0.861	0.711	81.5/84.1	80.6/83.9	0.58	0.703	-/82.5	-/82.3	Aligned
MISA	0.794	0.758	79.32/79.79	80.21/81.46	0.579	0.711	81.24/83.56	82.87/84.46	Aligned
MAG_BERT	0.765	0.774	82.43/83.49	82.87/83.81	0.566	0.748	83.66/84.76	83.68/84.48	Aligned
Self_MM	0.723	**0.797**	83.09/84.79	**83.03**/84.78	0.534	0.764	82.32/84.12	82.81/84.05	Unaligned
Ours	**0.704**	0.794	**83.65/85.06**	82.51/**85.43**	**0.523**	**0.766**	**82.98/85.01**	**83.26/84.89**	Unaligned

TABLE II
EXPERIMENTAL RESULTS FOR REGRESSION TASK AND CLASSIFICATION TASK ON SIMS DATASET. (2) INDICATES THAT THE RESULTS ARE FROM THE EXPERIMENTAL RESULTS OF WENGMENG YU ET AL [12].

Model	**MAE**	**Corr**	**Acc-2**	**F1-Score**
TFN(2)	0.428	0.605	79.86	80.15
LMF(2)	0.431	0.6	79.37	78.65
Self_MM(2)	**0.419**	0.616	80.74	80.78
Self_MM	0.4218	0.6092	79.89	79.94
Ours	0.423	**0.6198**	**81.25**	**81.25**

$$A^T U_a \varepsilon_a U_a^T A = U_{v[1:R]} \varepsilon_{v[1:R]} U_{v[1:R]}^T \tag{27}$$

$$(AU_a^T)^T \varepsilon_a (U_{.a}^T A) = E^T \varepsilon_a E \tag{28}$$

$$U_a^T A = E \tag{29}$$

So:

$$A = U_a E = (U_a \varepsilon_a^{+1/2} U_a^T)(\ U_{v[1:R]} \varepsilon_{v[1:R]}^{+1/2} U_{v[1:R]}^T) \tag{30}$$

In summary, it is clear that there is an optimal solution for this function.

H. Overall optimization objective function

The overall optimization objective function is divided into three categories, one for multimodal task, one for unimodal tasks, and one for modal deep feature alignment task.

For multimodal task:

$$l_1 = |y_{all}^{output\_i'} - y_{all}^{output\_i}| \tag{31}$$

For unimodal tasks:

$$l_2 = \sum_{s}^{\{t,a,v\}} \omega_s^i |y_s^{output\_i'} - y_s^{output\_i}| \tag{32}$$

$$\omega_s^i = tanh(|y_s^{output\_i'} - y_{all}^{output\_i}|) \tag{33}$$

For modal deep feature alignment task:

$$l_3 = \frac{1}{4d^2} ||C_a - C_v||_F^2 \tag{34}$$

The overall optimization objective function is:

$$L = \frac{1}{N} \sum_{i}^{N} (l_1 + l_2) + \ l_3 \tag{35}$$

where N is the number of training samples.

IV. EXPERIMENTAL SETUP

In this section, we provide a detailed account of the parameter settings used in our experiments, as well as the experimental setup data, including the datasets and baseline models employed. Our aim is to conduct a comparative study of our model with other existing models on three different baseline datasets for multimodal sentiment analysis. This will help establish the robustness and efficacy of our model in handling multimodal sentiment analysis tasks.

A. Introduction to the dataset

MOSI. CMU-MOSI [43] is a baseline dataset for multimodal sentiment analysis created by Zadeh et al. The dataset comprises multimodal observational data, including audio transcribed textual information and visual modal character gestures, as well as audio features. It also provides opinion-level subjective segmentation. The dataset features 93 YouTube videos of 89 English-speaking speakers, including 41 females and 48 males. Emotional intensity definitions range from strongly negative to strongly positive, with a linear scale of -3 to 3.

MOSEI. CMU-MOSEI [44] is the most extensive collection of data available for sentiment analysis and emotion recognition. The dataset comprises monologue videos of speakers, which were captured through YouTube, using face detection technology. With more than 1000 speakers and 250 testers, the dataset offers a whopping 65 hours of video content. It contains 3,228 videos and 23,453 sentences, covering various topics, including 250 topics such as product and service evaluations and topic debates. The dataset is diverse in its content and is an excellent resource for research on sentiment analysis and emotion recognition.

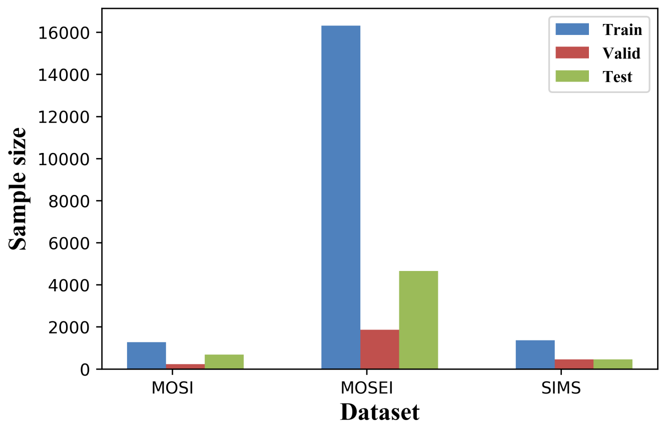

Fig. 3. Flowchart of the complete model architecture.

SIMS. SIMS [41] is a novel Chinese multimodal sentiment analysis dataset proposed by Yu et al. This dataset contains 60 original videos, from which 2281 video clips were extracted. SIMS boasts a rich character background with a wide age range, and it is of very high quality. The dataset includes a range of emotional intensities, spanning from strongly negative to strongly positive, and linearity scales from -1 to 1.

For these three datasets the sample division is shown below.

B. Baseline introduction

TFN [40]. The Tensor Fusion Network (TFN) uses a tensor fusion approach to model intermodal dynamics and learn intra- and intermodal dynamics end-to-end. The intra-modal dynamics are modeled by three modal embedding sub-networks, representing inter-modal interaction states descriptively.

LMF [45]. The low-rank Multimodal Fusion (LMF) focuses on modal fusion, with the weights decomposed into low-rank factors to effectively reduce the number of model parameters. The multimodal fusion of tensor representations is achieved using parallel decomposition of the low-rank weight tensor and the input tensor.

MNF [46]. The Memory Fusion Network (MFN) processes each modality singularly through an LSTM network. The Delta-memory Attention Network (DMAN) module is used to learn the interactions of cross-modal information, which is then stored in a multi-view gated memory module.

RAVEN [47]. The Recurrent Attended Variation Embedding Network (RAVEN) takes into account the different sampling rates of each modality and the long time dependency between them. To address these issues, RAVEN employs a Cross-modal Transformer module.

MulT [11]. The Multimodal Transformer (MulT) extracts regional vision features through Faster RCNN as fictitious word elements, which are then input into the multimodal self-attentive layer along with the text modality to adjust attention under the guidance of text.

MAG-BERT [48]. The Multimodal Adaptation Gate for Bert (MAG-BERT) maps multimodal information onto a vector using a tensor-based approach to deep model fusion, allowing models to learn from large amounts of data in an end-to-end fashion.

MISA [20]. Modality-invariant and specific representations (MISA) consist of two phases: modality feature learning and modality fusion. Features are extracted to learn modality representations under different subspaces in different modalities, and finally, the modality fusion of these representations is performed using Transformer.

Self-MM [12]. The Self-Supervised Multi-task Multimodal sentiment analysis network (Self-MM) designs a single-modal label generation module based on a self-supervised strategy to help multimodal tasks shift more attention to samples with greater modal variability in the multimodal task.

C. Model evaluation parameters

We validate our model with a regression task and a classification task, respectively. For the regression task, Mean Absolute Error (MAE) and Pearson Correlation (Corr) are used as performance evaluation parameters. For classification, the weighted F1-Score (F1-Score) and the Binary Classification Accuracy (Acc-2) are used as performance evaluation parameters.

V. Results and Discussion

Table 1 presents the results of our experiments on the multimodal sentiment analysis datasets (CMU-MOSI and CMU-MOSEI) of the English corpus, with annotations indicating whether the data is aligned or not. Our results demonstrate that aligned data can simplify the multimodal sentiment analysis task, while the use of unaligned data increases its complexity and difficulty.

As shown in Table 1, our models achieved significant improvements compared to the unaligned models, and were highly competitive even when compared to the data-aligned models. Our model outperformed many advanced multimodal sentiment analysis models in recent years, achieving the best results in the evaluation of many parameters. These results confirm the effectiveness of our model.

We also evaluated our model on the newly released multimodal sentiment analysis dataset (SIMS) for the Chinese corpus, which does not contain aligned data. Table 2 compares our model against three existing advanced multimodal sentiment analysis models based on unaligned data, and shows that our model outperforms TFN, LMF, and Self-MM in various metrics.

To further explore the performance and possibilities of our deep modal shared information learning module, we conducted additional experiments on the three baseline datasets. The results are presented in Table 3 and Table 4, where "A-B" indicates that the module enables the network to learn the shared information of modality A and modality B, "A+B" indicates that the module enables the network to learn the private information of each modality A and B, and "A-B/B+C" indicates that the module is used twice, enabling the network to learn both the shared information of modes A and B, and the private information of modes B and C.

TABLE III
EXPERIMENTAL RESULTS OF REGRESSION TASKS AND CLASSIFICATION TASKS BASED ON DIFFERENT USAGE OF MODULES ON MOSI AND MOSEI DATASETS.

Model	MOSI				MOSEI			
	MAE	Corr	Acc-2	F1-Score	MAE	Corr	Acc-2	F1-Score
V-A	**0.704**	0.794	**83.65**/85.06	82.51/**85.43**	**0.523**	**0.766**	**82.98**/85.01	**83.26**/84.89
T-A	0.716	0.722	82.71/84.60	82.62/84.58.	0.529	**0.766**	81.16/84.35	81.57/84.28
T-V	0.722	0.793	82.54/83.99	83.06/84.02	0.535	0.759	79.77/84.07	80.35/84.07
V+A	0.716	**0.798**	82.97/84.51	82.94/84.53	0.531	**0.766**	82.97/84.51	82.94/84.53
T+A	0.882	0.744	81.75/83.57	81.63/83.52	0.76	0.329	70.30/68.36	66.02/62.58
T+V	0.881	0.746	81.60/81.45	81.43/83.36	0.748	0.407	72.00/71.48	70.05/68.41
T-V/T-A	0.719	0.765	82.96/85.30	**83.24**/85.18	0.832	0.765	82.96/85.30	83.24/**85.18**
T+V/T+A	0.908	0.706	81.37/83.69	80.95/83.38	0.811	0.264	65.11/65.93	65.31/64.90
T+V/T-A	0.723	0.793	83.29/84.94	83.23/84.92	0.54	0.76	80.35/84.44	80.92/84.45
T-V/T+A	0.715	0.797	82.83/84.60	82.77/84.60	0.539	0.76	82.66/**85.33**	82.96/85.21
T-A/V-A	0.719	0.792	82.57/84.30	82.49/84.28	0.832	0.765	82.96/85.30	83.24/85.18
T-V/V+A	0.711	0.795	83.09/**85.15**	82.97/85.10	0.535	0.762	77.96/83.57	78.74/83.64
T-A/V+A	0.712	0.795	82.68/84.54	82.58/84.51	0.533	0.766	78.84/84.12	79.59/84.19
T-V/T-A/A-V	0.72	0.791	82.68/84.51	82.59/84.48	0.54	0.76	80.35/84.44	80.92/84.45
T+A/T+V/A+V	0.906	0.708	81.43/83.75	81.03/83.46	0.811	0.245	69.93/66.33	66.06/61.06
T-A/T-V/V+A	0.721	0.792	82.77/84.63	82.68/84.61	0.535	0.762	77.96/83.57	78.74/83.64

Based on Tables 3 and 4, the evaluation metrics demonstrate an improvement compared to the previous model, regardless of how the module is used. According to Yang Wu et al. [49], the text modality is the core modality in multimodal sentiment analysis tasks, while the non-text modality plays a more complementary role. ”V-A” enables the network to learn the shared information of visual and audio modalities, and it extracts more meaningful textual complementary information such as the demeanor and tone of voice in a particular emotion that would not be possible without using the module. However, some approaches such as ”T-V/V+A” demonstrated strong capability on the MOSI dataset but exhibited poor results on the MOSEI dataset. This can be attributed to the fact that the MOSI dataset has fewer topics and requires less generalization performance of the model compared to MOSEI. For the MOSEI dataset, which requires higher generalization performance, making the modal data search for more private information is not conducive for the model to capture the important information in the modalities.

In this paper, we introduce a novel deep modal information sharing module and employ a self-supervised strategy for multi-task learning to facilitate the multimodal sentiment analysis network in learning both shared and private information of the modalities. Our approach not only significantly reduces manual annotation costs but also considerably enhances the performance of the model in multimodal sentiment analysis tasks.

Furthermore, we aim to stimulate further research in the area of shared and private modality information representation, as we believe it is a crucial and promising avenue for exploration. Our approach enhances the interpretability of multimodal sentiment analysis tasks, making it a valuable contribution to the field.

Overall, our work demonstrates the potential benefits of incorporating shared and private information in multimodal sentiment analysis tasks. We hope that our approach will inspire further research in this area and lead to more effective and interpretable multimodal sentiment analysis models.

One limitation of our multimodal sentiment analysis model is its reliance on uniform multimodal labels, which may not always be available or appropriate for certain tasks. Additionally, the raw feature fusion approach used in our model may not always be the most effective way to combine modalities. Future research could explore alternative approaches to multimodal fusion, such as attention mechanisms or graph-based methods.

Another potential area for improvement is the label generation module based on self-supervised learning. While this approach has shown promise in capturing private information, it may not be optimal for all tasks or datasets. Future research could investigate alternative methods for capturing private information, such as adversarial training or unsupervised learning techniques.

Although our model outperforms current state-of-the-art methods on the three public datasets tested, it is important to evaluate its performance on a wider range of datasets and tasks to ensure its generalizability. Future research could explore the use of our deep modal shared information learning module in other multimodal tasks beyond sentiment analysis.

ACKNOWLEDGMENT

This work was supported in part by Joint found for smart computing of Shandong Natural Science Foundation under Grant ZR2020LZH013; open project of State Key Laboratory of Computer Architecture CARCHA202001; the Major Scientific and Technological Innovation Project in Shandong Province under Grant 2021CXG010506 and 2022CXG010504; ”New University 20 items” Funding Project of Jinan under Grant 2021GXRC108 and 2021GXRC024.

REFERENCES

[1] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, “Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research,” *IEEE Transactions on Affective Computing*, 2020.

[2] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, “Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8992–8999.

[3] G. Vinodhini and R. Chandrasekaran, "Sentiment analysis and opinion mining: a survey," *International Journal*, vol. 2, no. 6, pp. 282–292, 2012.

[4] J. Arevalo, T. Solorio, M. Montes-y Gomez, and F. A. González, "Gated multimodal networks," *Neural Computing and Applications*, vol. 32, pp. 10 209–10 228, 2020.

[5] J. O. Egede, S. Song, T. A. Olugbade, C. Wang, C. D. C. Amanda, H. Meng, M. Aung, N. D. Lane, M. Valstar, and N. Bianchi-Berthouze, "Emopain challenge 2020: Multimodal pain evaluation from facial and bodily expressions," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 2020, pp. 849–856.

[6] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Systems*, vol. 244, p. 108580, 2022.

[7] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition based on manifold learning and convolution neural network," *Multimedia Tools and Applications*, vol. 81, no. 23, pp. 33 253–33 268, 2022.

[8] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, 2022.

[9] Z. Xiaoming, Y. Yijiao, and Z. Shiqing, "Survey of deep learning based multimodal emotion recognition," *Journal of Frontiers of Computer Science & Technology*, vol. 16, no. 7, p. 1479, 2022.

[10] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: challenges, applications with datasets, recent advances and future directions," *Information Fusion*, vol. 81, pp. 203–239, 2022.

[11] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[12] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[13] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[14] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[15] D. Kim, K. Wang, S. Sclaroff, and K. Saenko, "A broad study of pre-training for domain generalization and adaptation," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*. Springer, 2022, pp. 621–638.

[16] C. Zhao and W. Shen, "A domain generalization network combing invariance and specificity towards real-time intelligent fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 173, p. 108990, 2022.

[17] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2539–2544.

[18] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[19] W. Wu, Y. Wang, S. Xu, and K. Yan, "Sfnn: Semantic features fusion neural network for multimodal sentiment analysis," in *2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE)*. IEEE, 2020, pp. 661–665.

[20] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.

[21] C. Edwards, "The best of nlp," *Communications of the ACM*, vol. 64, no. 4, pp. 9–11, 2021.

[22] M. Koroteev, "Bert: a review of applications in natural language processing and understanding," *arXiv preprint arXiv:2103.11943*, 2021.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[24] M. Müller, M. Salathé, and P. E. Kummervold, "Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter," *arXiv preprint arXiv:2005.07503*, 2020.

[25] Y. Yu, X. Si, C. Hu, and J. Zhang, "A review of recurrent neural networks: Lstm cells and network architectures," *Neural computation*, vol. 31, no. 7, pp. 1235–1270, 2019.

[26] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020.

[27] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with lstm," *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

[28] Z. Zhao, W. Chen, X. Wu, P. C. Chen, and J. Liu, "Lstm network: a deep learning approach for short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, 2017.

[29] K. Park, Y. Choi, W. J. Choi, H.-Y. Ryu, and H. Kim, "Lstm-based battery remaining useful life prediction with multi-channel charging profiles," *Ieee Access*, vol. 8, pp. 20 786–20 798, 2020.

[30] Z. Shen, J. Liu, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624*, 2021.

[31] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE transactions on neural networks*, vol. 22, no. 2, pp. 199–210, 2010.

[32] Y. Xu, J. Liu, Z. Wan, D. Zhang, and D. Jiang, "Rotor fault diagnosis using domain-adversarial neural network with time-frequency analysis," *Machines*, vol. 10, no. 8, p. 610, 2022.

[33] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[34] ——, "An overview of multi-task learning," *National Science Review*, vol. 5, no. 1, pp. 30–43, 2018.

[35] O. Sener and V. Koltun, "Multi-task learning as multi-objective optimization," *Advances in neural information processing systems*, vol. 31, 2018.

[36] B. Yang, L. Wu, J. Zhu, B. Shao, X. Lin, and T.-Y. Liu, "Multimodal sentiment analysis with two-phase multi-task learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2015–2024, 2022.

[37] Y. Zhang, L. Rong, X. Li, and R. Chen, "Multi-modal sentiment and emotion joint analysis with a deep attentive multi-task learning model," in *Advances in Information Retrieval: 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10–14, 2022, Proceedings, Part I*. Springer, 2022, pp. 518–532.

[38] Y. Song, X. Fan, Y. Yang, G. Ren, and W. Pan, "A cross-modal attention and multi-task learning based approach for multi-modal sentiment analysis," in *Artificial Intelligence in China: Proceedings of the 3rd International Conference on Artificial Intelligence in China*. Springer, 2022, pp. 159–166.

[39] D. S. Chauhan, S. Dhanush, A. Ekbal, and P. Bhattacharyya, "Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4351–4360.

[40] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[41] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 3718–3727.

[42] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 443–450.

[43] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *arXiv preprint arXiv:1606.06259*, 2016.

[44] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[45] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.

[46] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.
[47] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L.-P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7216–7223.
[48] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020. NIH Public Access, 2020, p. 2359.
[49] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 4730–4738.