Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 2105593, 14 pages
https://doi.org/10.1155/2022/2105593

*Research Article*

# Dynamic Invariant-Specific Representation Fusion Network for Multimodal Sentiment Analysis

**Jing He, Haonan Yanga, Changfan Zhang, Hongrun Chen, and Yifu Xua**

*College of Electrical and Information Engineering, Hunan University of Technology, Zhuzhou 412007, China*

Correspondence should be addressed to Changfan Zhang; zhangchangfan@263.net

Received 29 November 2021; Revised 31 December 2021; Accepted 6 January 2022; Published 24 January 2022

Academic Editor: Zhongxu Hu



Multimodal sentiment analysis (MSA) aims to infer emotions from linguistic, auditory, and visual sequences. Multimodal information representation method and fusion technology are keys to MSA. However, the problem of difficulty in fully obtaining heterogeneous data interactions in MSA usually exists. To solve these problems, a new framework, namely, dynamic invariant-specific representation fusion network (DISRFN), is put forward in this study. Firstly, in order to effectively utilize redundant information, the joint domain separation representations of all modes are obtained through the improved joint domain separation network. Then, the hierarchical graph fusion net (HGFN) is used for dynamically fusing each representation to obtain the interaction of multimodal data for guidance in the sentiment analysis. Moreover, comparative experiments are performed on popular MSA data sets MOSI and MOSEI, and the research on fusion strategy, loss function ablation, and similarity loss function analysis experiments is designed. The experimental results verify the effectiveness of the DISRFN framework and loss function.

## 1. Introduction

Multimodal sentiment analysis (MSA), as an emerging field of natural language processing (NLP), aims to infer the speaker's emotion by exploring clues in multimodal information [1–3]. Many methods in MSA focus on exploring the complex fusion mechanism to improve the performance of MSA [4–6]. However, these fusion technologies present a bottleneck due to the difficulty in obtaining interaction between heterogeneous modes. The common method to solve this problem is to map the heterogeneous feature to the common subspace in the representation learning process [7]. However, some unique features of each mode are ignored by those methods. These unique features can be used as complementary information between modes. Effective use of this complementary information can help the network improve performance. For this consideration, this paper intends to use supplementary information on the basis of shared representation. And then, a dynamic fusion mechanism is established to fuse the modal features to obtain the interactive information. This study mainly aims to explore a sentiment analysis framework based on multimodal representation learning and the dynamical fusion method.

For multimodal representation learning methods, since multimodal data is usually a sequence with different feature dimensions, long-short memory neural network (LSTM) is a powerful tool to deal with such problems [8]. Therefore, different LSTMs are used to extract features of different modalities in many methods, such as memory fusion network (MFN) [9], graph-memory fusion network (Graph -MFN) [10]. However, a single LSTM is difficult to apply to the feature distribution of each mode at the same time. Therefore, there are studies using different networks to represent different modal information, such as tensor fusion network (TFN) [11], low-rank multimodal fusion net (LMF) [12]. It is worth mentioning that the information between modalities was not used fully before fusion in these methods. The shared features and special features of two data sources are captured by domain separation network (DSN) using adversarial learning and soft orthogonal constraint [13]. And then, these features are used to perform domain adaptive

tasks. The combination of shared features and special features can effectively solve the problem that the redundant information between different data sources is not fully utilized. In other words, the DSN is improved and adopted to perform multimodal sentiment analysis tasks in this paper. It is named improved joint domain separation network (improved JDSN).

In this paper, the improved JDSN is adopted to learn the joint representation of modality-invariant and modality-specific of all modes in the common–special subspace. The former aims to map all the modes of discourse to the common subspace to shorten the distance between modes to effectively reduce the extra burden of fusion work. The latter aims to extract special representation from each mode as complementary information. Then, the combination of two representations can fully use the complementary information between modes. In addition, the modal interactions were mostly obtained by feature connection fusion in early work [14]. However, these methods are unable to dynamically adjust the contribution of each mode in the fusion process. Mai et al. assumed that the multimodal fusion process is a hierarchical interactive learning process [15, 16] and designed a ARGF network to solve the problem [15]. The ARGF was comprised of two stages: a joint embedding space learning stage and a hierarchical graph fusion net (HGFN) stage. In the HGFN stage, firstly, the unimodal dynamic layer, bimodal dynamic layer, and trimodal dynamic layer are modelled, and then the outputs of each dynamic layer are connected to obtain the interaction features of each mode. However, the method of joint embedding space learning also has a problem that the redundant information was not fully utilized. Therefore, the improved JDSN and HGFN are combined to optimize the network's ability to capture modal interactions by rationally using redundant information in this paper.

In summary, firstly, the applied DSN in this paper is improved in the aspects of the following: (1) The mode of DSN is extended; (2) The orthogonal constraint loss between special representations of different modes is additionally considered (See Section 3.3.1); (3) Adversarial loss is replaced by a more advanced similarity metric (CMD) (See Section 3.3.2); (4) Invariant and specific representation are jointed at the output of the network (see Section 3.2.3). Then, combining the improved JDSN and HGFN, a new framework (DISRFN) is proposed in this paper to deal with MSA problems. The main contributions are as follows:

(1) A multimodal sentiment analysis framework (DISRFN) is proposed in this study. It can perform the fusion of various representations dynamically while emphasizing learning invariant and specific joint representations of various modes.

(2) A new loss function is designed, which can improve the effect of semantic fusion clustering whilst assisting the model in learning the target subspace representation effectively.

(3) The performance analysis experiments of MSA tasks is designed on the benchmark data sets MOSI and MOSEI. The results confirm the advancement of the DISRFN model and fusion strategy, the effectiveness of the loss function, and the rationality of similarity loss function selection.

The remainder of this paper consists of the following parts. In Section 2, the correlation work is briefly reviewed. Section 3 introduces the structure of the DISRFN model and the proposed learning method in detail. Section 4 explains the experimental details, parameter settings, and network component design. The experimental results are analyzed in Section 5. Section 6 shows the summary and prospects.

## 2. Correlation Work

In multimodal sentiment analysis, the mainstream multimodal learning methods include multimodal fusion representation and multimodal representation learning, which will be discussed in this section.

*2.1. Multimodal Fusion Representation.* In recent years, some complex and efficient fusion representation mechanisms have been gradually proposed. Amir Zadeh et al. put forward TFN to obtain the trimodal fusion representations by using the outer product [11]. On this basis, a low-rank multimodal fusion net (LMF) was proposed. This network performs multimodal data fusion employing a low-rank tensor and obtains better results [12]. Mai et al. proposed a strategy "divide and rule, unite many into one" to transfer local tensor and global fusion, which was extended in multiconnected bidirectional long-short time memory network (Bi-LSTM) [17, 18]. In addition to the tensor fusion method, the recursive fusion method has been developed better. For example, a recursive multilevel fusion network (RMFN) is used for specialized and effective fusion through decomposing the fusion problems into several parts [19]. The more attention-based recursive network (MARN) is used to fuse cyclic memory representations of different modes of long-short term hybrid memory networks (LSTHM) by using a more attention block [20]. Hierarchical polynomial fusion network (HPFN) is used to recursively integrate and transfer the local correlation to the global correlation through multilinear fusion [21]. Moreover, the multiview learning method plays an important role in multimodal fusion [22]. For example, MFN designed by Amir Zadeh et al. is used to fuse the memory of different modes of LSTM system based on incremental attention memory network (DAMN) and gated memory network (MVGN) [9], and it is successfully used to solve multiview problems. Furthermore, to analyze the explainability of MFN, the dynamic fusion graph model (DFG) is embedded into MFN, and a Graph-MFN obtained finally has excellent performance and is explainable [10]. Recently, word-level fusion representation has also been a wide concern [23]. For example, a repeated participation variation network (RAVEN) is used to model multimodal language through work representation transfer based on facial expression [24]. Chen et al. modeled the time-dependent multimodal dynamics through cross-modal work alignment [25]. However, most

of these methods use complex fusion mechanisms or add additional fusion modules, which will increase the amount of calculation and slow down the speed of network convergence. In contrast, this paper uses a hierarchical mechanism to model the dynamics of each fusion layer, which can quickly fuse the information of each mode.

*2.2. Multimodal Representation Learning.* Multimodal representation learning is mainly divided into two types, namely, common subspace representations and factorised representations. The two types of study on common subspace representations amongst modes are the correlation-based model and adversarial learning-based model. In terms of a correlation-based model, Shu et al. proposed an extensible multilabel canonical correlation analysis (sml-CCA) for cross-modal retrieval [26]. Kaloga et al. proposed a multiview graph canonical correlation analysis based on variational graph neutral network for classification and clustering tasks [27]. Verma et al. proposed a deep network with high-order information and single sequence information (Deep-HOSeq) for fusing multimodal sentiment data [28]. Mai et al. learned the embedding space within invariant mode based on a new encoding-decoding classifier framework in confrontation [15]. Pham et al. proposed a robust joint representation method to learn by shifting between modes under the constraints of cyclic consistency loss [29]. In terms of the adversarial learning-based model, Wu and Qiang et al. proposed the generative adversarial net based on specific mode and sharing and the adversarial hashing algorithm based on deep semantic similarity, respectively, to obtain cross-modal invariance [30, 31]. However, these methods only learn about the shared representation of the model and lack the consideration of the special representation of the modal. For factorized representations, Amir Zadeh et al. proposed a multimodal factorized model (MFM) to factorize multimodal representations into multimodal discriminant factor and multimodal special generation factor [32]. Liang et al. proposed a multimodal baseline model (MMB) to learn the cases of multimodal embedding based on the factorized method [33]. Wang et al. proposed a joint and separate matrix factorized hashing method, which could be used to learn common and specific attributes of multimodal data at the same time [34]. Fang et al. proposed a new semantic enhanced discrete matrix factorized hashing (SDMFN), which could directly extract the common hashing representation from the reconstructed semantic polynomial similar graph, causing the hash code to be more discriminative [35]. Caicedo et al. proposed a multimodal image representation based on nonnegative matrix factorisation to synthesise visual features and text features [36]. However, most of these factorized methods adopt the form of matrix decomposition, which may have the problem of incomplete feature representation. In contrast, the improved JDSN designed in this paper can obtain a richer shared-special representation of each mode in a simpler way.

## 3. The Proposed Method

*3.1. Task Setting.* In general, the proposed framework is mainly used to study the trimodal data. Figure 1 shows the flowchart of the proposed multimodal fusion framework. This framework consists of two parts, as follows: (1) improved JDSN for learning trimodal data-specific shared subspace joint representation; (2) HGFN for fusing trimodal joint representation, thereby realizing dynamical effective semantic clustering. This study introduces this network framework in the following section.

Moreover, the discourse data are divided into N sequences composed of segment S to facilitate detecting emotion in video by using multimodal data. Each segment S includes three low-level feature sequences in linguistic (l), visual (v), and auditory (a) modes. These feature sequences are represented as $S_l \in \mathbb{R}^{t_l \times d_l}, S_v \in \mathbb{R}^{t_v \times d_v}, S_a \in \mathbb{R}^{t_a \times d_a}$. Amongst them, $t_m$ and $d_m$ ($m \in \{l, v, a\}$) represent the length of discourse and the dimension of the corresponding feature, respectively. Given this data sequence, the study aims to predict the emotional state of the predefined set. This emotional state is a continuous dense variable $y \in \mathbb{R}$. In addition, to effectively use multimodal data, linguistic (l), visual (v), and auditory (a) trimodal feature sequences, they should be aligned with emotional state label $y$.

The framework of DISRFN is shown in Figure 1: (1) The data of the three modes are fed into the corresponding Bi-LSTM and BERT models to obtain the discourse-level feature representations; (2) The discourse-level feature representations of each mode are fed into the corresponding MLP to obtain the representation of unified dimension; (3) The unified representations of each mode are fed into the corresponding encoder and shared encoder to obtain the shared representations and special representations; (4) The shared representations are added with a special representation of each modal to obtain the joint domain separation representations; (5) The joint domain separation representations of each mode are fed into the corresponding decoder to obtain the reconstruction loss; (6) The joint domain separation representations of each mode are fed into HGFN for dynamic fusion to perform MSA task.

### 3.2. Dynamic Invariant-Specific Representation Fusion Network

*3.2.1. Discourse-Level Feature Representation.* Firstly, the stacking bidirectional long-short time memory neural network (sLSTM) is used to map the feature sequence ($S_v$, $S_a$) in visual ($v$) and auditory ($a$) modes to obtain the underlying features of the sequence. Its output includes the hidden representations of LSTM end state, namely, $F_v$ and $F_a$, as follows:

$$\begin{aligned} F_v &= s\mathrm{LSTM}\left(S_v; \theta_v^{\mathrm{LSTM}}\right), \\ F_a &= s\mathrm{LSTM}\left(S_a; \theta_a^{\mathrm{LSTM}}\right), \end{aligned} \tag{1}$$

where $\theta_v^{\mathrm{LSTM}}$ and $\theta_a^{\mathrm{LSTM}}$ refer to the parameters of sLSTM on visual and auditory modes.

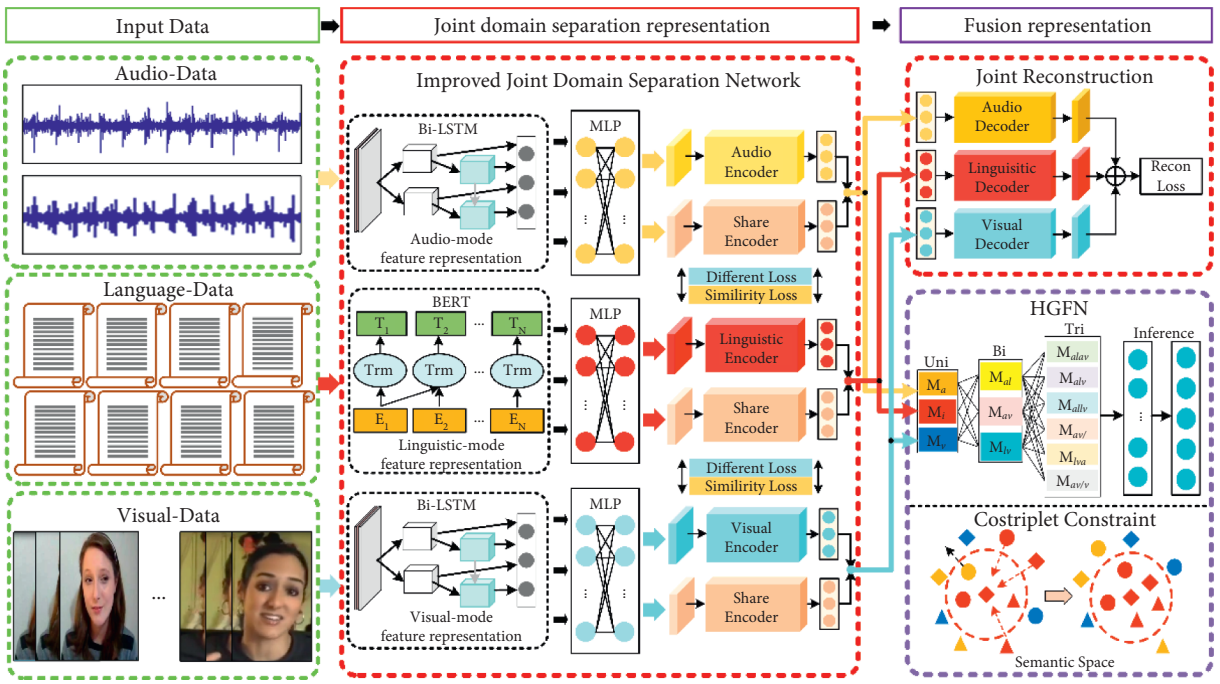


FIGURE 1: The framework of DISRFN. Note: Bi-LSTM: bidirectional short and long memory network; BERT: bidirectional encoder representation from transformers; MLP: multilayer perception; audio encoder (decoder): encoder (decoder) of auditory mode; linguistic encoder (decoder): encoder (decoder) of linguistic mode; visual encoder (decoder): encoder (decoder) of visual mode; share encoder: shared encoder of three modes; HGFN: hierarchical graph fusion net.

Secondly, for the text feature sequence ($S_l$) in linguistic mode, most linguistic features are embedded through Glove [37]. However, in recent studies [38], such as the advanced ICCN [39] model, the pretraining BERT model is used as the feature extractor of text discourse. A better result than the Glove method is obtained. Therefore, the feature representation $F_l$ of text is obtained through the pretraining BERT model, as follows:

$$F_l = \mathrm{BERT}\left(S_l; \theta_l^{\mathrm{BERT}}\right), \tag{2}$$

where $\theta_l^{\mathrm{BERT}}$ refers to the parameter of the BERT model.

*3.2.2. Unified Representation of Features.* The dimensions of discourse-level features are different. In order to facilitate the encoding-decoding operation in the back-end network, multilayer perceptron (MLP) is used to unify mapping these features to $O_m$, as follows:

$$O_m = \mathrm{MLP}\left(F_m; \theta_m^{MLP}\right), (m \in \{l, v, a\}), \tag{3}$$

where $\theta_m^{\mathrm{MLP}}$ refers to a parameter of multilayer perceptron networks in different modes; MLP consists of dense connection layers and a normalized layer activated by relu function.

*3.2.3. Improved Joint Domain Separation Representation.* In this part, based on the improved JDSN, the unified mapping representation of each mode is factorized into two parts, namely, modality-invariance and modality-specificity. Amongst them, the sharing encoder $E^c$ is used to learn invariant representation in the common subspace to narrow the gap in the heterogeneity between modes [40]. The specific encoder $E_m^p$ is used to capture the specific representation in a specific subspace. The process is as follows.

Firstly, after obtaining the unified mapping vector $O_m$ of each mode, the mode-sharing encoder $E^c$ (weight sharing) is used to obtain modality-invariant representation ($h_m^c$), and the mode-specific encoder $E_m^p$ is used to extract modality-specific representation ($h_m^p$), as follows:

$$h_m^c = E^c\left(O_m; \theta^c\right), h_m^p = E_m^p\left(O_m; \theta_m^p\right), (m \in \{l, v, a\}), \tag{4}$$

where $\theta^c$ refers to a parameter of mode-sharing encoder; $\theta_m^p$ refers to a parameter of mode-specific encoder; $E^c$ has the same structure as that of $E_m^p$, which is composed of a dense connection layer activated by sigmoid function.

Then, hidden layer vectors $h_m^p$ and $h_m^c$ are generated through feedforward propagation of neural network, and the joint domain separation representation is obtained through vector addition "+", as follows:

$$h_m = h_m^c + h_m^p, (m \in \{l, v, a\}), \tag{5}$$

where $h_m$ refers to the joint domain separation representation of mode $m$, and it has the feature representation of shared subspace and specific subspace characteristics.

*3.2.4. Hierarchical Graph Fusion Representation.* After obtaining the joint domain separation representation of each mode, it is necessary to fuse each representation to obtain the interaction information of each mode.

As shown in Figure 2, HGFN is composed of three dynamic layers (unimodal dynamic layer, bimodal dynamic layer, and trimodal dynamic layer). Unimodal dynamic layer is modeled by self-attention weighting each unimodal information vector. Bimodal dynamic layer is modeled by weighting bimodal information vectors (e.g., $M_{al}$) using the correlation weight between unimodal vectors. Trimodal dynamic layer is constructed through weighting trimodal information vectors (e.g., $M_{alv}$ or $M_{allv}$) by the correlation weight between unimodal vectors. Finally, three dynamic layers are used for vector connection and fusion to realize the dynamic fusion of multimodal features in HGFN. This hierarchical modeling method is more conducive to exploring the interaction between modes [12]. Therefore, HGFN, which can preserve all modal interactions, is introduced to fuse the obtained joint domain separation representations of different modes to explore multimodal interaction in this section. The fusion representation is as follows:

$$\text{Fusion} = \text{HGFN}\left(h_l, h_v, h_a; \theta^{\text{HGFN}}\right), \tag{6}$$

where "Fusion" refers to the output of HGFN; $\theta^{\text{HGFN}}$ refers to the parameters of HGFN. Then, the predictive neural network (P) is used for prediction, as follows:

$$\text{Pred} = P\left(\text{Fusion}; \theta^{\text{Pre}}\right), \tag{7}$$

where "Pred" refers to the output of the predictive network; "P" refers to a predictive network, including a standardized layer and the fully connected layers; $\theta^{\text{Pre}}$ refers to the parameter of the predictive network. Moreover, the specific parameters of the model are described in the experimental section.

*3.3. Learning Process.* A joint loss function is newly set to effectively learn the network model, as follows:

$$L_{\text{total}} = L_{\text{task}} + \alpha L_{\text{diff}} + \beta L_{\text{sim}} + \gamma L_{\text{recon}} + \eta L_{\text{trip}}, \tag{8}$$

where $\alpha$, $\beta$, $\gamma$, and $\eta$ refer to weights of the interaction. They determine the contributions of each loss $L_{\text{diff}}$, $L_{\text{sim}}$, $L_{\text{recon}}$, and $L_{\text{trip}}$ to total loss $L_{\text{total}}$. In addition, each loss is analyzed and introduced in the remaining section.

*3.3.1. Differential Loss.* Some studies have shown that a nonredundant effect can be achieved by applying soft orthogonality constraint to two representation vectors [13, 41]. Therefore, the constraint is used to drive the sharing-encoder $E^c$ and specific-encoder $E_m^p$ to perform encoding representation to different aspects, that is, modality-invariant and modality-specific representations. Soft orthogonality constraint is defined as follows.

When training a batch of data, $H_m^c$ and $H_m^p$ are set as the two matrices, respectively. The rows of the two matrices correspond to invariant representation $h_m^c$ and specific representation $h_m^p$ of mode $m$ in each batch of data, respectively. The orthogonality constraint of the modal vector is calculated as follows [13]:

$$L_{\text{diff}} = \sum_{m\in\{l,v,a\}} \left\| H_m^{c\text{T}} H_m^p \right\|_F^2 + \sum_{\substack{(m_1,m_2)\in\{(l,a),\\(l,v),(a,v)\}}} \left\| H_{m_1}^{p\text{T}} H_{m_2}^p \right\|_F^2, \tag{9}$$

where $||\cdot||_F^2$ refers to squared Frobenius norm.

*3.3.2. Similarity Loss.* Similarity loss ($L_{\text{sim}}$) used to constrain shared subspace can reduce the difference in the heterogeneity between the shared representations of different modes [42]. Central moment discrepancy (CMD) is used to measure the difference between two distributions by matching order-wise moment differences of two representations [43]. Compared with other methods (e.g., MMD and DANN), it is a more efficient and concise distance measurement. Therefore, CMD is selected as the similarity loss in this paper. It is defined as follows.

$X$ and $Y$ are set as bounded random samples with probability distributions $p$ and $q$ in a compact interval $[a, b]^N$, respectively. CMD is defined as follows [43]:

$$\begin{aligned} \text{CMD}(X, Y) &= \frac{1}{|b-a|}\|E(X) - E(Y)\|_2 + \sum_{k=2}^{K} \frac{1}{|b-a|^k}\left\|C_k(X) - C_k(Y)\right\|_2 \\ C_k(X) &= E\left((x - E(X))^k\right) \\ E(X) &= \frac{1}{|X|}\sum_{x\in X} x, \end{aligned} \tag{10}$$

where $E(X)$ refers to the empirical expectation vector of sample $X$; $C_k(X)$ refers to the vector of all k-order sample centre moments in the $X$ coordinate.

In this paper, the similarity loss is calculated by summing the CMD distances of the shared representations of every two modes. Its representation is as follows:

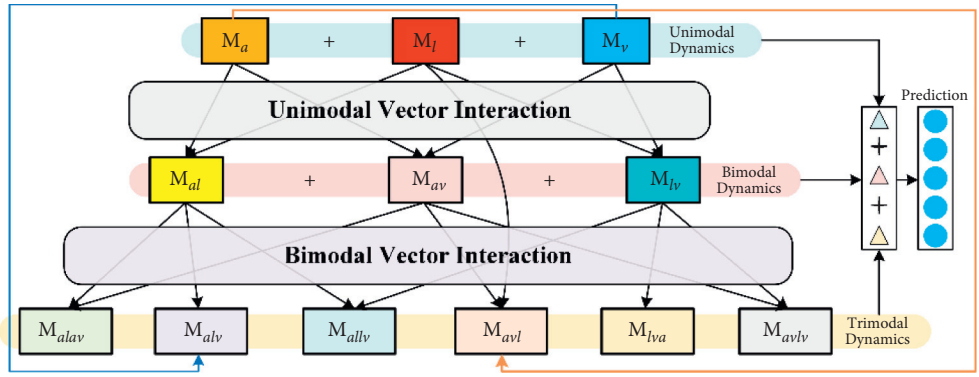


Figure 2: The framework of HGFN.

$$L_{\text{sim}} = \sum_{\substack{(m_1,m_2)\in\{(l,a),\\(l,v),(a,v)\}}} \text{CMD}\left(h^c_{m_1}, h^c_{m_2}\right), \tag{11}$$

Moreover, the reason for selecting CMD as the similarity loss will be discussed in Experimental part 5.4.

*3.3.3. Reconstruction Loss.* When soft orthogonality constraint is enforced, the risk of specific encoder learning trivial representation exists. However, the reconstruction loss can be added to ensure that the encoder can capture the details of each mode to solve these problems [13]. Initially, the modal decoder $D_m$ is used to reconstruct the joint domain separation representation vector $h_m$ of mode $m$, and the output of reconstruction is $\widehat{h}_m$. Then, the reconstruction loss is represented by the mean square error loss between $h_m$ and $\widehat{h}_m$, as follows [13]:

$$L_{\text{recon}} = \frac{1}{3}\left(\sum_{m\in\{l,v,a\}} \left\|h_m - \widehat{h}_m\right\|_2^2\right), \tag{12}$$

where $\|\cdot\|_2^2$ refers to squared $L_2$-norm.

*3.3.4. Cosine Triplet-Margin Loss.* In the fusion representation of joint domain separation representation vector, to ensure the high-level relationship of the similarity between all projects, the representation distance of discourse segments with similar semantics between different modes is minimized through cosine triplet-margin loss $L_{\text{trip}}$, and the distance between different discourse segments is maximized [44].

For example, in linguistic and visual modes, a triple representation $(h_l, h_v^+, h_v^-)$ is established. Amongst them, visual representation $h_v^+$ is positively correlated with linguistic representation $h_l$ in semantics. At the same time, visual representation $h_v^-$ is the contrary. Therefore, the cosine triplet-margin loss of linguistic mode is shown as follows [44]:

$$L^l_{\text{trip}} = \sum_{m\in\{v,a\}} \max\left(\cos\left(h_l, h_m^-\right) - \cos\left(h_l, h_m^+\right) + \text{margin}, 0\right), \tag{13}$$

where $h_m^+, h_m^-$ refers to the joint domain separation representation vector of mode $m$; "margin = 1" is a boundary parameter.

In the same way, the cosine triplet-margin loss of visual mode and auditory mode can be described as follows:

$$L^v_{\text{trip}} = \sum_{m\in\{l,a\}} \max\left(\cos\left(h_v, h_m^-\right) - \cos\left(h_v, h_m^+\right) + \text{margin}, 0\right), \tag{14}$$

$$L^a_{\text{trip}} = \sum_{m\in\{l,a\}} \max\left(\cos\left(h_v, h_m^-\right) - \cos\left(h_v, h_m^+\right) + \text{margin}, 0\right). \tag{15}$$

Based on formulas (13)–(15), the total cosine triple margin loss is represented as follows:

$$L_{\text{trip}} = L^l_{\text{trip}} + L^v_{\text{trip}} + L^a_{\text{trip}}. \tag{16}$$

*3.3.5. Task Loss.* The mean square error (MSE) is used as the task loss of the network to predict continuous dense variables. For $N_b$ discourse data in one batch, this loss calculation is as follows:

$$L_{\text{task}} = \frac{1}{N_b}\sum_{i=0}^{N_b} \left\|y_i - \widehat{y}_i\right\|_2^2. \tag{17}$$

where $y_i$ refers to the actual emotional label; $\widehat{y}_i$ refers to the predictive value of the network.

## 4. Experiment

In this section, the required data sets, evaluation index, and experimental details (experimental environment, experimental parameters, and network structure) are described.

*4.1. Datasets.* The data set is introduced in this section. This data set includes two parts, namely, CMU-MOSI and CMU-MOSEI.

CMU-MOSI data set: this data set is a collection of monologues on YouTube, including videos with 93 comments from different speakers. These common videos consist of 2199 subjective discourses. These discourses are manually

marked with continuous opinion scores in the range of −3 to 3. Amongst them, −3/+3 represents strong negative/positive emotions. A total of 1283 segment samples are used for training, 229 segments are used for verification, and 686 segments are used for testing.

CMU-MOSEI data set: it is an improved version of MOSI; it includes 23453 annotated discourse segments, which are from 5000 videos, 1000 different speakers, and 250 different topics. A total of 1283 segment samples are still used for training, 229 segments are used for verification, and 686 segments are used for testing.

The problems on multimodal signal (linguistic, visual, and auditory) acquisition and modal data pretreatment are solved based on CMU-Multimodal SDK[1] in many studies [45]. This tool library is a machine learning platform used for developing high-level multimodal models and acquiring and processing multimodal data by Amir Zadeh et al. It integrates the acquisition and alignment method of benchmark data sets (MOSI and MOSEI). Similarly, this tool library is used to solve the problems of data acquisition and alignment.

*4.2. Evaluation Index.* This experiment is a regressive task. Therefore, the mean absolute error (MAE) and Pearson correlation coefficient (Corr) are adopted to measure the test results. In addition, the classification index is considered in the experiment, including five-classification accuracy (Acc-5) in affection domain (−2,2), two-classification accuracy (Acc-2) including positive and negative emotion (p/g), and F1score (F1-Score).

*4.3. Experimental Settings.* This method is tested on Pytorch in this section. The grid searching of hyperparameter is performed in a data verification set to identify appropriate hyperparameter, and the best model and hyperparameter are saved. In grid searching, limited option sets for setting hyperparameters are as follows: $\alpha \in \{0.3, 0.4\}$, $\beta \in \{0.7, 0.8, 0.9, 1.0\}$, $\gamma \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, $\eta \in \{0.01, 0.1\}$ and drop$\in$\{0, 0.1, 0.2, 0.3, 0.4\}; the hidden layer sizes of the representation and predictive network can be reviewed from the following: Hid$\in$\{128, 256\}, P_h$\in$\{50, 64\}.

In the iterative optimization process, Adam optimizer with max epoch = 20, batch_size = 16, and learning rate of 0.0001 are used to train the network. The grid searching results of all data sets are shown in Table 1, and based on the hyperparameter settings, Figure 3 shows the model component structural diagram. Note: (1) FC Layer is the dimension of the fully connected layer; (2) LSTM is the dimension of the LSTM hidden layer; (3) Layer-Norm is a dimension of the batch normalization layer; (4) Dropout is the rate of dropout; (5) BERT is the output dimension of the BERT model; (6) Hid/drop/P_h is hyperparameters.

*4.4. Experimental Process.* This section mainly introduces the experimental process, the specific experimental steps are as follows:

Table 1: Hyperparameter settings in this article.

| Hyperparameter | MOSI | MOSEI |
|---|---|---|
| CMD K | 5 | 5 |
| Batch_size | 16 | 16 |
| $\alpha$ | 0.3 | 0.4 |
| $\beta$ | 1.0 | 0.8 |
| $\gamma$ | 0.4 | 0.4 |
| $\eta$ | 0.1 | 0.01 |
| Drop | 0.4 | 0.1 |
| Hid | 256 | 256 |
| P_h | 64 | 50 |

(1) Manual feature extraction of video and audio: for CMU-MOSI and CMU-MOSEI, Facet[2] and COVAREP [46] are used to extract the manual features of visual and auditory sequences. Amongst them, the dimensions $d_v$ of the visual feature are 47 and 35, respectively, and the dimension $d_a$ of the auditory feature is 74.

(2) Discourse-level feature extraction: for linguistic mode, because the BERT model has text embedding and representation functions, the pretraining model of BERT is directly used to extract linguistic features. Its discourse-level feature is represented as feature representation $F_l$ with dimension of 768 [47]. And then, visual and auditory features at the discourse-level $F_v$ and $F_a$ are obtained based on sLSTM.

(3) Unified representation mapping: MLP is adopted to map linguistic, visual, and auditory representation vectors $F_l$, $F_v$, and $F_a$ to an output $O_m$ with the unified dimension size.

(4) Improved joint domain separation representation: $O_m$ is input to sharing encoder and specific encoder to obtain hidden layer representation $h_m^c, h_m^p$. And then, an improved joint domain separation representation $h_m$ is obtained through vector addition $(h_m^p + h_m^c)$.

(5) Fusion inference: the joint domain separation representation vector is sent to the HGFN to perform fusion and prediction tasks.

(6) Calculating loss function and training: loss function is calculated to train the neural network and make cyclic iteration.

## 5. Results and Analysis

Model comparison experiments, research on fusion strategy, research on loss function ablation, and research on similarity loss selection are designed in this section. All experiments are discussed by combining visualization and quantitative analysis.

*5.1. Model Comparison Experiments Result.* In the comparison experiment, some classical models (TFN, LMF, MFN, Gragh-MFN, MARM, and MISA) are reproduced. In addition, some derived fusion model based on LSTHM [17]

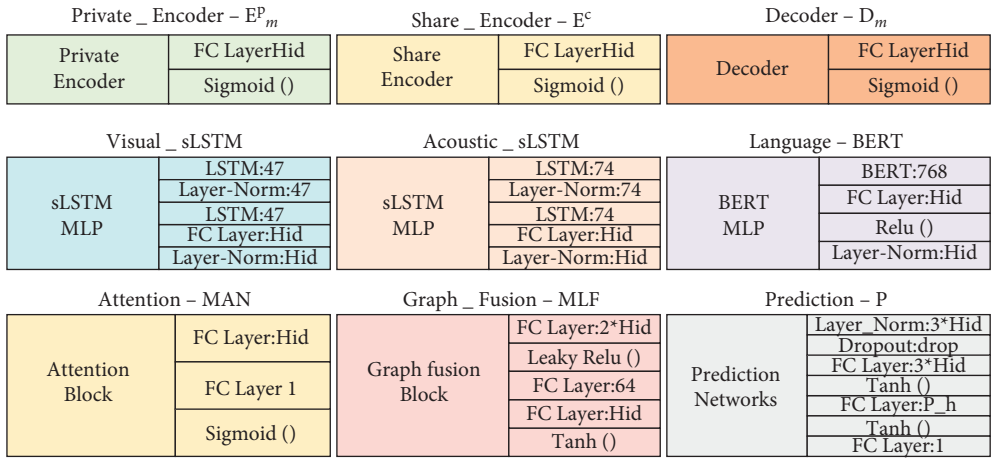


Figure 3: The parameter setting of modules.

is designed to comparison with the proposed framework (DISRFN). The result is shown in Tables 2 and 3.

Tables 2 and 3 show that our method achieves the best performance under two data sets. That is, it exceeds the comparison model in terms of MAE, Corr, Acc, and other comprehensive indexes. These results show that the proposed model exceeds some complex fusion mechanisms (e.g., TFN, MFN, and Gragh-MFN) in the performance. The reason is that these methods ignore the exploration of modal invariant space while the proposed method obtains a joint representation of invariant-specific space.

Moreover, it can be seen from the "CPU Clock" items in Tables 2 and 3. Compared with the model that also applies mechanism fusion (TFN, LMF, MFN, Gragh-MFN, MARM, ARGF, LSTHM-DFG, LSTHM-Out Product), the proposed method is at a disadvantage in the aspect of real-time due to the relatively large number of parameters in the representation learning. However, compared with the model that uses additional networks in the fusion part (MISA, LSTHM-AttFusion, LSTHM -Concat), the proposed method has an advantage when it comes to real-time. Therefore, compared with the baseline model, the proposed method has moderate real-time performance when the various MSA indicators are optimal.

In Section 3.2.1, the reason for using the BERT pretraining model to extract discourse-level features of language modality instead of Glove method is explored. Tables 2 and 3 show that, compared with the baseline model based on the Glove word embedding method, and LSTHM-derived fusion model, various evaluation indexes are improved significantly by the model using BERT (DISRFN and MISA). It proves that the application of the BERT method is reasonable. Moreover, compared with the MISA model using BERT, the proposed model still has a slight advantage. The difference is probably caused by different fusion strategies. The comparative experiment is carried out in the next section to further discuss the effectiveness of the fusion strategy of this model.

*5.2. Fusion Strategy Comparison Result.* In this section, a fusion strategy comparison experiment is designed in the MOSI data set to verify the effectiveness of the HGFN fusion strategy. The improved JDSN component remains unchanged in the experiment, and the fusion component is replaced with Multi-Attention Fusion (AttFusion), vector concatenation fusion (Concate), dynamic fusion net (DFN), and other strategies. Then, the results are concluded, as shown in Tabel 4.

The results shown in Tabel 4 indicate that HGFN has a significantly improved performance compared with other fusion methods. The reason for these results is that HGFN not only models single-modal, bimodal, and trimodal layers dynamically but also obtains trimodal fusion representations more comprehensively by the splicing mode of various modal layers. Moreover, to verify the dynamicity of the graph fusion network, the weight change of the fusion process is visualized as follows.

As shown in Figure 4, the vertical axis represents the iteration order, and the horizontal axis represents the interaction information vector in the dynamic layer. The value in the figure represents the weight of the corresponding information vector. The results of vertical axis analysis indicate that the contributions of different discourse segments to the same modal interaction information vector are almost unchanged. The reason is that the modal data are affected by the similarity constraint in the domain separation representation learning prior to fusion, which reduces the fluctuation in the difference amongst all sample representations. Through the observation of the horizontal axis, for single-modal vector weight (the first three columns), the contributions of linguistic mode to the prediction result are the most evident. The reason is that language text is usually the most important information in MSA. For bimodal vector weight (fourth–sixth column), weight "tv" is closer to "ta" and significantly greater than weight "va". The reason may be that linguistic mode plays a more important role in bimodal fusion than other modes. Through observation of the trimodal vector weight (the seventh–twelfth column), the vector weight obtained by fusing one bimodal vector and one single-modal vector is close to 0. However, the vector weight obtained by fusing two bimodal vectors is dominant in the trimodal information. It indicates that modeling the interaction process of every two bimodal vectors is

Table 2: Comparison experiments of multimodal models in MOSI

| Model | MAE | Mul_Acc2 | Mul_Acc5 | Corr | F1_Score | CPU_Clock |
|---|---|---|---|---|---|---|
| TFN [11] | 1.016 | 0.765 | 0.386 | 0.604 | 0.765 | 0.404 |
| LMF [12] | 1.009 | 0.767 | 0.362 | 0.604 | 0.769 | 0.395 |
| MFN [9] | 1.007 | 0.773 | 0.329 | 0.632 | 0.773 | 0.379 |
| ARGF [15] | 0.857 | 0.814 | 0.423 | 0.712 | 0.815 | 0.147 |
| Gragh-MFN [10] | 1.003 | 0.784 | 0.360 | 0.623 | 0.785 | 0.454 |
| MARM [20] | 1.028 | 0.756 | 0.351 | 0.625 | 0.755 | 0.345 |
| LSTHM [20]-AttFusion | 1.087 | 0.745 | 0.375 | 0.608 | 0.744 | 1.527 |
| LSTHM [20]-Concat | 1.056 | 0.750 | 0.370 | 0.581 | 0.752 | 1.524 |
| LSTHM [20]-DFG | 0.992 | 0.758 | 0.401 | 0.626 | 0.757 | 0.357 |
| LSTHM [20]-Out_Product | 1.092 | 0.764 | 0.332 | 0.569 | 0.764 | 0.708 |
| MISA [41] | 0.827 | 0.819 | 0.440 | 0.726 | 0.819 | 0.839 |
| **DISRFN (ours)** | **0.798** | **0.834** | **0.468** | **0.734** | **0.836** | **0.737** |

Table 3: Comparison experiments of multimodal models in MOSEI.

| Model | MAE | Mul_Acc2 | Mul_Acc5 | Corr | F1_Score | CPU_Clock |
|---|---|---|---|---|---|---|
| TFN [11] | 0.714 | 0.760 | 0.443 | 0.507 | 0.761 | 0.417 |
| LMF [12] | 0.729 | 0.761 | 0.436 | 0.520 | 0.760 | 0.412 |
| MFN [9] | 0.715 | 0.773 | 0.432 | 0.530 | 0.772 | 0.418 |
| Gragh-MFN [10] | 0.714 | 0.765 | 0.448 | 0.526 | 0.766 | 0.46 |
| MARM [20] | 0.708 | 0.772 | 0.449 | 0.530 | 0.773 | 0.363 |
| LSTHM [20]-AttFusion | 0.852 | 0.733 | 0.383 | 0.403 | 0.733 | 1.585 |
| LSTHM [20]-Concat | 0.861 | 0.704 | 0.383 | 0.383 | 0.721 | 1.6 |
| LSTHM [20]-DFG | 0.837 | 0.748 | 0.391 | 0.437 | 0.748 | 0.369 |
| LSTHM [20]-Out_Product | 0.905 | 0.722 | 0.383 | 0.405 | 0.723 | 0.715 |
| MISA [41] | 0.600 | 0.858 | 0.538 | 0.776 | 0.857 | 0.975 |
| **DISRFN (ours)** | **0.591** | **0.875** | **0.541** | **0.781** | **0.875** | **0.948** |

Table 4: Experiments of fusion methods.

| Method | MAE (↓) | Mul_Acc2 (↑) | Mul_Acc5 (↑) | Corr (↑) | F1_Score (p/g) (↑) |
|---|---|---|---|---|---|
| JDSN-AttFusion | 0.924 | 0.791 | 0.378 | 0.687 | 0.782 |
| JDSN-concat | 0.839 | 0.814 | 0.443 | 0.724 | 0.813 |
| JDSN-DFG | 0.825 | 0.816 | 0.459 | 0.727 | 0.817 |
| **DISRFN (ours)** | **0.798** | **0.834** | **0.468** | **0.734** | **0.836** |

necessary. And it is also verified that the fusion network can dynamically fuse the multimodal data.

*5.3. Ablation Study.* The loss functions of various components discussed in Section 3.3 play an important role in the implementation of an improved joint domain separation network in Section 3.2. Therefore, the loss function is analyzed and discussed, and visualised and quantitative analysis is conducted based on ablation study.

*5.3.1. Visual Presentation.* An ablation experiment is designed in this section. The network is retrained after obtaining a zero setting of the loss weights $(\alpha, \beta, \lambda, \eta)$ of other components except for the basic task loss $L_{task}$, and the best performance model parameters are saved. Moreover, to intuitively observe the effects of various loss functions on the model results, the fusion representation of MOSI test samples is visualized by T-SNE, as shown in Figure 5.

As shown in Figure 5, the red spots represent positive emotions, and the blue ones represent negative emotions. When the distance between spots of the same color is shorter and the distance between spots of different colors is farther, the effect of semantic clustering and emotion analysis is better. The figure shows the T-SNE graph of the test data fusion representation, showing different distribution features under different loss function training. When all component losses exist, the model has the best semantic clustering effect. When the weight $\gamma$ of the reconstruction loss $L_{recon}$ is zero, it has the suboptimal clustering effect. When similarity loss $L_{sim}$ does not exist, the clustering effect of the model is the most divergent. The impact of the loss $L_{diff}$ and $L_{trip}$ is between similarity loss and reconstruction loss. Furthermore, to explore the effect of each loss more

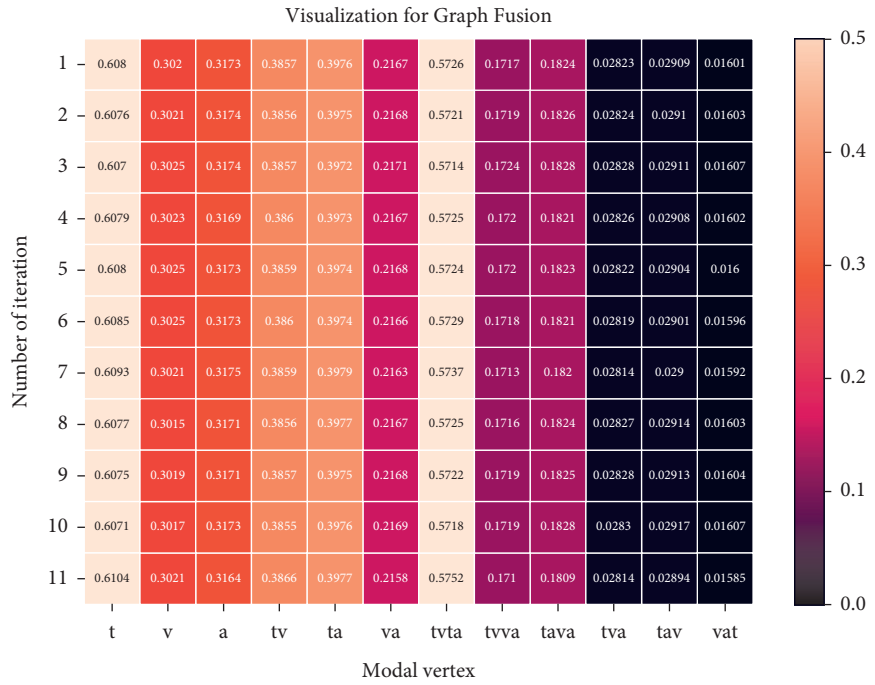


Figure 4: Visualization for Graph Fusion in MOSI sentiment analysis task.

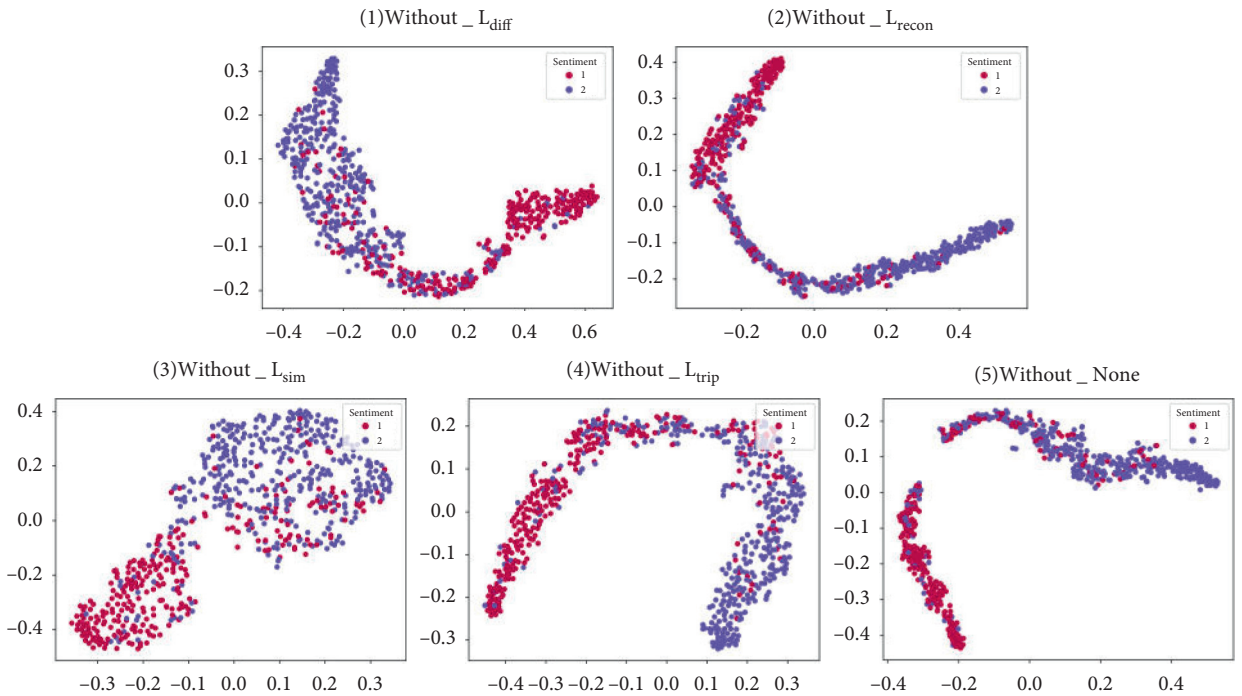


Figure 5: Visualization of sentiment semantic distribution under different loss. Notes: (1) lack of loss function Ldiff; (2) lack of loss function Lrecon; (3) lack of loss function Lsim; (4) lack of loss function Ltrip; (5) full configuration of loss function.

specifically, the evaluation indexes of the best model of each experiment are recorded in Table 5 for quantitative analysis.

*5.3.2. Quantitative Analysis.* As shown in Table 5, the model achieves the best performance when all losses are involved. This finding indicates that each component loss is effective. The observation results show that the model is sensitive to $L_{sim}$ and $L_{diff}$. It means that decomposing modes into independent space is conducive to the performance improvement of the model. The effect of cosine triplet-margin loss on the model is smaller than $L_{sim}$ and $L_{diff}$. Because

Table 5: Experiments of ablation study.

| Method | MAE | Mul_Acc2 | Mul_Acc5 | Corr | F1_Score |
|---|---|---|---|---|---|
| Without diff loss | 0.868 | 0.811 | 0.404 | 0.728 | 0.816 |
| Without sim loss | 0.999 | 0.784 | 0.351 | 0.723 | 0.782 |
| Without recon loss | 0.833 | 0.817 | 0.464 | 0.711 | 0.816 |
| Without CosineTriplet loss | 0.857 | 0.799 | 0.469 | 0.705 | 0.798 |
| **ALL loss** | **0.798** | **0.834** | **0.468** | **0.734** | **0.836** |

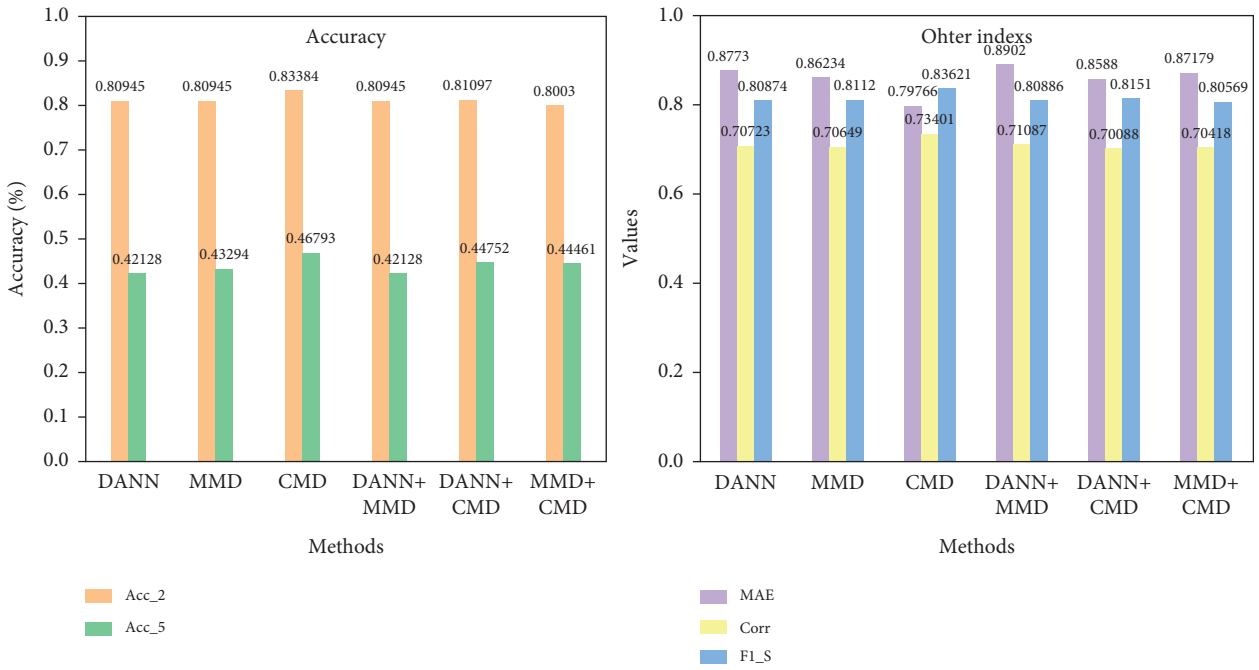


Figure 6: Visualization of performance comparison under different similarity loss.

semantic clustering effect is observed in the process of modal similarity feature acquisition. Therefore, the effect of this loss is weakened. In addition, the model is less dependent on reconstruction loss. The reason is that the trivial representation features of a specific encoder can be learned by $L_{task}$ in the absence of reconstruction loss. The model is most sensitive to similarity loss; thus, the selection of similarity loss is very important. Therefore, an in-depth analysis is discussed in the following section.

*5.4. Comparison of Similarity Measures.* In this section, the selection of similarity loss function in 3.4.2 is discussed. For this reason, the following experiment is designed. Domain adversarial loss (DANN) [48], maximum mean square measure (MMD) [49], CMD, and their combinations are used for network training tests, as shown in Figure 6. The first three columns in the figure show that the performance of CMD in a single form is better than that of MMD and DANN in various indexes.

The reasons are summarised in the following points: (i) CMD can directly perform exact matching of the high-order moment without expensive distance and kernel matrix calculation; (ii) compared with CMD, DANN obtains modal similarity through minimax game using discriminator and shared encoder. However, in adversarial training, additional parameters are added, and fluctuations may be encountered in training. Moreover, through the observation of joint form (the last three columns), the effect of similarity loss with CMD is better than that of the loss without CMD but worse than that of single CMD loss. This finding indicates that the increase in computation cost reduces the efficiency of network learning and further verifies the rationality of selecting CMD as similarity loss.

## 6. Conclusions

This paper studies multimodal emotion analysis. In the research, we have the following findings: (1) feature representation with more comprehensive information can reduce the burden of fusion network; (2) the redundant information of each mode can be used more effectively by jointing modality-invariance and modality-specificity representations of each mode; (3) simple dynamic fusion mechanism can obtain the interaction between modes more efficiently. Thus, this study puts forward a multimodal sentiment analysis framework consisting of two parts, namely, improved JDSN and HGFN. Firstly, modal invariant-specific joint representation of each mode is obtained through an improved JDSN module to effectively

utilize the complementary information amongst modes and reduce the heterogeneity gap between modes. Then, the joint representation of each mode is input to the HGFN for fusion to provide input for the prediction network. Moreover, a new combined loss function is designed to encourage the DISRFN model to learn the representation of expectation. Finally, the performance analysis experiment is carried out on MOSI and MOSEI data sets, obtaining acceptable results. In practice, the multimodal data usually have an unbalanced phenomenon, which will lead to the task bottleneck of the model. However, the study does not consider this issue. Therefore, we plan to study the problems of multimodal imbalance in the future.

## Data Availability

The data used includes MOSI and MOSEI. The address of the MOSI dataset is correct. The MOSEI dataset address is as follows: http://immortal.multicomp.cs.cmu.edu/raw_datasets/CMU_MOSEI.zip.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Authors' Contributions

Changfan Zhang and Haonan Yang conceived and designed the experiments; Haonan Yang proposed the method; Jing He performed the experiments; Yifu Xu analyzed the data; Hongrun Chen prepared the original draft.

## Acknowledgments

This work was supported by the Natural Science Foundation of China (U1934219, 52172403, and 62173137), Hunan Provincial Natural Science Foundation of China (2021JJ50001and 2021JJ30217), and Project of Hunan Provincial Department of Education (19A137).

## References

[1] Q. Li, D. Gkoumas, C. Lioma, and M. Melucci, "Quantum-inspired multimodal fusion for video sentiment analysis," *Information Fusion*, vol. 65, pp. 58–71, 2021.

[2] K. Huang, W. Zhou, and M. Fang, "Deep multimodal fusion autoencoder for saliency prediction of RGB-D images," *Computational Intelligence and Neuroscience*, vol. 2021, pp. 1–10, 2021.

[3] J. Zhou, M. Ye, J. Ding, S. Mao, and H. J. Zhang, "Rapid and robust traffic accident detection based on orientation map," *Optical Engineering*, vol. 51, no. 11, Article ID 117201, 2012.

[4] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429–439, 2020.

[5] S. Mao, M. Ye, X. Li, F. Pang, and J. Zhou, "Rapid vehicle logo region detection based on information theory," *Computers & Electrical Engineering*, vol. 39, no. 3, pp. 863–872, 2013.

[6] S. Mao, H. Wu, and M. Lu, "Multiple 3D marker localization and tracking system in image-guided radiotherapy," *International Journal of Robotics and Automation*, vol. 32, no. 5, pp. 517–523, 2017.

[7] Y. Zhang, D. Song, X. Li et al., "A Quantum-Like multimodal network framework for modeling interaction dynamics in multiparty conversational sentiment analysis," *Information Fusion*, vol. 62, pp. 14–31, 2020.

[8] S. Agethen and W. H. Hsu, "Deep multi-kernel convolutional LSTM networks and an attention-based mechanism for videos," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 819–829, 2020.

[9] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multiview sequential learning," in *proceedings of the thirty-second AAAI conference on artificial intelligence (AAAI-2018)*, vol. 32, no. 1, pp. 5634–5641, New Orleans, Louisiana, USA, 2018.

[10] A. Zadeh, P. P. Liang, J. Vanbriesen, S. Poria, E. Cambria, and L. 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 (ACL-2018)*, pp. 2236–2246, Melbourne, Australia, 2018.

[11] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 1, pp. 1103–1114, Copenhagen, Denmark, 2017.

[12] 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," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 2247–2256, Melbourne, Australia, 2018.

[13] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proceedings of the 30th conference on neural information processing systems (NIPS-2016)*, vol. 3, pp. 343–351, Barcelona, Spain, 2016.

[14] R. F. Silva, S. M. Plis, T. Adali, M. S. Pattichis, and V. D. Calhoun, "Multidataset independent subspace analysis with application to multimodal fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 588–602, 2021.

[15] S. Mai, H. Hu, and S. Xing, "Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, pp. 164–172, New York, NY, USA, 2020.

[16] J. Kim, T. Kim, S. Kim, and C. Yoo, "Edge-labeling graph neural network for few-shot learning," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[17] S. J. Mai, H. F. Hu, and S. L. X.. Divide, "Conquer and combine: hierarchical feature fusion network with local and global perspectives for multimodal affective computing," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 481–492, Florence Italy, 2019.

[18] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 122–137, 2020.

[19] P. P. Liang, Z. Liu, A. Zadeh, and L. P. Morency, "multimodal language analysis with recurrent multistage fusion," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 1, pp. 150–161, Brussels, Belgium, 2018.

[20] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pp. 5642–5649, New Orleans, Louisiana, USA, 2018.

[21] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," in *Proceedings of the advances in neural information processing systems*, vol. 32, pp. 1–10, Vancouver, Canada, 2019.

[22] S. Mai, H. Hu, J. Xu, and S. Xing, "Multi-fusion residual memory network for multimodal human sentiment comprehension," *IEEE Transactions On Affective Computing*, p. 1, 2020.

[23] Y.-J. Zhang and Z.-H. Ling, "Extracting and predicting word-level style variations for speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1582–1593, 2021.

[24] 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, pp. 7216–7223, Honolulu, Hawaii, 2019.

[25] M. H. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI-17)*, pp. 163–171, Glasgow, Scotland, 2017.

[26] X. Shu and G. Zhao, "Scalable multilabel canonical correlation analysis for cross-modal retrieval," *Pattern Recognition*, vol. 115, Article ID 107905, 2021.

[27] Y. Kaloga, P. Borgnat, S. P. Chepuri, P. Abry, and A. Habrard, "Variational graph autoencoders for multiview canonical correlation analysis," *Signal Processing*, vol. 104, Article ID 108182, 2021.

[28] S. Verma, J. W. Wang, Z. F. Ge et al., "Deep-HOSeq: Deep Higher Order Sequence Fusion for Multimodal Sentiment Analysis," in *Proceedings of the 2020 IEEE International Conference on Data Mining (ICDM)*, Sorrento, Italy, 2020.

[29] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: learning robust joint representations by cyclic translations between modalities," in *Proceedings of the thirty-third AAAI conference on artificial intelligence (AAAI-19)*, vol. 33, no. 01, pp. 6892–6899, Honolulu, Hawaii, USA, 2019.

[30] H. Qiang, Y. Wan, L. Xiang, and X. Meng, "Deep semantic similarity adversarial hashing for cross-modal retrieval," *Neurocomputing*, vol. 400, pp. 24–33, 2020.

[31] F. Wu, X.-Y. Jing, Z. Wu et al., "Modality-specific and shared generative adversarial network for cross-modal retrieval," *Pattern Recognition*, vol. 104, Article ID 107335, 2020.

[32] Y. H. H. Tsai, P. P. Liang, A. Zadeh, L. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *Proceedings of the International Conference on Learning Representations (ICLR-2019)*, New Orleans, Louisiana, USA, 2019.

[33] P. P. Liang, Y. C. Lim, Y. H. Tsai, R. Salakhutdinov, and L. Morency, "Strong and simple baselines for multimodal utterance embeddings," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, vol. 1, pp. 2599–2609, Minneapolis, Minnesota, 2019.

[34] D. Wang, Q. Wang, L. He, X. Gao, and Y. Tian, "Joint and individual matrix factorization hashing for large-scale cross-modal retrieval," *Pattern Recognition*, vol. 107, Article ID 107479, 2020.

[35] Y. Fang, Y. Ren, and J. H. Park, "Semantic-enhanced discrete matrix factorization hashing for heterogeneous modal matching," *Knowledge-Based Systems*, vol. 192, Article ID 105381, 2020.

[36] J. C. Caicedo, J. BenAbdallah, F. A. González, and O. Nasraoui, "Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization," *Neurocomputing*, vol. 76, no. 1, pp. 50–60, 2012.

[37] Y. Wu, Y. Zhao, X. Lu et al., "Modeling incongruity between modalities for multimodal sarcasm detection," *IEEE Multimedia*, vol. 28, pp. 86–95, 2021.

[38] F. Chen, Z. Luo, and Y. Xu, "Complementary Fusion of Multi-Features and Multi-Modalities in Sentiment Analysis," in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, New York, NY, USA, 2020.

[39] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis the thirty-fourth AAAI conference on artificial intelligence (AAAI-20)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, pp. 8992–8999, New York, NY, USA, 2020.

[40] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: a survey," *IEEE Access*, vol. 7, Article ID 63373, 2019.

[41] 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 (ACM MM)*, pp. 1122–1131, Seattle, USA, 2020.

[42] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba, "Cross-modal scene networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2303–2314, 2018.

[43] W. Zellinge, T. Grubinger, E. Lughofer, T. Natschlager, and S. S. Platz, "CMD for Domain-Invariant Representation learning," in *Proceedings of the 5th International Conference on Learning Representations (ICLR-2017)*, Toulon, France, 2017.

[44] W. Gu, X. Y. Gu, J. Z. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval (ICMR-2019)*, pp. 159–167, Ottawa, ON, Canada, 2019.

[45] J. He, S. Mai, and H. Hu, "A unimodal reinforced transformer with time squeeze fusion for multimodal sentiment analysis," *IEEE Signal Processing Letters*, vol. 28, pp. 992–996, 2021.

[46] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP-A collaborative voice analysis repository for speech technologies," in *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 960–964, Florence, Italy, 2014.

[47] R. Aharoni and Y. Goldberg, "Unsupervised domain clusters in pretrained language models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020)*, vol. 1, pp. 7747–7763, Seattle, Washington, USA, 2020.

[48] H. Tang and K. Jia, "Vicinal and categorical domain adaptation," *Pattern Recognition*, vol. 115, Article ID 107907, 2021.
[49] J. Pomponi, S. Scardapane, and A. Uncini, "Bayesian neural networks with maximum mean discrepancy regularization," *Neurocomputing*, vol. 453, pp. 428–437, 2021.