A Multi-Focus-Driven Multi-Branch Network for Robust Multimodal Sentiment Analysis

Chuanqi Tao*, Jiaming Li*, Tianzi Zang, Peng Gao

College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
{taochuanqi, ljiaming, zangtianzi, gaopengnj}@nuaa.edu.cn

Abstract

Multimodal sentiment analysis aims to integrate diverse modalities for precise emotional interpretation. However, external factors such as sensor malfunctions or network issues may disrupt certain modalities. This may lead to missing data, which poses challenges in real-world deployment. Most existing approaches focus on designing feature reconstruction strategies, overlooking the collaborative integration of reconstruction and fusion strategies. Moreover, they fail to capture the relationships between features in the global dimension and those in the local dimension. These limitations hinder the full capture of the complex nature of multimodal data, especially in scenarios involving missing modalities. To address the above issues, this paper proposes a robust model named MFMB-Net with multiple branches for feature multi-focus fusion and reconstruction. We design a two-stream fusion branch where macro-fusion focuses on the fusion of features in the global dimension and micro-fusion targets local dimension features. This dual-stream fusion branch distributes multi-focus across both pathways, simultaneously capturing global coarse-grained and local fine-grained features. Additionally, the reconstruction branch interacts collaboratively with the fusion branch to reconstruct and enhance the missing data. It integrates the reconstructed feature information with the fused information thus refining the representation fidelity of the missing information. Experiments performed on two benchmarks show that our approach obtains results superior to state-of-the-art models.

Code — https://github.com/MFMB-Net/MFMB-Net

Introduction

Recently, with the increased use of social media and the improvement of video quality, multimodal sentiment analysis (MSA) has become a significant domain in the field of sentiment analysis and has attracted widespread attention (Morency, Mihalcea, and Doshi 2011). MSA broadens traditional text-based sentiment analysis by adding audio and visual modalities, enabling the detection of subtle cues missed in text-only analysis. This comprehensive approach provides insights into online behaviors and emotional expressions.

[1]Corresponding authors.

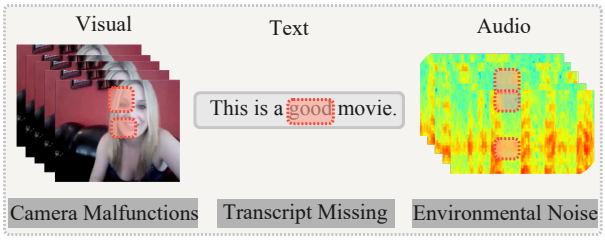

Figure 1: Scenarios that may result in modality missing.

The core of MSA lies in multimodal representation learning and multimodal fusion. These components are crucial to MSA's effectiveness in modeling and analyzing emotional expressions. The former is the process of extracting features from different sources and transforming these features into representations that can be processed by the model. The latter combines representations learned from different modalities to form a unified, integrated representation that more comprehensively reflects the complexity of emotions. And so far, many models have been proposed to address multimodal representation learning (Hazarika, Zimmermann, and Poria 2020; Yu et al. 2021; Lin and Hu 2022; Zhang et al. 2023; Oord, Li, and Vinyals 2018) and multimodal fusion (Han et al. 2021; Lv et al. 2021; Nagrani et al. 2021; Cheng et al. 2021) through the tireless efforts of researchers.

The basic premise of the above is that the multimodal data is complete. However, in real-world deployment, we often encounter situations where modalities are missing. As shown in Figure 1, for instance, camera malfunctions may result in occluded visual features and extensive environmental noise can render audio information unusable. These can affect model performance, therefore, addressing missing modalities is indispensable in practical applications.Recently, several translation-based (Pham et al. 2019; Tang et al. 2021) and generative model-based (Zhao, Li, and Jin 2021; Zeng, Zhou, and Liu 2022) approaches have been proposed to address MSA in missing modalities to retain the maximum information from all modalities. While these approaches are attractive, the former is developed to deal with the total missing of one or more modalities, which is unlikely to happen in the real world. And thus we mainly

address the case of random missing modalities. The latter approaches predominantly concentrate on the singular aspect of reconstructing missing features, neglecting the importance of collaboratively well-designed fusion strategies in missing modalities. Overall, on the one hand, most of existing approaches focus primarily on designing feature reconstruction strategies independently, neglecting the crucial issue of synergistically integrating reconstruction and fusion strategies. On the other hand, existing research on missing modalities usually adopts simple and identical symmetrical parallel processing structures, such as three identical cross-modal attention layers for interaction. It ignores the special processing requirements of different information densities of different modalities and fails to capture the relationship between global and local dimensional features. Specifically, there are natural differences in the information density of the three modalities: text modality typically contains more useful core information, while the audio and video modalities contain complementary information. Furthermore, text modalities are often processed by powerful feature extractors like BERT (Devlin et al. 2018), resulting in more informative content. In contrast, features extracted from audio and visual modalities by less sophisticated extractors such as COVAREP (Degottex et al. 2014) and Facet[1] may contain partial redundancy and noise. We will verify this in subsequent experiments.These limitations hinder fully capturing the complexity of multimodal data with missing modalities.

To address the above problems, we propose a robust multi-focus multi-branch fusion model (MFMB-Net). It contains multiple branches, and the fusion branch collaborates with the reconstruction branch to reconstruct and enhance the missing modalities, thus improving the fidelity of the missing information representation. Further, the well-designed dual-stream fusion branch assigns multiple focus to capture features at different granularities. The macro-fusion stream contains a multi-focus module that utilizes a coarse hub to assign focus to critical information in the global dimension, while the micro-fusion stream is used to integrate critical information in the local dimension. The novel contributions of our work can be summarized as follows:

- We propose a robust model for multi-focus and multi-branch fusion, which comprehensively considers the integration and interaction of features in the global dimension and features in the local dimension under missing data conditions. Meanwhile, our model utilizes multiple branch streams to coordinate the fusion and reconstruction process.
- We design a two-stream fusion branch where macro-fusion focuses on the fusion of features in the global dimension and micro-fusion targets local dimension features. This fusion strategy disperses multi-focuses on global coarse-grained features and local fine-grained features.
- We conduct extensive experiments on the CMU-MOSI and CMU-MOSEI datasets under both complete and missing modalities conditions, and we gain superior results to the state-of-the-art models.

[1]https://imotions.com/

Related Work

Multimodal Sentiment Analysis

Multimodal sentiment analysis integrates information from various modalities, such as text, audio, and visual cues, to accurately assess and interpret emotions (Morency, Mihalcea, and Doshi 2011). Multimodal fusion is a central aspect of the MSA framework, which aims to integrate representations from different modalities into a unified and comprehensive representation. It is common to use Transformer (Vaswani et al. 2017) to fuse multimodal representations. Zadeh et al. (Zadeh et al. 2018a) used LSTM to model each modality in the temporal dimension. Zadeh et al. (Zadeh et al. 2017) proposed a Tensor Fusion Network to explicitly capture unimodal, bimodal, and trimodal interactions through a 3-fold Cartesian product from modality embedding. Tsai et al. (Tsai et al. 2019) used cross-modal attention mechanisms to effectively capture complex bimodal interactions. Hazarika et al. (Hazarika, Zimmermann, and Poria 2020) used mode-invariant and mode-specific representations by projecting each modality into two different subspaces to reduce the gap between modalities and capture modality. Han et al. (Han et al. 2021) designed an innovative end-to-end bimodal fusion network that conducts fusion and separation on pairs of modality representations. Paraskevopoulos et al. (Paraskevopoulos, Georgiou, and Potamianos 2022) proposed a feedback module named MMLatch that allows modeling top-down cross-modal interactions between higher and lower-level architectures. Nagrani et al. (Nagrani et al. 2021) proposed MBT which is a new transformer architecture for audio-visual fusion that restricts the cross-modal attention flow to the latter layers of the network. Zhang et al. (Zhang et al. 2023) proposed ALMT with an Adaptive Hyper-modality Learning (AHL) component, which leverages language features across scales to filter irrelevant and conflicting information from visual and audio inputs.

Missing Modality Problem in MSA

The extensive research on complete modalities has reached a significant depth of understanding. However, in real-world deployments or under specific conditions, data may suffer from absence or corruption. Prior works in this domain are primarily categorized into two main strategies: 1) Generative methods, which are designed to produce new data that conforms to the observed distribution, including techniques such as Variational Auto-encoders (VAEs) and Generative Adversarial Networks (GANs). 2) Joint learning approaches, which strive to extract latent representations from the available data, utilizing strategies such as translation-based and cycle-consistency-based learning methods. Pham et al. (Pham et al. 2019) proposed MCTN to learn robust joint representations through inter-modality information translation, believing that source-to-target transitions effectively capture shared modal information. Tang et al. (Tang et al. 2021) proposed CTFN to achieve robustness

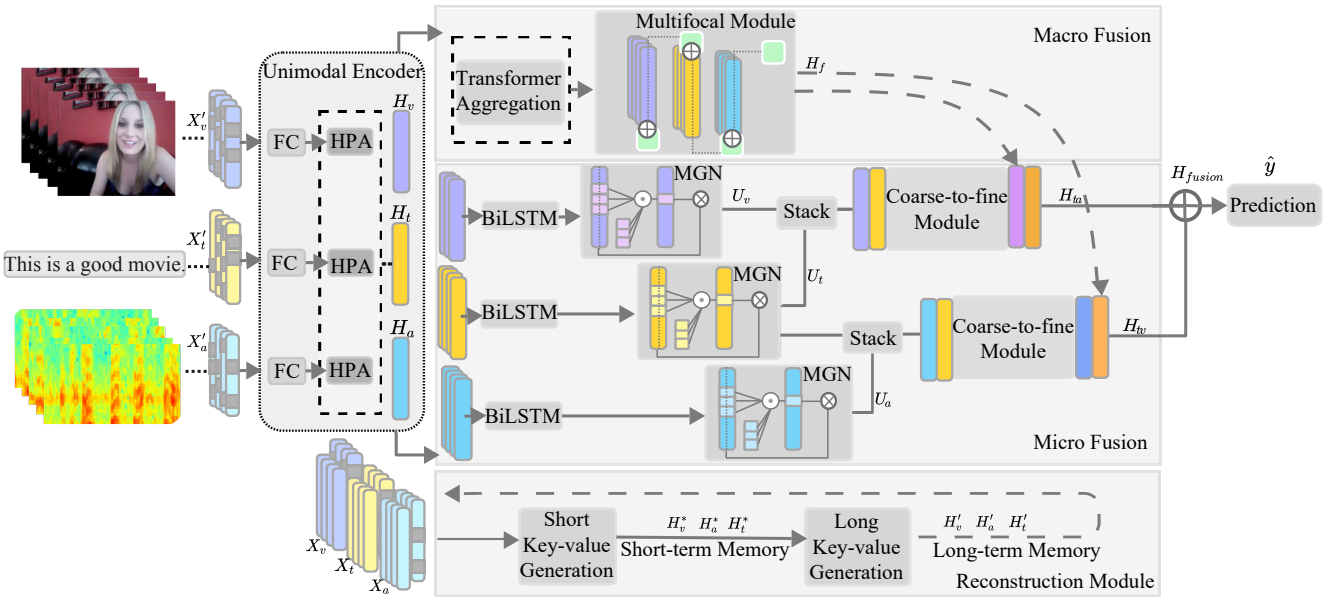

Figure 2: The framework of MFMB-Net contains unimodal encoder, two-stream fusion, and reconstruction module.

against missing data through coupled translation fusion and cyclic consistency constraints. Zhao et al. (Zhao, Li, and Jin 2021) introduced MMIN, which also introduces cycle consistency learning for imputing missing modalities and devises a CRA-based module for cross-modality imagination using paired data, further proposing a network to address uncertain missing cases. Zeng et al. (Zeng, Zhou, and Liu 2022) proposed EMMR to employ an encoder-decoder structure to identify and restore essential semantics from the absent modality, effectively tackling MSA's inconsistency and prioritizing the influential modality for accurate sentiment analysis. Yuan et al. (Yuan et al. 2021) introduced TFR-Net to strengthen model robustness against random missing data in nonaligned modality sequences by using a transformer-based framework for relation extraction and a feature reconstruction network to recover missing semantics. Zeng et al. (Zeng, Liu, and Zhou 2022) proposed TATE, which incorporates an innovative tag encoding module designed to handle various missing modalities. Sun et al. (Sun et al. 2023) used EMT-DLFR to improve model performance on incomplete data by implicitly extracting semantics and explicitly aligning high-level representations between complete and incomplete datasets using siamese representation learning.

Methodology

As an overview of the model, Figure 2 illustrates the architecture of MFMB-Net.

Problem Definition

Complete modality data consists of unimodal raw sequences $X_m \in R^{l_m \times d_m}$ from the same video clip, where l_m is the sequence length and d_m is the dimension of the representation vector of modality $m \in \{t, v, a\}$, respectively. In this paper, we tackle the missing modality issue. The input data to the model is incomplete modality data with randomly missing from modality $m \in \{t, v, a\}$, denoted as $X'_m \in R^{l_m \times d_m}$. The goal of the task is to learn a mapping $f\left(X'_t, X'_v, X'_a\right)$ using incomplete data to infer the sentiment score $\hat{y} \in R$. Moreover, during the training phase, leveraging the full data features along with the locations of missing features helps to guide the learning process of representations.

Feature Extraction

To ensure consistent and accurate performance comparisons across different approaches, we utilize the features commonly adopted by most MSA methods.

Text: We employ the bert-base-uncased model to extract feature representations of text. It converts text into vectors with 768-dimensions.

Audio: We apply COVAREP to to extract features from the CMU-MOSI and CMU-MOSEI datasets, resulting in 5- and 74- dimensions, respectively.

Visual: We extract 20-dimensional features from the CMU-MOSI dataset and 35-dimensional features from the CMU-MOSEI dataset using Facet.

Unimodal Encoders

We first encode the input X_m into the length representations as H_m. Specifically, features extracted from the feature extraction module are individually fed into fully connected networks. It can distill high-dimensional features into a more concise and abstract form, and it incorporates non-linearity to capture complex patterns.

$$h_m = W_m X'_m + b_m, \quad m \in \{t, v, a\} \tag{1}$$

where W_m, b_m are the parameters of the linear layer.

Alignment

Pre-fusion alignment allows better utilization of complementary multimodal information, inspired by (Han et al. 2021), we use self-attention as the soft alignment strategy to align the features before fusion. The shallow representations h_m are fed into the hybrid parallel attention module (HPA), which first uses self-attention to align features and learn the dynamics within each modality, and then uses cross-attention to learn the dynamics between modalities.

$$H_m = HPA(h_m, h_m, h_m), \quad m \in \{t, v, a\} \tag{2}$$

Fusion

Our fusion module is designed with two distinct parallel streams that cater to the unique processing demands of three different modalities.

Macro-fusion

The first stream allocates focus to global information and is equipped with an efficient multi-focal module. It processes all modalities in a waterfall flow to integrate the information, as shown in the Macro Fusion in Figure 2.

Firstly, we feed all modalities into the Transformer encoder separately to process and capture the complex dependencies between elements in the input sequences.

$$\overline{H}_m = Transformer(H_m, H_m, H_m), \quad m \in \{t, v, a\} \tag{3}$$

Secondly, mapping high-dimensional data to a compact latent space helps the network focus on key information and filter out noise, improving accuracy and efficiency (Oord, Li, and Vinyals 2018). We extend the progressive fusion strategy in (Nagrani et al. 2021) to integrate visual, audio, and text modalities in a multi-stage manner, utilizing a coarse hub to guide the fusion and only extract fused features compressed into a compact space.

Specifically, we introduce a coarse hub H_{ker} of length k, which is sequentially concatenated with visual, audio, and text features to transport integrated key information. The visual features are first compressed and processed to extract the key information and imported into the hub. Subsequently, this visual key information interacts with the key information of the audio features, which is imported into the hub through further compression and fusion processes to realize the integration of the information of both. Finally, this fused information is communicated with the key information of the text features to complete the final information. This structure progressively condenses high-dimensional data into a compact latent space through a cascading, phased approach. The coarse hub gradually absorbs coarse-grained information, guiding the integration and interaction of multimodal information within the latent space.

The coarse hub flows between three modalities to compress the high-dimensional semantics and guide the fusion process as shown below:

$$[H'_v|H'_{ker}] = Transformer\left(\left[\overline{H}_v|H_{ker}\right]; \theta_v\right) \tag{4}$$

$$[H'_a|H''_{ker}] = Transformer\left(\left[\overline{H}_a|H'_{ker}\right]; \theta_a\right) \tag{5}$$

$$[H'_t|H'''_{ker}] = Transformer\left(\left[\overline{H}_t|H''_{ker}\right]; \theta_t\right) \tag{6}$$

where θ_v,θ_a,θ_t are the parameters of the model corresponding to visual, audio, and text modality respectively. H'''_{ker} is the coarse hub with coarse-grained information. Finally, we obtain the representations of the macro-fusion stream:

$$H_f = H'''_{ker} \tag{7}$$

Micro-fusion

The second stream assigns focus to local information. We design a pseudo-Siamese network with a memory-gating network and a coarse-to-fine fusion module to gradually integrate fine-grained information. As shown in the Micro Fusion in Figure 2.

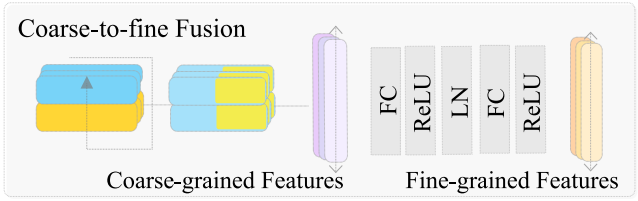

Figure 3: Coarse-to-fine Fusion Module.

Firstly, features are fed into a BiLSTM layer followed by a tanh activation function, which captures the contextual information.

$$\overline{H'}_m = tanh(BiLSTM(\overline{H'}_m)), \quad m \in \{t, v, a\} \tag{8}$$

Secondly, a one-dimensional convolutional network is used to focus on the local receptive field to ensure that each element in the input sequence can recognize its adjacent elements. Inspired by (Yuan et al. 2021), a feed-forward network acting as a memory-gated network (MGN) is then employed. MGN is used to filter out irrelevant contextual information and initially generate local coarse-grained features.

$$MGN = FFN(Conv1d(\overline{H'}_m)) \tag{9}$$

$$U_m = (\overline{H'}_m) \times MGN, \quad m \in \{t, v, a\} \tag{10}$$

where $\times$ means element-wise product. FFN is a feed-forward network with two linear transformations and a ReLU activation function.

Recently, many MLP-based models like MLP-Mixer (Tolstikhin et al. 2021), HireMLP (Guo et al. 2022), Cyclemlp (Chen et al. 2023) have been proposed. This is a competitive but conceptually and technically simple alternative to MLP alone to achieve performance comparable to CNN, Transformer. Finally, we design a coarse-to-fine module, which based on a simple and effective MLP. Thanks to the excellent performance of the powerful encoder and the rich information of the modality itself, we use the text modality as a flag to guide the visual and audio modalities through local fusion. As shown in Figure 3. The text modality is stacked with the visual modality and audio modality along the spatial dimension respectively to obtain stacked features.

$$H_{stack_{tq}} = stack(U_t, U_q), \quad q \in \{v, a\} \tag{11}$$

Next, the features are rearranged and normalized to achieve coarse interactions in the spatial dimension. We get the further processed coarse-grained fusion features H_{co_k}.

$$H_{co_k} = BN(\phi(H_{stack_k})), \quad k \in \{tv, ta\} \tag{12}$$

where ϕ is rearrange function. BN is the BatchNorm operation.

Then, the integrated features are stretched to the temporal dimension and input into coarse-to-fine MLP layers. By leveraging the aforementioned nonlinear and hierarchical structure for effective learning, it is possible to achieve intricate fine-grained fusion interactions along the temporal dimension. In this way, we obtain text-guided local fine-grained features.

$$H_k = ReLU(W_2(LN(ReLU(W_1 H_{co_k} + b_1))) + b_2) \quad (13)$$

where LN is LayerNorm operation. W_1, b_1, W_2, b_2 are the parameters of the linear layers. $k \in \{tv, ta\}$.

Finally, the local fine-grained features are concatenated to form the final representation H_{fusion} of the micro-fusion.

Reconstruction

Inspired by (Geva et al. 2021), the feed-forward network (FFN) can act as a key-value memory network. We employ self-attention to extract information, followed by a feed-forward network consisting of two fully connected layers with a Rectified Linear Unit (ReLU). The former integrates high-density information from the short-term contextual memory, and the latter taps into the long-term memory and compresses the feature dimensions into the output space. The front and back work synergistically to enhance each other, providing support on the reconstruction branch to restore high-fidelity features H'_m.

$$H^*_m = \text{softmax}\left(\frac{(W^Q H_m)(W^K H_m)^T}{\sqrt{d_m}}\right)(W^V H_m) \quad (14)$$

$$H'_m = FFN(H^*_m), \quad m \in \{t, v, a\} \quad (15)$$

where W^Q, W^K, and W^V are corresponding mapping matrices that project query, key and value vectors to different low-dimensional spaces.

By systematically adjusting the gradients of network parameters for local short-term and global long-term memory dimensions along the backpropagation trajectory, the model effectively inhibits overfitting due to task-specific loss. And it can capture the latent patterns in both missing and complete data.

Prediction

Finally, the coarse global fusion features and fine local fusion features obtained from each stream are concatenated to form the final representation $H = H_f + H_{\text{fusion}}$. Then we feed H into a layer of a fully connected network to obtain the final sentiment prediction $\hat{y}$.

$$\hat{y} = W_H H + b_H \quad (16)$$

where W_H is the weight vectors, b_H is the bias.

Objective Function

Our objective function comprises both task loss and generative loss. In our experiments, to comprehensively evaluate model performance, we consider two tasks: a regression task and a classification task, with losses computed separately as:

$$L_{\text{task}} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| & \text{For regression} \\ -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) & \text{For classification} \end{cases} \quad (17)$$

where N is the size of a mini-batch, y_i and $\hat{y}_i$ represent the true label and the predicted label of the sample, respectively.

The generative loss between the original and the reconstructed representation under missing conditions is as follows:

$$L^m_{\text{gen}} = \text{Smooth}_{L1}(H'_m \otimes \text{Mask}, X_m \otimes \text{Mask}), m \in \{t, v, a\} \quad (18)$$

where $Mask$ locates missing positions in input sequences.

The final objective function:

$$L = L_{\text{task}} + \sum_{m \in \{t,v,a\}} \alpha_m L^m_{\text{gen}} \quad (19)$$

where α_m are the weights that balance the contribution of different modalities.

Experiments

Datasets

We conduct experiments on public multimodal sentiment analysis benchmark datasets. The basic statistics of each dataset are shown in Table 1.

CMU-MOSI (Zadeh et al. 2016) contains 2199 short monologue video clips taken from 93 YouTube videos. The utterances are manually annotated with a sentiment score from -3 (strongly negative) to 3 (strongly positive).

CMU-MOSEI (Zadeh et al. 2018b) contains 22856 annotated video segments from 1000 distinct speakers and 250 topics gathered from online websites. Each utterance is annotated with a sentiment intensity from [-3, 3].

Datasets	Train	Val	Test
MOSI	552/53/679	92/13/124	379/30/277
MOSEI	4738/3540/8048	506/433/932	1350/1025/2284

Table 1: Statistics from the two MSA benchmark datasets. Each dataset is divided into sample sizes for different emotional tendencies, which are negative, neutral, and positive.

Model	Acc-2	Acc-7	F1
FC(V)	44.79/42.3	15.48	61.45/59.03
FC(A)	44.94/42.43	15.43	61.54/59.08
FC(T)	66.97/67.18	28.92	69.34/69.49
Concat(V,A,T)	68.15/68.33	29.07	72.60/72.78

Table 2: Unimodal experiments on CMU-MOSI dataset. FC denotes a fully connected network of unimodal data, and Concat denotes a fully connected network of three modalities connected after simple concatenation.

Model	CMU-MOSI						CMU-MOSEI					
	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1
TFN[1]	0.901	0.698	34.9	-	-/80.8	-/80.7	0.593	0.700	50.2	-	-/82.5	-/82.1
LMF[1]	0.917	0.695	33.2	-	-/82.5	-/82.4	0.623	0.677	48.0	-	-/82.0	-/82.1
MulT[1]	0.846	0.725	40.4	46.7	81.7/83.4	81.9/83.5	0.564	0.731	52.6	54.1	80.5/83.5	80.9/83.6
MISA[1]	0.804	0.764	-	-	80.8/82.1	80.8/82.0	0.568	0.724	-	-	82.6/84.2	82.7/84.0
Self-MM[1]	0.717	0.793	46.4	52.8	82.9/84.6	82.8/84.6	0.533	**0.766**	53.6	55.4	82.4/85.0	82.8/85.0
TFR-Net[1]	0.721	0.789	46.1	53.2	82.7/84.0	82.7/84.0	0.551	0.756	52.3	54.3	81.8/83.5	81.6/83.8
MMIM[1]	0.712	0.790	**46.9**	53.0	**83.3**/85.3	**83.4**/85.4	0.536	0.764	53.2	55.0	82.5/85.0	82.4/85.1
Ours	**0.709**	**0.798**	45.8	**53.7**	82.7/**85.7**	83.2/**86.0**	**0.532**	0.758	**54.2**	**55.9**	**84.7/85.1**	**85.0/85.1**

Table 3: Overall performance comparison on CMU-MOSI and CMU-MOSEI datasets under complete data setting. [1] means the results provided by (Sun et al. 2023).

Metrics

We evaluate the mean absolute error (MAE) and pearson correlation coefficient (Corr) for regression, seven-class accuracy (Acc-7), five-class accuracy (Acc-5), binary accuracy (Acc-2), and F1 score on CMU-MOSI and CMU-MOSEI. Following (Yuan et al. 2021), we compute the Area Under Indicators Line Chart (AUILC) value for each above metric to evaluate the overall performance of the model under different modality missing rates. Suppose that the performance on a metric under increasing missing rates $\{r_0, r_1, \ldots, r_t\}$ is $\{x_0, x_1, \ldots, x_t\}$. The AUILC value is formally defined as follows:

$$AUILC = \sum_{i=0}^{t-1} \frac{1}{2}(x_i + x_{i+1})(r_{i+1} - r_i) \qquad (20)$$

Implementation Details

We implement the proposed model using the PyTorch framework. The model is trained with a single Nvidia GeForce RTX 3090 GPU. We employ the technique of Random Masking to simulate incomplete modal data sets by independently generating random temporal masks for each modality. For the audio and video modality, we add white Gaussian noise on the original feature with a zero vector while for the text modality (Hazarika et al. 2022), and we replace the original token with the [UNK] token. For model training, we employ an Adam optimizer and adopt an early-stopping strategy with a patience of 6 epochs. In both datasets, the learning rate is set to 0.002 and the batch size is 24.

Baselines

We consider general MSA methods and targeted methods as baselines to assess the comprehensive performance of our methods in both incomplete and complete dataset settings.

TFN (Zadeh et al. 2017) uses a multi-dimensional tensor by calculating the outer product among different modalities to capture unimodal, bimodal and trimodal interactions.

LMF (Liu et al. 2018) leverages low-rank multimodal fusion methods that utilize low-rank tensors to improve efficiency and reduce the complexity of tensor fusion.

MulT (Tsai et al. 2019) integrates directional pairwise cross-modal attention to enhance three sets of Transformers, facilitating end-to-end multimodal data interaction and latent stream adaptation without explicit alignment.

MISA (Hazarika, Zimmermann, and Poria 2020) addresses the distributional gaps in multimodal signals by projecting each modality into two subspaces: a shared, modality-invariant space that minimizes gaps and a unique, modality-specific space for distinct features.

Self-MM (Yu et al. 2021) uses a self-supervised label module for unimodal supervision and combines multimodal and unimodal training to learn shared and distinct features.

TFR-Net (Yuan et al. 2021) uses a reconstruction module with SmoothL1Loss to effectively address random missing data in non-aligned modality sequences and learn semantic features in the case of missing modalities.

MMIM (Han, Chen, and Poria 2021) improves performance by maximizing mutual information for task-relevant multimodal fusion, using parametric and non-parametric approximations, and co-training with the main task.

Unimodal Experients

As mentioned earlier, due to the significant differences in information density across modalities, different modalities require varying levels of processing. To verify this idea, we conducted experiments with unimodal and simple fused multimodal representations. Specifically, unimodal features were encoded through the fully connected layer and then classified, while multimodal features were simply concatenated together and fed into the fully connected layer for classification. Table 2 indicates that textual information is richer than visual or audio, suggesting noise or redundancy in the latter. This indicates the need for denoising and effective feature integration across modalities to enhance overall performance. Simple concatenation-based fusion improves unimodal performance, highlighting the importance of multimodal fusion. This demonstrates the need for cross-modal denoising and effective multimodal fusion to enhance overall performance.

Complete Modality Setting

We compare our model with MSA baselines using complete data to demonstrate its validity and competitiveness.

From Table 3, we observe that our model achieves the best performance in almost all metrics. We attribute these

Model	CMU-MOSI						CMU-MOSEI					
	MAE	Corr	Acc-7	Acc-5	Acc-2	F1	MAE	Corr	Acc-7	Acc-5	Acc-2	F1
TFN[1]	1.316	0.308	22.3	23.7	61.0/60.9	59.7/59.7	0.695	0.500	46.1	46.6	75.2/74.1	73.4/71.5
LMF[1]	1.310	0.299	21.5	22.7	59.7/59.3	56.4/56.1	0.718	0.447	45.3	45.7	72.2/73.9	69.5/69.4
MulT[1]	1.263	0.348	23.1	24.6	63.1/63.2	60.7/61.0	0.700	0.504	46.3	46.8	74.4/75.1	72.9/72.6
MISA[1]	1.202	0.405	25.7	27.4	63.9/63.7	59.0/58.8	0.698	0.514	45.1	45.7	**75.2/75.7**	**74.4**/74.0
Self-MM[1]	1.162	0.444	**27.8**	30.3	**66.9**/67.5	65.4/66.2	0.685	0.507	46.7	47.3	75.1/75.4	73.7/72.9
TFR-Net[1]	1.156	0.452	27.7	30.5	67.6/**67.8**	65.7/66.1	0.689	0.511	46.9	47.3	74.7/74.2	73.5/73.4
MMIM[1]	1.168	0.450	27.0	29.4	66.8/66.9	64.6/65.8	0.694	0.502	45.9	46.4	74.9/72.4	74.4/69.3
Ours	**1.115**	**0.461**	27.5	**30.5**	66.4/67.0	**69.3/69.8**	**0.680**	**0.524**	**47.0**	**47.9**	72.8/75.0	73.0/**76.2**

Table 4: Overall performance comparison on CMU-MOSI and CMU-MOSEI datasets in the incomplete modality setting. The reported results are the AUILC values for each evaluation metric, calculated under missing rates of $\{0, 0.1, \ldots, 1.0\}$.

pleasing results to the effective multi-focus fusion strategy and multi-branch collaborative structure of the model. The macro-fusion stream and micro-fusion stream of the former disperse the focus to local and global features, while the latter comprehensively integrates Local and global features in each fusion stream. For example, compared to the second-ranked model in each metric, our model improves by 0.005 in Corr, 0.6% in F1, 0.5% in Acc-5 on the CMU-MOSI dataset. On the CMU-MOSEI dataset, it improves by 0.6% in Acc-7, 0.5% in Acc-5, and 2.2% in F1. In conclusion, the above results demonstrate the effectiveness of the model.

Model Robustness Study

Then, we focus on evaluating the robustness of the model under conditions of random modality missing. In the CMU-MOSEI dataset, as shown in Table 4, our model outperforms the state-of-the-art methods in nearly all metrics. For instance, there is a 0.6% improvement in Acc-5 and a 2.2% increase in the F1 score. This demonstrates that the network comprehensively considers the fusion and interaction of global dimensional features and local dimensional features under missing data conditions. Meanwhile, it successfully utilizes multiple branch to coordinate the fusion and reconstruction process to reconstruct and enhance insufficient data. The results in the CMU-MOSI dataset also show that our model surpasses state-of-the-art methods in metrics such as MAE, Corr, Acc-5, and F1. At the same time, we observed that the model's performance on the Acc-2 metric is relatively weak. By analyzing the statistical features of the dataset samples, we infer the possible reasons: Firstly, the sample polarity distribution is uneven, with more positive and negative data, which causes data imbalance and may lead the model to favor the more frequent class. Secondly, as a binary classification metric, Acc-2 has low class distinguishability and struggles to capture the subtle differences between the two classes. These factors together explain the model's underperformance on this metric. However, these experimental results demonstrate the superiority and robustness of our model in unstable environments.

Ablation Study

We investigate the contribution of the macro fusion and the pseudo-Siamese micro fusion as well as the reconstruction module in the multi-branch strategy. From the Table 5, we observe that removing any module results in varying degrees of performance degradation. Specifically, the impact of the absence of the reconstruction module is the most pronounced, resulting in a 1.4% degradation for Acc-7 and a 0.7% degradation for Acc-5. This underscores the importance of the reconstruction module when dealing with missing data. Compared with micro-fusion stream, the effect of removing macro-fusion stream is not significant. This shows that macroscopic global features carry less useful information density than microscopic local features, but they are indeed necessary. These studies confirm the effectiveness of the proposed module in improving model performance.

Model	Corr	Acc-2	Acc-5	Acc-7	F1
w/o F-1	0.451	66.4/66.4	29.3	27.1	69.1/69.0
w/o F-2	0.461	66.1/66.0	29.0	26.3	68.0/67.7
w/o Recon	0.422	65.8/66.0	29.0	26.1	68.2/68.4
MFMB-Net	**0.461**	**66.4/67.0**	**30.3**	**27.5**	**69.3/69.8**

Table 5: Ablation study on CMU-MOSI dataset. F-1 means the macro fusion while F-2 means the micro fusion. Recon means the reconstrcution module.

Conclusion and Future Work

In this paper, we propose MFMB-Net, a multi-focus-driven multi-branch network. It synergistically integrates fusion and reconstruction to simultaneously capture global and local dynamics within and between modalities. Specifically, it abandons the same symmetrical parallel processing structure and adopts a multi-branch structure. The multi-focus module in the macro fusion stream uses coarse hubs to allocate focus to key information in the global dimension, and the micro fusion stream is used to integrate key information in the local dimension. The two-stream fusion branch collaborate with the reconstruction process to reconstruct and enhance the missing data, thereby improving the representation fidelity of the lost information. In future work, we will explore more efficient strategies for handling missing data and extend the network's applicability to address data imbalance issues. Additionally, we plan to incorporate external knowledge to improve robust representation and fusion.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (No. NT2024020), the Open Fund of the State Key Laboratory for Novel Software Technology (No. KFKT2024B27), the National Science Foundation of China (No. 62402215), and China Postdoctoral Science Foundation (No. 2023M741685).

References

Chen, S.; Xie, E.; Ge, C.; Chen, R.; Liang, D.; and Luo, P. 2023. CycleMLP: A MLP-Like Architecture for Dense Visual Predictions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45: 14284–14300.

Cheng, J.; Fostiropoulos, I.; Boehm, B.; and Soleymani, M. 2021. Multimodal phased transformer for sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2447–2458.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, 960–964. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Geva, M.; Schuster, R.; Berant, J.; and Levy, O. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5484–5495.

Guo, J.; Tang, Y.; Han, K.; Chen, X.; Wu, H.; Xu, C.; Xu, C.; and Wang, Y. 2022. Hire-MLP: Vision MLP via Hierarchical Rearrangement. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 816–826.

Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.-p.; and Poria, S. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 international conference on multimodal interaction*, 6–15.

Han, W.; Chen, H.; and Poria, S. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9180–9192.

Hazarika, D.; Li, Y.; Cheng, B.; Zhao, S.; Zimmermann, R.; and Poria, S. 2022. Analyzing Modality Robustness in Multimodal Sentiment Analysis. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 685–696.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, 1122–1131.

Lin, R.; and Hu, H. 2022. Multimodal Contrastive Learning via Uni-Modal Coding and Cross-Modal Prediction for Multimodal Sentiment Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 511–523.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors.

Lv, F.; Chen, X.; Huang, Y.; Duan, L.; and Lin, G. 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2554–2562.

Morency, L.-P.; Mihalcea, R.; and Doshi, P. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, 169–176.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. volume 34, 14200–14213.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paraskevopoulos, G.; Georgiou, E.; and Potamianos, A. 2022. Mmlatch: Bottom-Up Top-Down Fusion For Multimodal Sentiment Analysis. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4573–4577.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 6892–6899.

Sun, L.; Lian, Z.; Liu, B.; and Tao, J. 2023. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*.

Tang, J.; Li, K.; Jin, X.; Cichocki, A.; Zhao, Q.; and Kong, W. 2021. Ctfn: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5301–5311.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. volume 34, 24261–24272.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000–6010.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 10790–10797.

Yuan, Z.; Li, W.; Xu, H.; and Yu, W. 2021. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, 4400–4407.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114.

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018b. 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)*, 2236–2246.

Zeng, J.; Liu, T.; and Zhou, J. 2022. Tag-assisted multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1545–1554.

Zeng, J.; Zhou, J.; and Liu, T. 2022. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2924–2934.

Zhang, H.; Wang, Y.; Yin, G.; Liu, K.; Liu, Y.; and Yu, T. 2023. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 756–767.

Zhao, J.; Li, R.; and Jin, Q. 2021. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2608–2618.