Multi-level Contrastive Learning: Hierarchical Alleviation of Heterogeneity in Multimodal Sentiment Analysis

Cunhang Fan, *Member, IEEE,* Kang Zhu, Jianhua Tao, *Senior Member, IEEE,* Guofeng Yi, Jun Xue, *Student Member, IEEE,* Zhao Lv, *Member, IEEE,*

Abstract—Recently, multimodal fusion efforts have achieved remarkable success in Multimodal Sentiment Analysis (MSA). However, most of the existing methods are based on model-level fusion, and the challenge of heterogeneity between modalities is not well resolved. Heterogeneity lies in the different feature distributions and distinct representation spaces among different modalities. To mitigate this problem, we propose that fusion is a progressive process, and we introduce a novel multi-level contrastive learning and multi-layer convolution fusion (MCL-MCF) method for MSA. Due to the relationships among multimodal data, the fusion process that involves single-modal to single-modal, single-modal to bimodal or trimodal, and higher-level fused modality semantic consistency is divided into three levels. The first-level contrast learning alleviates heterogeneity between unimodal modalities at the early level of multimodal feature fusion. The second-level contrast learning mitigates heterogeneity between unimodal and fused modalities. At the third level, we introduce a tensor convolution fusion (TCF) module that extracts high-level semantic features from the fused modalities and mitigates heterogeneity at the higher feature level through contrastive learning. To simulate fusion as a progressive process, MCF is proposed to fuse shallow and deep features to model complex relationships among modalities. Experiments on three public datasets show our approach's state-of-the-art performance.

Index Terms—multimodal sentiment analysis, multi-level contrastive learning, convolution fusion, heterogeneity.

1 Introduction

MULTIMODAL sentiment analysis (MSA) aims to predict emotional scores from audio, visual, and text features. MSA has been widely used and has become a popular topic of research. It has been widely applied in areas such as marketing management [1] [2], social media analysis [3] [4], and human-computer interaction [5] [6]. Although it is easy for humans to perceive the world through comprehensive information acquired via multiple sensory organs [7], the question of how to endow machines with analogous cognitive capabilities is still unresolved. One of the challenges we are facing is the heterogeneity gap in multimodal data [8]. This gap arises from the initial unequal subspaces of feature vectors extracted from different modalities, leading to completely different vector representations for semantically similar elements. This phenomenon poses a challenge to the comprehensive utilization of multimodal data by subsequent machine learning modules [9]. Researchers have made remarkable strides in the realm of designing multimodal feature fusion methods [10]–[14]. Nevertheless, limited consideration has been given to addressing the disparities in heterogeneity among multimodal features. Currently, there are two main methods in MSA: the first involves geometric operations performed on feature vectors to achieve feature fusion, while the second involves the use of a transformer (attention) to design complex feature fusion methods.

There are numerous ways to perform geometric operations on eigenvectors, such as simple splicing of eigenvectors, outer products, stacking and vector offset correction, or weighted summation. The tensor fusion network (TFN) [15] uses three modalities and adds a dimension to perform three geometric outer product operations to achieve the fusion of three modal features. The TFN has a good fusion effect when using the vector outer product, but it does not consider the heterogeneity of multimodal features that could hinder its effect. In [16], the authors proposed the multimodal adaptation gate (MAG) mechanism. They used attention gating to generate features that shift the position of linguistic features in the vector space to shift the linguistic feature vectors to the optimal position for fusion. The offset used by MAG is derived from the fusion of attention,

- *This work is supported by the STI 2030-Major Projects (No. 2021ZD0201500), the National Natural Science Foundation of China (NSFC) (No. 62201002), the Excellent Youth Foundation of Anhui Scientific Committee (No. 2208085J05), the Special Fund for Key Program of Science and Technology of Anhui Province (No. 202203a07020008), the Open Fund of Key Laboratory of Flight Techniques and Flight Safety, CAAC (No. FZ2022KF15).*(Cunhang Fan and Kang Zhu contribute equally and share the co-first authorship.)(Corresponding authors: Jianhua Tao and Zhao Lv.)

- *Cunhang Fan and Zhao Lv are with the Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China and with the Key Laboratory of Flight Techniques and Flight Safety China (e-mail:cunhang.fan@ahu.edu.cn; kjlz@ahu.edu.cn).*

- *Kang Zhu, Guofeng Yi, and Jun Xue are with the Anhui Province Key Laboratory of Multimodal Cognitive Computation, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: E22201061@stu.ahu.edu.cn; e21301183@stu.ahu.edu.cn; E21201068@stu.ahu.edu.cn;).*

- *Jianhua Tao is with the Department of Automation, Tsinghua University, China (e-mail:jhtao@tsinghua.edu.cn).*

and the heterogeneity between multimodal features hinders the generation of the optimal offset, thereby affecting the effectiveness of subsequent operations.

Moreover, transformers (attention) have proven to be effective in the fields of natural language processing and computer vision. Researchers at MSA have proposed a variety of excellent models [17] [18] [19]. [20] proposed that the message hub mechanism adopts a multi-layer cross-attention method to solve the problem of time asynchrony between multimodal features, but the significant differences in data distribution between multimodal features may lead to uneven attention weight allocation, thereby affecting the fusion effect and hindering cross-modal attention. Multi-layer operation may become increasingly hindered, which will eventually affect the experimental results. [21] used a shared-private mask and cross-attention mechanism to extract features and used a linear layer in the fusion stage, but the heterogeneity between multimodal features hindered the fusion of the linear layer, reducing the effect. Although these methods have strengths in feature fusion, the heterogeneity between multimodal features can hinder feature fusion at different locations and stages of the features. A common approach to address this problem is to project the heterogeneous features into a shared subspace, where multimodal data with similar semantics are represented by similar vectors [22]. Therefore, the main objective of multimodal representation learning is to narrow the distribution gap in a joint semantic subspace while preserving intact modality-specific semantics [8].

Multimodal feature fusion encompasses the integration of both basic information features and semantic information features. Basic information features, such as texture and details, capture low-level characteristics. On the other hand, semantic information features represent high-level features obtained through multiple stages of feature extraction. Despite expressing the same sentiment, different modalities exhibit significant differences in their forms. The heterogeneity of basic information and semantic information features lies in the different feature distributions and distinct representation spaces among different modalities. This heterogeneity exists not only within unimodal modalities but also across unimodal and multimodal modalities, as well as among different multimodal combinations. The challenge lies in devising approaches that can mitigate the heterogeneity of basic information and semantic information features among different sentiment modalities, thereby facilitating effective multimodal fusion. This remains a significant challenge in the field.

In this paper, we propose the novel concept that multimodal feature fusion is a progressive process. The fusion process, which encompasses the relationships between multimodal data, is categorized into three levels (early, middle, and late). These levels involve the fusion of single-modal to single-modal, and single-modal to bimodal or trimodal features and the maintenance of higher-level fused modality semantic consistency. Both multi-level contrastive learning (MCL) and multi-layer convolution fusion (MCF) have been designed with multiple levels or layers to align with this framework. The MCL consists of three levels. The first and second level primarily focus on addressing the heterogeneity of basic information features, while the third level primarily addresses the heterogeneity among high-level semantic information features. The first level aids in alleviating heterogeneity in individual modalities, making it easier to fuse information between modalities at the beginning of fusion. The second level reduces heterogeneity between single modalities and bimodal or trimodal data, further enhancing the fusion process. For the third level, we introduce the tensor convolution fusion (TCF) module, inspired by the TFN. The multimodal matrix obtained through the outer product is referred to as a multimodal "image" which may contain redundant information. Therefore, we utilize multi-layer convolution to extract more meaningful features. These extracted features are closer to the late fusion features, thereby facilitating the promotion of feature fusion through third-level contrastive learning. Following the reduction of multimodal heterogeneity, we implement a two-layer convolutional fusion approach: 1) the fusion of unimodal features results in the acquisition of the fused features of the first layer; 2) we independently extract advanced features from each unimodal feature and subsequently fuse them with the fused features obtained from the first layer. Benefiting from the assistance of multi-level contrastive learning, the fusion of multimodal features occurs progressively, starting from low-level features and advancing to high-level features, resulting in excellent outcomes.

Experiments on three datasets demonstrate that our method achieves impressive results. Multi-level contrastive learning is effective in alleviating the heterogeneity between multiple modalities, and multi-level convolutional fusion is the icing on the cake. The novel contributions of our work can be summarized as follows:

- Inspired by the fusion of objects in the natural world, we conceptualize multimodal fusion as a continuous process, dividing the entire procedure into three steps. We designed MCL-MCF. MCF simulates the continuous fusion process, while MCL, through multi-level alleviation of heterogeneity, assists MCF in achieving multi-level fusion. Their collaborative operation yields optimal fusion results.
- Taking inspiration from the TFN, we design a TCF module and apply it for high-level feature extraction. We conduct a comprehensive experimental analysis to evaluate the effectiveness of the multimodal "image".
- Extensive experiments on three public datasets, CMU-MOSI, CMU-MOSEI, and CH-SIMS were performed. We outperform prior methodologies and achieve results to those of superior state-of-the-art models. [1]

2 Related Works

2.1 Multimodal Sentiment Analysis (MSA)

In sentiment analysis, there are various methods of multimodal fusion, roughly divided into tensor-based [15] [23] [24], GAN-based [9], attention-based [25] [26] [27] [16]

1. Codes are released at https://github.com/Zhudogsi/MCL-MCF

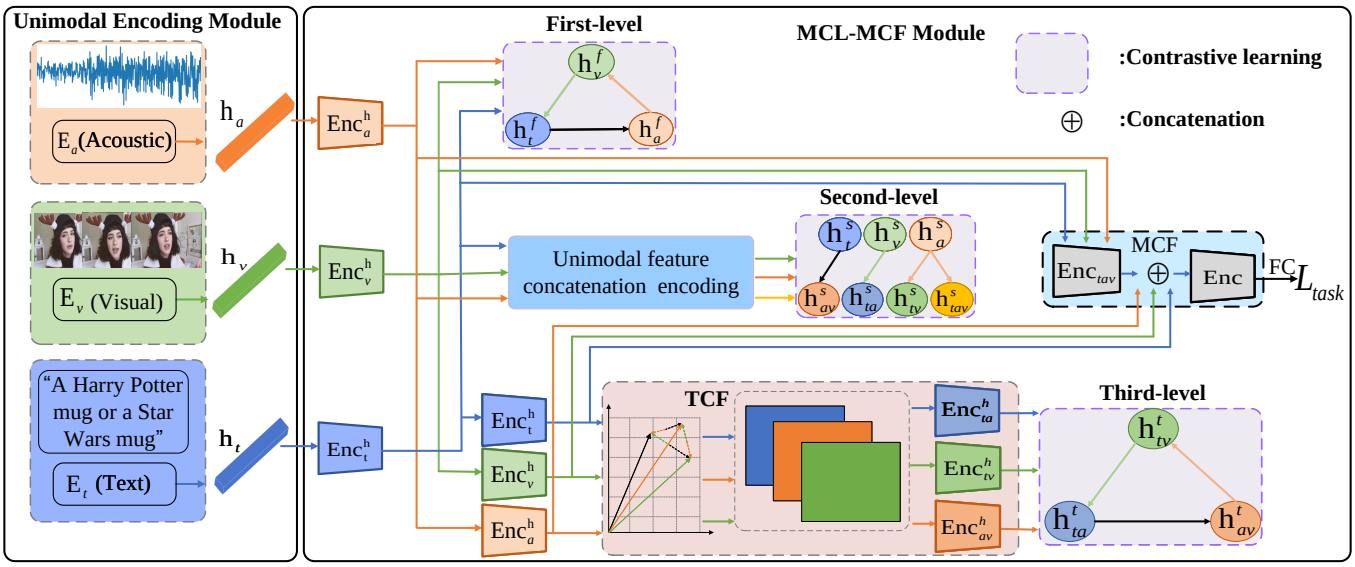


Fig. 1: Overall architecture of the MCL-MCF model. MCL-MCF mainly consists of the unimodal encoding module, MCL, MCF, and TCF; MCL has three levels of contrastive learning and the unimodal feature concatenation encoding module for concatenating and encoding the unimodal features; TCF is the advanced feature extraction module, which extracts high-level features required for third-level contrastive learning; and MCF has two layers of convolutional fusion for primary and advanced feature fusion and sentiment intensity prediction for sentiment prediction. The encoder in TCF is a two-dimensional convolution, while other encoders are one-dimensional convolutions.

[28], graph-based [29] [30], operation-based [31], translation-based [32] [33] [34], routing-based methods [35], etc. [36] employ information theory to quantify interactions within multimodal data and the interplay captured by multimodal models. [37] proposed a disentanglement translation network (DTN) with slack reconstruction to capture essential information attributes and reduce redundancy. [38] proposed a novel meta-learning-based paradigm that preserves the advantages of unimodal approaches, further enhancing the performance of multimodal fusion. To produce a comprehensive fusion representation, these models employ complex methods to design multimodal fusion. However, many of these models overlook the substantial heterogeneity differences among multimodal data. The use of complex fusion methods alone may not effectively address the inherent heterogeneity between different modalities, thereby impacting the quality of multimodal data fusion. Incorporating additional auxiliary tasks to alleviate the heterogeneity differences between modalities is considered an effective strategy.

2.2 Contrastive Learning

In recent years, contrastive learning has gained prominence among researchers and has found widespread application in MSA. [39] introduced multimodal infomax (MMIM), which utilizes contrastive predictive coding (CPC) for single-modal predictive contrastive learning. [40] proposed multimodal contrastive learning (MMCL), which also uses the fused modality to perform predictive contrastive learning on a single modality. [41] proposed the contrastive learning and multi-layer fusion (CLMLF) method, which employs labeled and unlabeled contrastive learning for emotion-related tasks. [42] presented a framework hycon for hybrid contrastive learning of tri-modal representation, utilizing various positive and negative sample selection methods for contrastive learning for multimodal sentiment analysis. FACTORCL decomposes information into shared and unique representations, achieving optimal results through the maximization and minimization of mutual information. By capturing both shared and unique information, it achieves optimal outcomes [43]. [44] proposed face-to-face contrastive learning (F2F-CL), which models social interactions by decomposing nodes and contextualizing multimodal face-to-face interactions along the boundaries of conversational turns. Contrastive learning in the image field [45] [46], such as [47] [48] [49] [50], is widely used and often requires data enhancement methods [50] [51] [52] to generate positive and negative samples. However, most contrastive learning-based MSA models do not include a data augmentation component. Since multiple modalities inherently contain the same semantic information, they can effectively serve as augmented modalities for each other. Hence, contrastive learning is well suited for MSA.

3 Method

3.1 Problem Definition

The task of MSA is to use multimodal information to identify the polarity of expressed emotions. The input signal consists of three parts, corresponding to three modes: text(t), visual(v), and acoustic(a). The input to the model is unimodal sequences $X_m \in R^{l_m \times d_m}$, where l_m is the sequence length and d_m is the representation vector dimension of modality m, $m \in (t, a, v)$. The goal of designing a model is to extract and integrate task-related information from these input vectors, form a unified representation, and use it to

make accurate predictions of the true value y reflecting the intensity of emotions.

3.2 Overall Architecture

Figure 1 shows the structure of the entire model, which consists of three parts: a unimodal encoding module, an MCL-MCF module, and a sentiment intensity prediction task. In the unimodal encoding module, the text modality utilizes pre-trained BERT [53] to extract features from raw text, while the audio and visual modalities are extracted using long short-term memory (LSTM) [54]. Subsequently, these features are mapped to the same dimension via one-dimensional convolution and input into the MCL-MCF module. The MCL-MCF module consists of two parts: multi-level contrastive learning and multi-layer convolution fusion. The MCL consists mainly of three levels. The first-level contrastive learning is performed between single modalities. The second-level contrastive learning first concatenates and encodes modalities with the same dimensions for preliminary fusion and then performs contrastive learning with single modalities. In the third level of contrastive learning, the single modality is first input into the TCF module, where an outer product operation is performed to generate a matrix. Subsequently, feature extraction entails a three-layer convolution process, followed by flattening and contrastive learning. The feature fusion module is composed of two layers of convolution fusion, with the first layer being the fusion between single modalities and the second layer being the fusion of high-level features, extracted from single modalities, with the output of the first layer. Ultimately, sentiment intensity prediction is executed.

$$h_t = BERT(X_t; \theta_t^{BERT}) \tag{1}$$

$$h_m = LSTM(X_m; \theta_m^{LSTM}) \quad m \in \{a, v\} \tag{2}$$

where h_t is the head embedding extracted from the output of the last layer of BERT and h_m is the feature of the last time step of LSTM, $m \in \{a, v\}$.

3.3 Multi-level Contrastive Learning

For the multi-level contrastive learning module, we design three contrastive learning tasks, namely, multi-level contrastive learning. The first level is unimodal contrastive learning, the second level is bimodal or trimodal contrastive learning, and the third level is a convolution operation after the outer product of two modalities to extract more advanced fusion features for contrastive learning. As shown in Figure 1, the multi-level contrastive learning module consists of three parts, where h^f represents the characteristics of the input of the first-level contrastive learning, h^s represents the characteristics of the input of the second-level contrastive learning and h^t represents the characteristics of the input of the third-level contrastive learning.

3.3.1 First-level Contrastive Learning

In the early stage of fusion, first-level contrastive learning is employed to mitigate heterogeneity among single modalities, assisting the initial fusion among single modalities. The first-level contrastive learning is shown in Figure 1.

Algorithm 1: Contrastive Learning Method

Require: $X_1 \in R^{B \times D_1}$, $X_2 \in R^{B \times D_2}$ The Project is a linear project; B denotes of batch size; D denotes of dimensions.;

$X_1^{'} = Project(X_1)$;
$X_2^{'} = Project(X_2)$;
$X = L2\_normalize(X_1^{'}, axis = 1)$;
$Y = L2\_normalize(X_2^{'}, axis = 1)$;
$CL\_label = arange(B)$;
$M_{XY} = dot(X, Y.T) * exp(\tau)$;
$loss\_\alpha = Cross\_Entropy(M_{XY}, CL\_label)$;
$loss\_\beta = Cross\_Entropy(M_{XY}.T, CL\_label)$;
$Loss = (loss\_\alpha + loss\_\beta)/2$;
return $Loss$

TABLE 1: Explanation of symbols used.

	Symbol Definition
t, a, v	text, audio, visual
M	the elemental matrix used for contrastive learning
M_{ij}	the element in the ith row and jth column of matrix M
X^f	variable used for first-level contrastive learning
X^s	variable used for second-level contrastive learning
X^t	variable used for third-level contrastive learning
Lf	the loss for the first-level contrastive learning
Ls	the loss for the second-level contrastive learning
Lt	the loss for the third-level contrastive learning

$h_m^f \in R^{b \times d_m} m \in (t, a, v)$ is projected to the same dimension via convolution, where b is the batch size. These vectors are then fed into Algorithm 1 to compute the infoNCE loss. The difference in heterogeneity between unimodal and multimodal learning is reduced by first-level contrastive learning, laying the foundation for unimodal fusion. Because there are three modes, there will be three losses, namely, Lf^{ta}, Lf^{tv}, and Lf^{av}. The total loss of first-level contrastive learning is as follows:

$$\mathrm{M}_{mn}^f = h_m^f \times (h_n^f)^T \tag{3}$$

$$Lf^{mn} = -\frac{1}{b} \sum_{i=1}^{b} \log \left(\frac{\exp(\frac{M_{ii}^f}{\tau})}{\sum_{j=1}^{b} \exp(\frac{M_{ij}^f}{\tau})} \right) \tag{4}$$

$$Lf_{total} = Lf^{ta} + Lf^{tv} + Lf^{av} \tag{5}$$

where b is the batch size; M_{ij} represents the element in the ith row and jth column of matrix M; $i \neq j$; τ is the temperature coefficient; Lf^{ta} is the loss generated by contrastive learning between text and acoustics; $Lf^{tv}$and Lf^{av} are calculated in the same way; and $m, n \in (t, a, v)$ $m \neq n$.

3.3.2 Second-level Contrastive Learning

In the middle stage of fusion, the second-level contrastive learning begins with enhanced preliminary fusion, supported by first-level contrastive learning. It was subsequently applied to alleviate heterogeneity between single and multimodal data, laying the foundation for the fusion

of single and multimodal information. The structure of the second-level contrastive learning is depicted in Figure 1. Unimodal feature concatenate encoding involves the concatenation and encoding of unimodal features. We aggregate multiple unimodal features and then apply one-dimensional convolution for feature fusion, subsequently projecting them to the same dimension. To fully slow the heterogeneity difference between unimodal and fused multimodal features, we utilize the fused features from two modalities and combine them with another modality, which is then input into Algorithm 1. We employ equation (10) (for the trimodal fusion feature, use equation (11)) to calculate the Ls loss. The experimental findings suggest that interaction between the audio modality and trimodal features can effectively reduce the heterogeneity issue at the second-level features. Second-level contrastive learning contributes to mitigating the heterogeneity between single-modality and multimodal feature fusion. The total loss of second-level contrastive learning is as follows:

$$h^s_{mn} = Conv1D([h_m, h_n]) \tag{6}$$

$$h^s_{tav} = Conv1D([h_t, h_a, h_v]) \tag{7}$$

$$\mathrm{M}^s_{mn-k} = h^s_{mn} \times (h^s_k)^T \tag{8}$$

$$\mathrm{M}^s_{tav-a} = h^s_{tav} \times (h^s_a)^T \tag{9}$$

$$Ls^{mn-k} = -\frac{1}{b}\sum_{i=1}^{b} \log\left(\frac{\exp(\frac{M^s_{mn-k(ii)}}{\tau})}{\sum_{j=1}^{b}\exp(\frac{M^s_{mn-k(ij)}}{\tau})}\right) \tag{10}$$

$$Ls^{tav-a} = -\frac{1}{b}\sum_{i=1}^{b} \log\left(\frac{\exp(\frac{M^s_{tav-a(ii)}}{\tau})}{\sum_{j=1}^{b}\exp(\frac{M^s_{tav-a(ij)}}{\tau})}\right) \tag{11}$$

$$\mathrm{Ls}_{total} = \mathrm{Ls}^{ta-v} + \mathrm{Ls}^{tv-a} + \mathrm{Ls}^{av-t} + \mathrm{Ls}^{tav-a} \tag{12}$$

where b is the batch size; $m, n, k \in (t, a, v)$ $m \neq n \neq k$; M_{ij} represents the element in the ith row and jth column of matrix M; $i \neq j$; τ is the temperature coefficient; h^s_{mn} represents the result of the initial fusion of h_m and h_n after convolution (h^s_{tav} similar operation); $\mathrm{M^s_{mn-k}}$ represents the result of the matrix product of fusion mode h^s_{mn} and single mode h_k (M^s_{tav-a} similar operation); and Ls^{ta-v} represents the result of contrastive learning between the fusion model h^s_{ta} and the single modality h^s_t (Ls^{tv-a} , Ls^{av-t} , and Ls^{tav-a} are similar).

3.3.3 Third-level Contrastive Learning

In the late stage of fusion, the TCF module is designed for advanced feature extraction. The acquired advanced features are utilized for third-level contrastive learning, mitigating heterogeneity between multimodal sources in the later stages of fusion. The third-level contrastive learning is shown in Figure 1. Unlike with the first and second level, we initially expanded the dimension of the single modality to 512 and input it into the TCF encoder module. The TCF module comprises two key components: 1) geometric operations conducted on single modalities and 2) feature extractors. Geometric operations conducted on single modalities refer to using the outer product of single modalities to obtain a multimodal "image" of shape (1, 512, 512), which is the result of fusion. Experimental findings have indicated that this multimodal "image" often contains redundant information. A method akin to image analysis is employed for feature extraction. This process amplifies the number of channels and extracts features from different perspectives of the multimodal "image". Consequently, advanced features are derived after modality fusion, with subsequent channel reduction to mitigate the computational cost of 1. These processed data are then input into Algorithm 1. Third-level contrastive learning focuses on contrastive learning at the high-level feature fusion level, which contributes to mitigating the heterogeneity between modalities at the later level of fusion. The total loss of the third-level contrastive learning is as follows:

$$\mathrm{h}^t_{cm} = \mathrm{Conv1D}(\mathrm{h}^t_m) \tag{13}$$

$$M^t_{\mathrm{cmcn}} = h^t_{cm} \otimes h^t_{cn} \tag{14}$$

$$h^t_{mn} = Flatten(Conv2d(M^t_{cmcn})) \tag{15}$$

$$M = h^t_{mn} \times (h^t_{nk})^T \tag{16}$$

$$Lt^{mn-mk} = -\frac{1}{b}\sum_{i=1}^{b} \log\left(\frac{\exp(\frac{M_{ii}}{\tau})}{\sum_{j=1}^{b}\exp(\frac{M_{ij}}{\tau})}\right) \tag{17}$$

$$Lt_{total} = Lt^{ta-tv} + Lt^{tv-av} + Lt^{ta-av} \tag{18}$$

where b is the batch size; $m, n, k \in (t, a, v)$ $m \neq n \neq k$; $\otimes$ is the vector outer product; M_{ij} represents the element in the ith row and jth column of matrix M; $i \neq j$; τ is the temperature coefficient; h^t_{cm} represents the result of one-dimensional convolution of single-mode m; M represents the matrix obtained by matrix multiplication of h^t_{mn} and $(h^t_{nk})^T$; M^t_{cmcn} represents the result of the outer product of the single mode m and n, respectively, after the formula (13); and Ls^{ta-tv} is the loss generated by contrastive learning between ta and tv, Ls^{tv-av}, and Ls^{ta-av} are the same.

3.4 Multi-layer Convolution Fusion

Figure 2(a) illustrates the conventional use of one-dimensional convolution. Taking a 1×1 convolution kernel as an example, features are continuously extracted as the kernel moves. However, this method solely captures data features merged within the convolution kernel's size and might not adequately encompass the overall contextual information of the data. In contrast, (b) in Figure 2 depicts our application of one-dimensional convolution. The number of convolution channels is the same as the number of word lengths, which is equivalent to observing data features from multiple perspectives. The size of the convolution kernel is fixed at 1×1.

Fusion is an incremental process characterized by its division into multiple levels. single-level fusion is inadequate, and adopting a multi-level fusion strategy becomes essential. Hence, we utilized this approach to design a two-layer multimodal fusion method. The first layer involves the

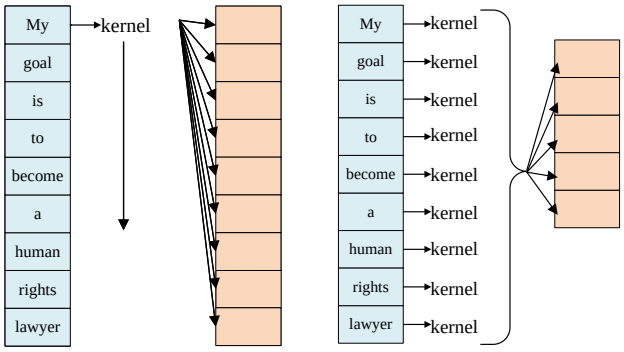
fusion of three single modalities to obtain a fused modality. The second layer extracts high-level features from the single modalities first and then fuses them with the fused modality obtained from the first layer. With the aid of three-level contrastive learning, the fusion of the first and second layers can achieve excellent results.

$$h^{'} = Conv1D([h_t, h_a, h_v]) \tag{19}$$

$$h_m^{'} = Conv1D(h_m) \quad m \in (t, a, v) \tag{20}$$

$$h = Conv1D([h_t^{'}, h_a^{'}, h_v^{'}, h^{'}]) \tag{21}$$

h_t, h_a and h_v are obtained via equation (1) or (2); $h_t^{'}$, $h_a^{'}$, and $h_v^{'}$ are calculated via equation (20); and h is used for sentiment score prediction (after FC).



(a) with 1 kernel of size 1x1 (b) with 9 kernels of size 1x1

Fig. 2: The total length of the data is 9, with a kernel size of 1x1 and a stride of 1 for both convolutions. (a) performs a standard one-dimensional convolution, while (b) has a kernel number identical to the length of the data.

3.5 Relationships among MCL, MCF, and TCF

The concept of continuous fusion is introduced, with the MCF simulating the continuous fusion of multiple modalities. In this continuous fusion process, heterogeneity poses a challenge. Mitigating heterogeneity at a single level is not sufficient to achieve the optimal fusion effect. The reduction in heterogeneity between modalities is achieved through multimodal shifting methods [55]. Contrastive learning is utilized to alleviate heterogeneity between single-modal and multimodal data [56]. Therefore, MCL is divided into three levels to alleviate modality heterogeneity at different stages. The alleviation of multi-level modality heterogeneity contributes to improving the overall multimodal fusion effectiveness of MCFs. First-level and second-level contrastive learning are applied to the early and middle stages, respectively. To better address multimodal heterogeneity and acquire advanced features closer to the later stages of fusion, inspired by the TFN, we designed a TCF for third-level contrastive learning, which achieved promising results. Although the first level alleviates heterogeneity between single modalities, merely mitigating heterogeneity within single modalities is insufficient for achieving optimal results in multimodal fusion [40] [41]. Therefore, we designed the second and third level. In the middle stage, the second level merged single modalities into multimodalities, reducing heterogeneity between single modalities and multimodalities. In the late stage, the third level reduces heterogeneity between multiple modalities. While each step of MCL has minimal differences, they each play distinct roles. Their collaborative efforts contribute to achieving the optimal fusion result. It is inappropriate for multimodal fusion to only consider reducing differences; the complementarity between modalities is equally important [57] [58] [59]. To avoid the confusing impact of complementary loss and model parameter quantity on model performance, we select fusion features from both shallow and higher-level features, excluding features for conducting comparative learning. The simultaneous application of the above steps optimizes the fusion effect.

4 Sentiment Intensity Prediction

Finally, after the effect of three-level contrastive learning, these three modalities improve in terms of the fusion effect, the mixed-modal fusion period, and the advanced feature fusion period, and the heterogeneity between the modalities greatly decreases. We perform multimodal fusion and calculate the prediction result loss, as in equation (22). The total loss of the model is shown in equation (23).

$$L_{reg} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{22}$$

where n is the number of training samples, the truth sentiment label is y_i, and the prediction of the final sentiment score is $\hat{y}_i$.

$$\mathrm{L}_{total} = L_{reg} + \alpha Lf_{total} + \beta Ls_{total} + \gamma Lt_{total} \tag{23}$$

The contribution weights of α, β, and γ multi-level contrastive learning are used to mitigate modal heterogeneity at different levels.

TABLE 2: Dataset split.

Dataset	Train	Valid	Test	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856
CH-SIMS	1368	456	457	2281

5 Experiments

In this section, we empirically evaluate the performance of MCL-MCF on MSA tasks with three publicly available academic datasets and present the experimental details, including datasets, baselines, and results.

5.1 Datasets and Evaluation

The CMU-MOSI [61] dataset is a popular benchmark database in MAS research. The data were collected from YouTube and consisted of 93 monologs in which speakers commented on specific topics. The dataset contains clips of 26,295 total words in 2,199 opinion video utterances annotated with sentiment strength labels ranging from -3

TABLE 3: Results on CMU-MOSI and CMU-MOSEI; All models use bert-base-uncased as the text encoder; In Acc-2 and F1-Score, the left of the "/" is calculated as negative/non-negative and the right is calculated as negative/positive; † indicates that the corresponding result is significantly better than the MMIM with p-value < 0.05 based on paired t-test. Performance Comparison between MCL-MCF and baselines on CMU-MOSI and CMU-MOSEI datasets. Baselines from [60]

	CMU-MOSI				CMU-MOSEI			
Models	MAE ↓	Corr↑	Acc-2↑	F1↑	MAE ↓	Corr↑	Acc-2↑	F1↑
TFN	0.925	0.662	78.3 / 80.2	78.2 / 80.1	0.570	0.716	81.0 / 82.6	81.1 / 82.3
LMF	0.931	0.670	77.5 / 80.1	77.3 / 80.0	0.568	0.727	81.3 / 83.7	81.6 / 83.8
MFN	0.951	0.665	77.9 / 80.0	77.8 / 80.0	0.575	0.720	81.8 / 84.0	81.9 / 83.9
MFM	0.948	0.664	77.7 / 80.0	77.7 / 80.1	0.580	0.722	80.3 / 83.4	80.7 / 83.4
MulT	0.918	0.685	79.0 / 80.5	79.0 / 80.5	0.564	0.732	81.3 / 84.0	81.6 / 83.9
MAG-BERT	0.730	0.789	82.4 / 84.6	82.2 / 84.6	0.558	0.761	81.9 / 85.1	82.0 / 84.3
MISA	0.752	0.784	81.8 / 83.50	81.7 / 83.5	0.550	0.758	81.6 / 84.3	83.8/85.3
Self-MM	0.731	0.785	82.7 / 84.9	82.6 / 84.8	0.540	0.763	82.6 / 85.2	82.6 / 85.2
MMIM	0.738	0.781	83.0 / 85.1	82.9 / 85.0	0.547	0.752	81.9 / 85.1	82.3 / 85.0
MTMD	0.705	**0.799**	84.0 / 86.0	83.9 / 86.0	**0.531**	**0.767**	**84.8** / 86.1	**84.8** / 86.1
MCL-MCF(Ours)	**0.692**†	**0.799**†	**84.9** / **87.3**†	**84.7/87.2**†	0.536†	**0.767**†	84.2/**86.4**†	84.4/**86.3**†

TABLE 4: Results on CH-SIMS. All models use bert-base-chinese as the text encoder; † means the corresponding result is significantly better than the Self-MM with p-value < 0.05 based on paired t-test.

Models	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
TFN	43.22	59.1	78.38	78.62
LMF	44.12	57.59	77.77	77.88
MFN	43.49	58.24	77.90	77.88
MulT	45.32	56.41	78.56	79.66
Self-MM	42.50	**59.52**	80.04	80.44
MCL-MCF(Ours)	**41.00**†	58.88†	**81.84**†	**81.82**†

(strongly negative) to +3 (strongly positive).

The CMU-MOSEI [62] dataset is a large-scale MSA and emotion recognition dataset consisting of 23,454 YouTube monolog video clips covering 250 different topics from 1,000 different speakers. The utterance dataset is composed of randomly selected review topics in various movies, annotated with sentiment scores between -3 and +3 and 6 different sentiment categories.

The CH-SIMS [63] dataset is a Chinese MSA dataset that not only contains unified multimodal annotations but also introduces independent unimodal annotations. The dataset consists of 2281 refined video clips from different movies, TV series and variety shows. Each sentiment score ranges from -1 (strongly negative) to 1 (strongly positive).

We use the same set of metrics that have been proposed and compared before the mean absolute error (MAE), which is the average mean absolute difference between the predicted and true values; Pearson correlation (Corr), which measures the degree to which predictions are skewed; binary classification accuracy (Acc-2); and F1 scores, which are calculated for non-negative/negative results.

5.2 Baselines

To fully validate the performance of MCL-MCF, we compare our model with several baselines. This earlier work is included, as are recent and more competitive baselines. The models we compare are as follows:

- **TFN** [15] is a tensor fusion network that uses a multidimensional tensor to capture interactions between different modalities, including unimodal, bimodal, and trimodal modalities. This is achieved by calculating the outer product between the different modalities within the tensor.
- **LMF** [23] is a low-rank multimodal fusion approach that involves decomposing stacked high-rank tensors into multiple low-rank factors, followed by an efficient fusion process using these factors.
- **MFM** [64] is a multimodal factorization model that connects an inference network and a generative network with intermediate modality-specific factors to facilitate the fusion process of reconstruction and recognition losses.
- **MFN** [65] is a memory fusion network that leverages LSTM-encoded information from each modality separately and a virtual attention network with multiview gated memory to explicitly account for cross-view interactions.
- **MulT** [27] is a multimodal transformer that builds a network of unimodal and cross-modal transformers and a complete fusion process.
- **MAG-BERT** [16] MAG is fine-tuned on specific datasets but struggles with multimodal language tasks. It introduces the multimodal adaptation gate for BERT and XLNet, enabling them to incorporate visual and acoustic data.
- **MISA** [57] represents item patterns in two distinct subspaces, modality-invariant and specific, to

provide a holistic view of multimodal data.

- **Self-MM** [66] is a self-supervised multitask learning approach that automatically generates unimodal labels weighted by multimodal labels to learn the similarities and differences between different modalities.

- **MMIM** [39] is a multimodal mutual information maximization method that maintains task-related information by maximizing unimodal input pairs and mutual information between multimodal fusion outputs and unimodal inputs.

- **MTMD** [60] views the learning process of modalities as multiple subtasks and introduces an innovative approach called multitask momentum distillation to narrow the gap between different modalities. This method employs a unimodal momentum model to consider modality-specific knowledge and utilizes adaptive momentum fusion factors in learning robust multimodal representations.

5.3 Implementation Details and Results

In all experiments, we utilized unaligned data provided by the open source [39]. The training set, validation set, and test set divisions for the three datasets are shown in Table 2. We used a single RTX 3090 for training. For unimodal feature extraction from text data, we use pre-trained BERT models with bert-base-uncased and bert-base-Chinese files, with an output dimension of 786. For sound and vision data, we use unidirectional LSTMs with an output dimension of 64 for feature extraction. In the TCF module, for the three datasets CMU-MOSI, CMU-MOSEI, and CH-SIMS, we use a convolution kernel size of 3×3 for the three-layer two-dimensional convolutional neural network, a step size of 1, and numerous channels of $[(1,4),(4,6),(6,1)]$, $[(1,2),(2,4),(4,1)]$, and $[(1,4),(4,7),(7,1)]$, respectively. We use Xavier-normal for parameter initialization. The learning rate hyperparameters for the three datasets are 1e-4. On the CMU-MOSI dataset, α, β, and γ are all set to 0.05. On the CMU-MOSEI dataset and CH-SIMS dataset, α, β, and γ are all set to 0.02. The results of our model are shown in Table 3. Compared with previous works, our model achieves state-of-the-art results with multiple indicators.

5.4 Experiments Results Summary

Table 3 and Table 4 present the experimental results of the MCL-MCF model. Our model achieves competitive results with both English and Chinese datasets (note that this assessment does not take into consideration comparisons of our model with excellent models that have significant differences in original data preprocessing and that are not publicly available). Across various metrics, including accuracy, correlation (Corr), and F1 score, our model outperforms the current state-of-the-art models. This suggests the effectiveness of our proposed method for mitigating the heterogeneity of multimodal features. Further feature analysis will be needed to provide deeper insights into this aspect.

TABLE 5: Ablation study of MCL-MCF on CMU-MOSI; MCL-F, MCL-S, and MCL-T respectively represent the first-level contrastive learning, the second-level contrastive learning and the third-level contrastive learning in the MCL module.

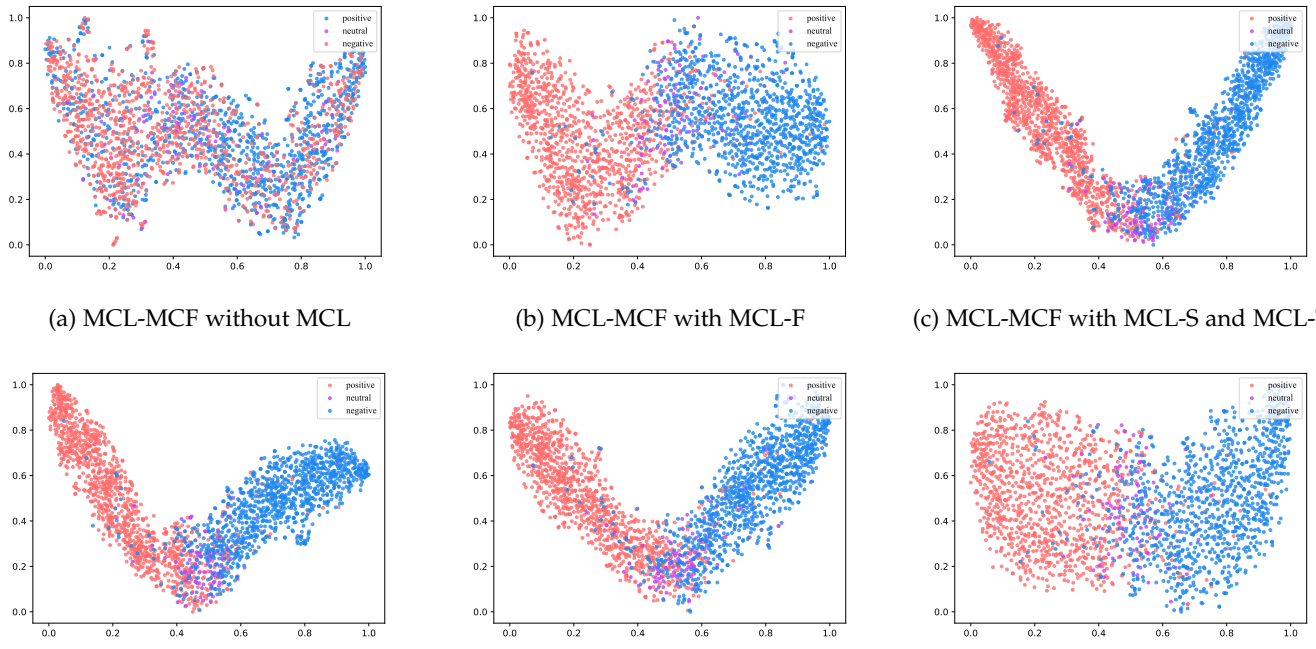
Description	MAE ↓	Corr ↑	Acc-2 ↑	F1 ↑
MCL-MCF	**0.692**	**0.799**	**0.849 / 0.873**	**0.847 / 0.872**
w/o MCL-F	0.720	0.782	0.836 / 0.859	0.834 / 0.858
w/o MCL-S	0.727	0.783	0.8418 / 0.862	0.839 / 0.862
w/o MCL-T	0.721	0.783	0.838 / 0.8581	0.836 / 0.857
w/o MCL	0.727	0.791	0.833 / 0.855	0.832 / 0.854
w/o MCF	0.717	0.793	0.841 / 0.862	0.838 / 0.861

5.5 Ablation Study

To substantiate the effectiveness and usefulness of MCL-MCF, we conducted a series of ablation experiments and random seed sampling experiments with the CMU-MOSI dataset. The results from different ablation configurations are presented in Table 5.

Initially, we performed hierarchical ablation experiments on multi-level contrastive learning. Table 5 shows the first-level, second-level, and third-level contrastive learning results. The experimental results were lower than those without ablation experiments, indicating that single-level contrastive learning is not enough to alleviate the heterogeneity differences between different layers of multiple modalities. This proves that multi-level contrastive learning methods complement each other, and contrastive learning works simultaneously at each level to alleviate the heterogeneity between multimodal fusion features at different levels. After completely ablating multi-level contrastive learning, only the feature extraction and convolution fusion modules remained in the model. The experimental results indicate that the fusion method does not have the ability to alleviate the heterogeneity difference between multimodal features, and the fusion ability of multi-layer convolution without the help of multi-level contrastive learning is hindered by the heterogeneity between multimodal features. Under the action of multi-level contrastive learning, the heterogeneity is greatly reduced to promote multi-layer convolution fusion. To validate the usefulness of the model, we randomly sampled 10 seeds and calculated the variance for each metric in the ablation experiments. The small fluctuations in the variance of each metric indicate the effectiveness of the ablation experiments and demonstrate that the model exhibits good stability and usefulness.

Furthermore, we conducted an ablation experiment pertaining to the multi-layer convolution fusion module for multimodal fusion, where linear layers were utilized instead of multi-layer convolution. The experimental outcomes indicate that multi-level contrastive learning can alleviate heterogeneity among multimodal features, thereby assisting the fusion technique in achieving commendable results. This further underscores that even in the absence of intricate fusion methodologies, as long as heterogeneity among multimodal features is reduced, the model can achieve satisfactory outcomes. Thus, multi-level contrastive learning serves as the foundation of MCL-MCF,



(a) MCL-MCF without MCL (b) MCL-MCF with MCL-F (c) MCL-MCF with MCL-S and MCL-T

(d) MCL-MCF with MCL-F and MCL-T (e) MCL-MCF with MCL-F and MCL-S (f) MCL-MCF with MCL

Fig. 3: t-SNE [67]visualization of multimodal representation in the embedding space on CMU-MOSI. a, b, c, d, e, and f correspond to the visualizations of ablation experiments on the MCL.

with optimal experimental results stemming from the amalgamation of multi-level contrastive learning and multimodal fusion through multi-layer convolution.

5.6 Visualization

To verify whether multi-level contrastive learning can significantly enhance the effectiveness of multimodal fusion, we conducted visualization experiments with the CMU-MOSI dataset. By visualizing the data feature vectors of the last layer of the model through dimensionality reduction, we used the t-SNE [67] algorithm to obtain 2D features for visualization. As depicted in 3, all the plots feature the MCF module, and they showcase the visualization outcomes of level-by-level ablation experiments conducted on multi-level contrastive learning. (a) corresponds to the absence of any contrastive learning at any level, relying solely on MCF for the distribution of features, which results in a relatively disordered feature distribution. (b) demonstrates the inclusion of first-level contrastive learning atop (a), leading to a considerably clearer feature distribution. (e) incorporates both first-level and second-level contrastive learning; the feature distribution is akin to that of (b), with features from same-polarity samples clustering together. (f) integrates first-level, second-level, and third-level contrastive learning, revealing a discernible shift in feature distribution compared to (e), where the distribution of features from same-polarity samples becomes more scattered. (c) and (d) show the visualizations of ablation experiments carried out on first-level and second-level contrastive learning, respectively. The distributions of their features show relatively minor deviations from that of (e). Upon analyzing (a), (b), (e), and (f), despite the addition of an extra layer of contrastive learning in (f), its feature distribution does not exhibit a markedly denser pattern compared to that of (e). This observation suggests that the multi-level contrastive learning (MCL) approach not only decreases the distances between same-polarity sample features but also decreases the distances between different-polarity sample features but also has the ability to alleviate heterogeneity among multimodal features. In the context of the analysis of (c), (d), and (e), the division of the fusion levels into early, middle, and late phases appears to be a reasonable approach. Optimal results are achieved when all three levels are active, thereby maximizing the alleviation of heterogeneity among modalities.

5.7 Loss analysis

To investigate the model's loss profile, we present the results of a 200-epoch run, as depicted in Figure 5. In general, all the losses converge well without encountering conflicts, indicating the integrity of the model design. The trends of label loss and overall loss are similar, rapidly reaching convergence, suggesting that alleviating heterogeneity in the multimodal fusion process is immensely beneficial for multimodal integration. After the convergence of the label loss and overall loss, the other three losses continue to decrease. However, the model's accuracy does not improve with the reduction of these losses. This is attributed to the excessive mitigation of heterogeneity leading to the loss of private information in each modality, which plays a crucial role in emotion analysis recognition. Therefore, in the process of multimodal fusion, we opt for shallow-

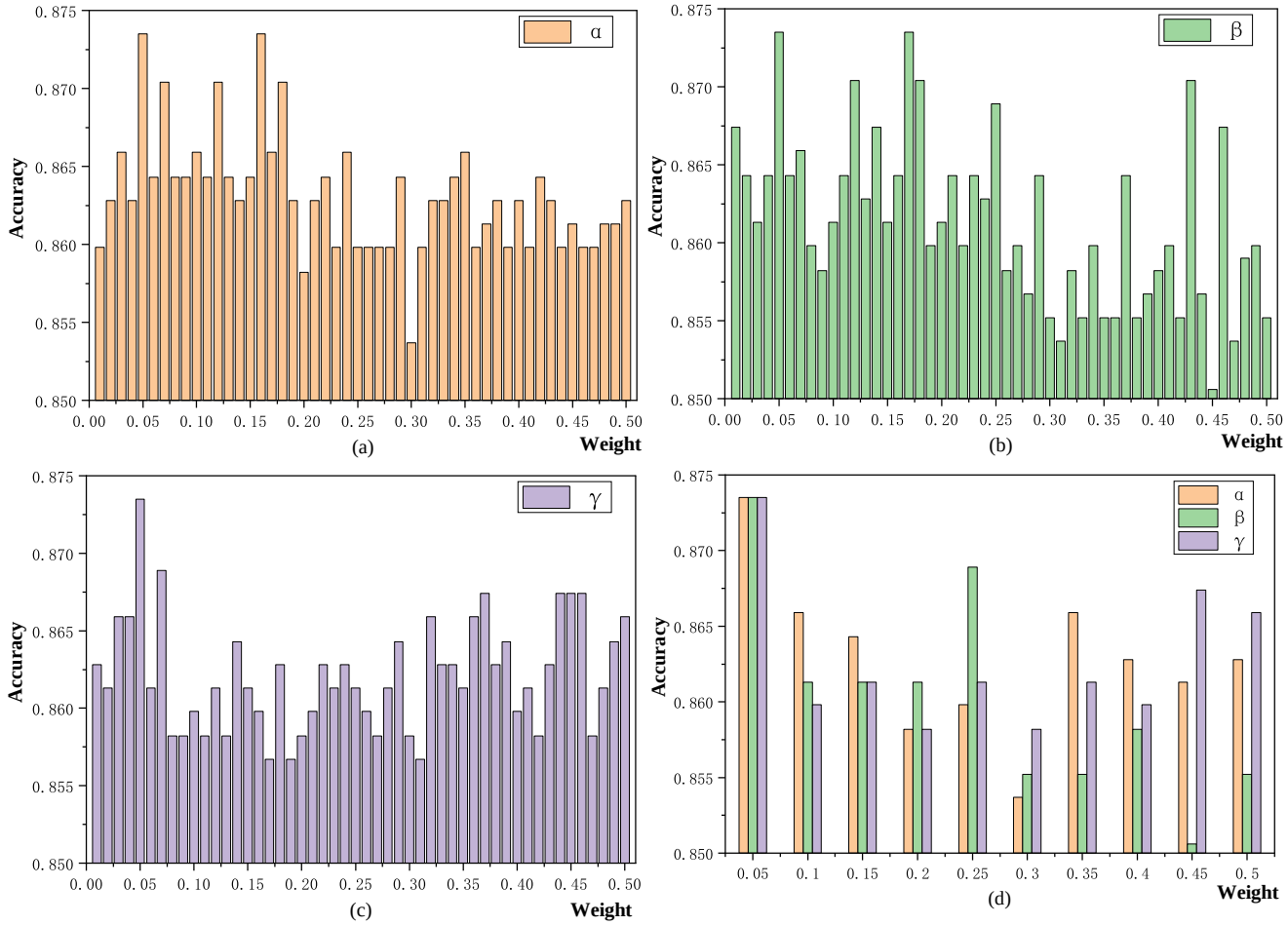


Fig. 4: Visualization of the impact of weight changes in multi-level contrastive learning on accuracy, where α, β, and γ are the weights for first-level, second-level, and third-level contrastive learning, respectively. (a) shows the variation in α when β and γ are held constant at 0.05. (b) illustrates the change in β when α and γ remain fixed at 0.05. (c) shows the fluctuation of γ with α and β fixed at 0.05. (d) involves extracting some data from (a), (b), and (c).

level features to facilitate multimodal integration, thereby preventing the loss of private information.

6 Further Analysis

6.1 Case Study

Four samples were randomly selected from CMU-MOSI, MCL-MCF and MMIM to predict these four samples (MMIM was trained to achieve the best performance according to the open source code), and the results are shown in Table 6. From the four samples in the table, it can be seen that MCL-MCF is better than the current optimal MMIM model in the prediction of positive samples, negative samples, and neutral samples.

6.2 Weight Analysis

For further analysis of the MCL module, Figure 4 provides an investigation into the variations in the weights α, β, and γ. As their respective weights increase, noticeable fluctuations are observed in (a), (b), and (c), all exhibiting an overall decreasing trend. Unilateral improvements in α, β, and γ disrupt the balance of constraints and lead to a decrease in model performance. As seen in (d), α, β, and γ collectively attain their optimum when equal and set at 0.05, with subsequent variations failing to surpass this outcome. The insights from both Figure 3 and Figure 4 affirm that each layer of multi-level contrastive learning is indispensable. Balancing the weights of hierarchical contrastive learning is also essential.

6.3 Positive and Negative Pair Analysis

To further substantiate the validity of our framework design, we devised three distinct configurations for positive and negative sample pairs in the context of contrastive learning, as delineated in Table 7. Herein, A, B, and C represent three respective types, with comprehensive explanations of their nuances provided in Table 7. Our experiments encompassed the exploration of these three distinct design

TABLE 6: In MMIM and MCL-MCF, the left of the "/" is predict value,the right is truth value .

Text	Visual	Acoustic	MMIM	MCL-MCF
(A):Its completely different from anything we've ever seen him do before	Smile	Slightly rising tone	-0.0637/0.4000	0.4004/0.4000
(B):And she was kind of wierd during twilight	Look up Turn head	Peaceful tone Slight pause	-0.6119/-1.0000	-1.0004/-1.0000
(C)And I was pretty open to to it being good	Raise eyebrows Turn head	Slightly rising tone Slight pause	1.4106/0.6000	0.5996/0.6000
(D):Because I sure didn't see the end coming	Glance	Peaceful tone	-0.1139/0.0000	0.0002/0.0000

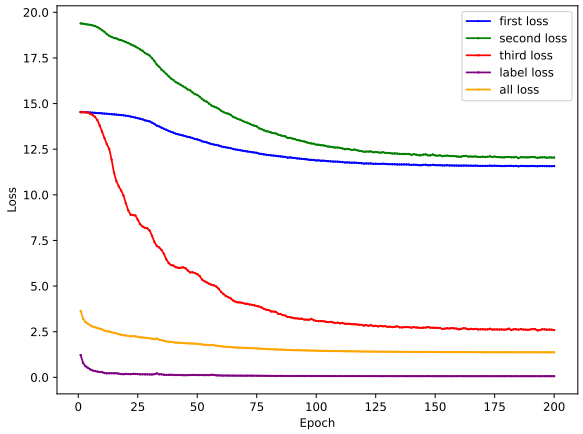


Fig. 5: The first loss, second loss, and third loss correspond to the contrastive learning losses at the first, second, and third level, respectively. The label loss represents the loss associated with labels, while all losses encompass the overall loss of the entire model.

TABLE 7: Positive sample pairs with different modalities within the same sample (PSA), positive sample pairs with the same sentiment across different samples (PSB), negative sample pairs with the same sentiment (NSA), negative sample pairs without the same sentiment (NSB). A, B, and C represent three different designs for positive and negative sample pairs, respectively.

	A	B	C
PSA	✓	✓	×
PSB	×	×	✓
NSA	✓	×	×
NSB	×	✓	✓

approaches, and the ensuing results are presented in Table 8. Analysis of the results (A, B, C) reveals that employing different modalities within the same sample as positive pairs in contrastive learning yields superior outcomes. This is attributed to their intrinsic suitability as positive pairs, a phenomenon substantiated by prior research demonstrating their efficacy [40]. [42] Conversely, employing samples with identical emotional expressions as positive pairs manifests marginally diminished results. This discrepancy may arise from the substantial inherent differences among samples, despite sharing identical emotional attributes. The commendable outcomes observed across all three ABC experiments underscore the overarching efficacy and rationale behind our integrated design.

TABLE 8: Experimental results of three different positive and negative sample pairs on CMU-MOSI.

	MAE ↓	Corr ↑	F1 ↑	Acc-2 ↑
A	0.692	0.799	84.7 / 87.2	84.9 / 87.3
B	0.728	0.798	84.4 / 86.5	84.6 / 86.7
C	0.769	0.790	83.8 / 86.2	83.9 / 86.2

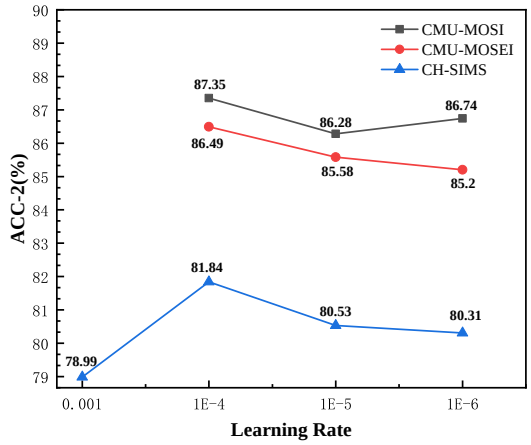
6.4 Comparative Analysis

Four graphs are shown in Figure 6: (a) and (b) are the accuracy distributions on three public datasets, while (c) and (d) are the ACC-2, Corr, and MAE distributions under different conditions on CMU-MOSI.

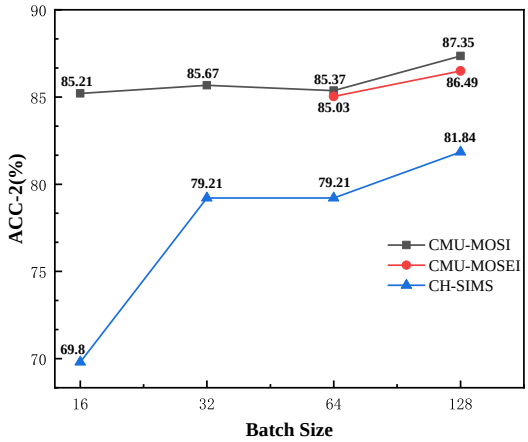
In (a), we can see that the accuracy rate of CMU-MOSI and CMU-MOSEI is NaN under the condition of (a) a learning rate of 1e-3, and the learning rate of 1e-3 in (c) results in NaN values not only for accuracy but also for Corr and MAE. This shows that the MCL-MCF model is sensitive to the learning rate. In contrast, in (b), only the CMU-MOSEI dataset has an accuracy of NaN when the batch size is 8 or 16, while the other two datasets are normal. The MCL-MCF model is also sensitive to the batch size. Under the condition of using multi-level contrastive learning, the model should be given a larger batch size to provide more negative samples to avoid model collapse. As the batch size increases in (b), the accuracy of the MCL-MCF model on the three datasets continues to increase. From (d), we can see that as the batch size increases, ACC-2 and Corr increase, while MAE decreases. This is consistent with the notion that the more negative samples given to the contrastive learning model, the better it is. It also shows that the multi-level contrastive learning module in the MCL-MCF model is effective in promoting multimodal fusion. Due to hardware limitations, a larger batch size could not be tested. The four graphs also show that the limitation of MCL-MCF is that it requires considerable memory and is very sensitive to the learning rate hyperparameters.

6.5 Accuracy Analysis

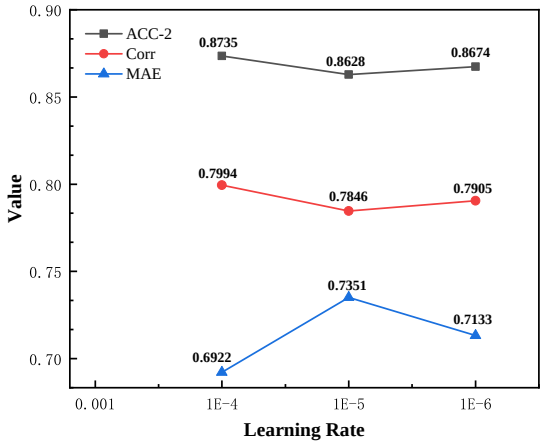
To investigate the generalizability and sensitivity of the model to different samples, we visualized the confusion matrices of the MCL-MCF model in three datasets (CMU-MOSI, CMU-MOSEI, and CH-SIMS) for three-class classifi-



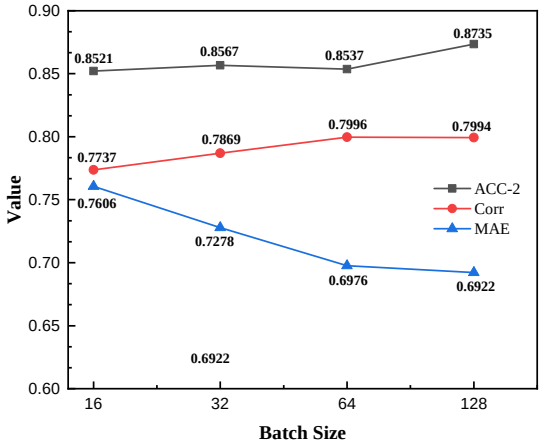
(a) Accuracy distribution of three public datasets under different learning rate conditions.



(b) Accuracy distribution of three public datasets under different batch size conditions.



(c) The distributions of ACC-2, Corr, and MAE under different learning rate conditions with the CMU-MOSI dataset.



(d) The distribution of ACC-2, Corr, and MAE under different batch size conditions with the CMU-MOSI dataset.

Fig. 6: (a) and (b) show the accuracy of different datasets under the same batch size and learning rate conditions; the other conditions are the same. (c) and (d) are the learning rates and batch sizes with the CMU-MOSI dataset, respectively, and the ACC-2, Corr, MAE, and other conditions are the same. In (a), the accuracy values of CMU-MOSI and CMU-MOSEI under the condition of 1e-3 learning rate are Nan. In (b), the accuracy rate of CMU-MOSEI under the condition of an 8,16 batch size is Nan. In (c), the values of Acc, Corr and the MAE of CMU-MOSI under the condition of 1e-3 learning rate is Nan.

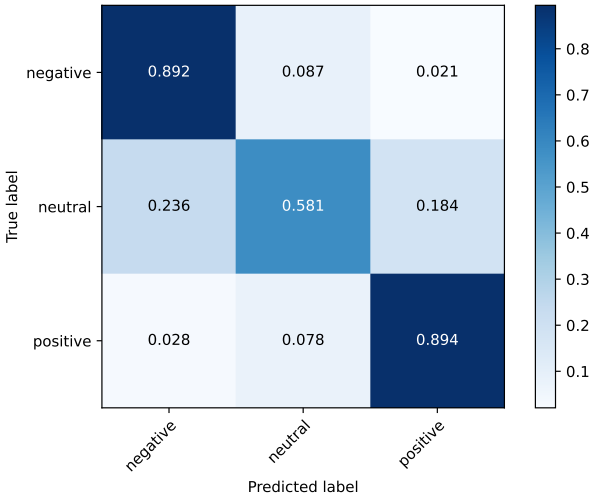
cation, as shown in Figure 7 (a, c, d). To facilitate comparison, we also visualize the three-class confusion matrix of the MMIM [39] model on CMU-MOSI, as shown in Figure 7 (b).

By comparing (a) and (b), we can conclude that the MCL-MCF model outperforms the MMIM model on three types of samples. It also performs well in processing positive and negative samples in the CMU-MOSI dataset but not neutral samples. From (a) and (c), we can infer that MCL-MCF may not handle neutral samples from CMU-MOSI well because there are too few neutral samples in the dataset, which makes it difficult for the model to be sufficiently trained. From (a), (c), and (d), we can see that MCL-MCF has good generalizability, as it performs well from the small CMU-MOSI dataset to the large CMU-MOSEI dataset and from the English CMU-MOSI dataset to the Chinese CH-SIMS dataset.

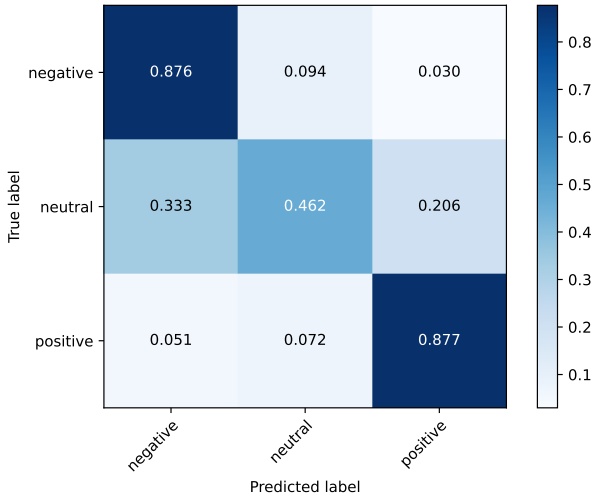
Overall, multi-level contrastive learning can handle the heterogeneity between multimodal data features well, thereby helping the model achieve excellent results and good generalizability.

6.6 Multimodal "image" Analysis

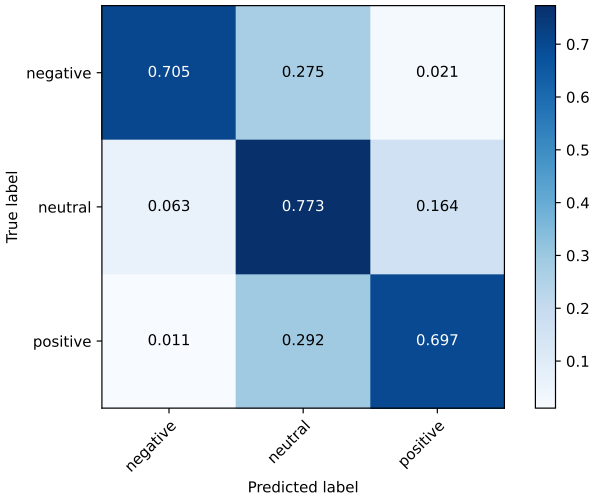
To validate the efficacy of two-dimensional convolution applied to multimodal "images" generated by the outer product of two modalities and whether subsequent two-dimensional convolutional feature extraction can yield a



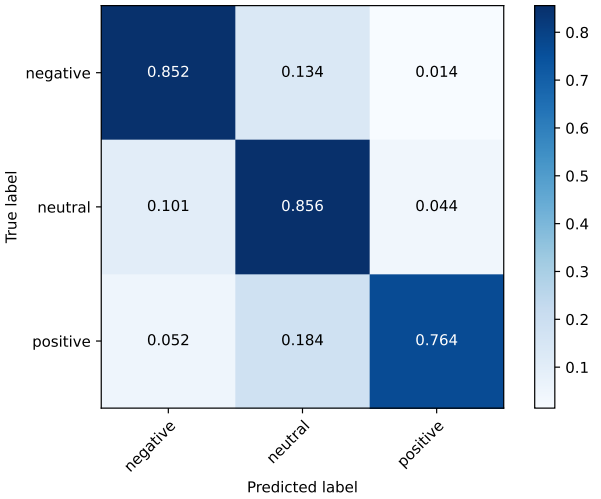
(a) Three-category confusion matrix visualization of MCL-MCF on CMU-MOSI.



(b) Three-category confusion matrix visualization of MMIM on CMU-MOSI.



(c) Three-category confusion matrix visualization of MCL-MCF on CMU-MOSEI



(d) Three-category confusion matrix visualization of MCL-MCF on CH-SIMS

Fig. 7: (a,c,d) represent the three-class visualization confusion matrix for MCL-MCF, and b represents the three-class visualization confusion matrix for MMIM. (a,c,d) compare the performance and generalizability of MCL-MCF on three datasets, while (a,d) compare the sensitivity of MCL-MCF and MMIM to sample categories.

robust fusion feature representation, we examined scenarios involving both single-channel and multichannel processing. We obtained three one-channel multimodal "images" from the outer products of h_t, h_a, and h_v, which are the same as the inputs of the contrastive learning module.

We stack the three one-channel multimodal "image" to generate a three-channel multimodal "image", which were then directly sent to ResNet18, ResNet34, ResNet50, a vision transformer (Vit) [68] and masked autoencoders (MAE) [69], but three single-layer convolutions, which use the three single-channel convolutions in this article. The objective was to assess the validity and effectiveness of convolution on multimodal "image". Table 9 presents the results of this analysis. ResNet [70] series models consistently outperform previous models across multiple metrics, confirming the strength of employing three-channel multimodal "images" for two-dimensional convolution operations. The three single-layer convolutions also exhibited promising outcomes, supporting the feasibility and effectiveness of applying two-dimensional convolution to single-channel multimodal "images". The ResNet18, ResNet34, ResNet50 results and the three single-layer convolutions underscored the effectiveness of convolution in extracting local features from the multimodal "image". However, importantly, deeper convolutional layers did not necessarily correlate with superior performance. From the Vit and MAE results, we make two hypotheses: 1) the three-channel multimodal "image" can contain redundant and conflicting information unsuitable for direct global feature extraction, and 2) the timing of the three-channel multimodal "image" obtained from the outer product may be damaged to a certain extent, making it unsuitable as a model input for the transformer or Vit architecture.

Multimodal "images" themselves are obtained through

TABLE 9: MAE* means masked autoencoders instead of evaluation indicators; Results of ResNet18, ResNet34, ResNet50, vision transformer (Vit) [68] and masked autoencoders (MAE) [69] on CMU-MOSI, all of which input three-channel multimodal "image"; TSLC means that three single-channel multimodal "image" is calculated by 1-channel two-dimensional convolution. TSLC is short for Three single layer convolutions; In Acc-2 and F1-Score, the left of the "/" is calculated as negative/non-negative and the right is calculated as negative/positive, TPs represents trainable parameters.

ModelsΔ	MAE$\downarrow$	Corr$\uparrow$	Acc-2$\uparrow$	F1$\uparrow$	TFLOPs/M	TPs/M
ResNet18	0.834	0.773	**84.84 / 86.59**	**84.71 / 86.51**	596.08	**11.82**
ResNet34	0.765	0.789	84.69 / 86.43	84.65 / 86.44	1201.69	21.93
ResNet50	0.777	0.776	83.82 / 86.13	83.64 / 86.03	1351.30	25.69
Vit	0.825	0.763	82.94 / 85.52	82.67 / 85.35	11921.54	184.34
MAE*	1.125	0.491	72.16 / 75.00	71.26 / 74.37	20331.97	329.34
TSLC	**0.735**	**0.794**	83.38 / 86.13	83.08 / 85.94	**26.52**	24.77

the interaction of single-modal fine-grained details containing rich interactive information. However, some interactions between elements are meaningless, leading to considerable redundancy similar to that of visual images. Tremendous success has already been achieved with local extraction through convolution in the image domain. Through experiments, it has been found that performing local information extraction on multimodal "images" is also very effective. Although significant success has been achieved with transformer architecture models in the image domain, our experimental results suggest that multimodal "image" is not well suited for direct input into models based on the transformer architecture. The reason is that multimodal "images" obtained from different modalities contain more redundancy and conflicting emotional information than visual images. These conflicts can affect the model's analysis of emotions. Therefore, local convolutional networks are more suitable for multimodal "images". In the future, a combination of both approaches can be explored to achieve a fusion of local and global features. In the experiments with multimodal "image", we utilized the built-in ResNet series models in PyTorch. The ResNet series models have an input size of 128×128×3. For ViT, the input size is also 128×128×3, with a patch size of 16×16 and 6 layers. MAE is a pre-trained MAE-based model with an input size of 224×224×3 and a patch size of 16×16. To further contrast the performance of global and local models, we established a standardized encoding environment, concentrating exclusively on the computation metrics following the encoding process. A comparison of the accuracy results of the global models (VIT and MAE) and local models (ResNets and TSLC) indicates that multimodal "images" are more redundant. Global models have a significantly larger parameter count than local models, which paradoxically exacerbates the introduction of redundant noise, leading to a decrease in the results. This suggests that local feature aggregation is more effective and highlights the rationale behind TCF design.

Multimodal "images" overcome the one-dimensional limitation of previous multimodal fusion methods, improve multimodal fusion to a two-dimensional level, and enable the direct application of excellent models from the image field, greatly expanding the range of multimodal fusion methods. We believe this work can inspire creativity in the field of multimodal learning and MSA in the future.

7 Conclusion

This paper introduces the MCL-MCF framework as a solution to mitigate the heterogeneity among multimodal features during the process of multimodal feature fusion, aiming to enhance fusion effects. Fusion is considered a gradual process, with MCL implementing hierarchical processing to mitigate heterogeneity across various feature levels. In the process of obtaining advanced features, we designed the tensor convolution fusion (TCF) module and found through experiments that performing two-dimensional convolution operations conducted on multimodal "image" can yield excellent advanced features. To simulate fusion as a progressive process, the MCF is designed to employ hierarchical fusion. We found through our experiments that the two are complementary and thus can be combined for better performance. Ultimately, our experiments with three datasets show that the method achieves state-of-the-art performance across the English and Chinese languages, as well as with both small and large datasets. The analysis of the visualization experiments shows that our model alleviates the heterogeneity among multimodal features, improves the fusion effect of fusion methods, and has good generalizability.

In future work, we intend to explore the application of MCL-MCF in different multimodal learning methods. The multimodal "images" overcome the limitations of one-dimensional and two-dimensional methods. In the future, we will further study multimodal "images". While high-level features extracted from multimodal "images" are effective for contrastive learning, they consume considerable memory, and the model becomes highly sensitive to the learning rate. The MAE and Vit do not handle multimodal "images" well. Discovering how to solve these problems remains an attractive direction for future research.

References

[1] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

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

[3] K. Somandepalli, T. Guha, V. R. Martinez, N. Kumar, H. Adam, and S. Narayanan, "Computational media intelligence: human-centered machine analysis of media," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 891–910, 2021.

[4] L. Stappen, A. Baird, L. Schumann, and S. Bjorn, "The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements," *IEEE Transactions on Affective Computing*, 2021.

[5] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," *A practical guide to sentiment analysis*, pp. 1–10, 2017.

[6] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.

[7] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[8] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *Ieee Access*, vol. 7, pp. 63 373–63 394, 2019.

[9] Y. Peng and J. Qi, "Cm-gans: Cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1, pp. 1–24, 2019.

[10] J. Tang, D. Liu, X. Jin, Y. Peng, Q. Zhao, Y. Ding, and W. Kong, "Bafn: Bi-direction attention based fusion network for multimodal sentiment analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1966–1978, 2022.

[11] R. Chen, W. Zhou, Y. Li, and H. Zhou, "Video-based cross-modal auxiliary network for multimodal sentiment analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 12, pp. 8703–8716, 2022.

[12] L. He, Z. Wang, L. Wang, and F. Li, "Multimodal mutual attention-based sentiment analysis framework adapted to complicated contexts," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[13] C. Fan, J. Wang, W. Huang, X. Yang, G. Pei, T. Li, and Z. Lv, "Light-weight residual convolution-based capsule network for eeg emotion recognition," *Advanced Engineering Informatics*, vol. 61, p. 102522, 2024.

[14] C. Fan, H. Xie, J. Tao, Y. Li, G. Pei, T. Li, and Z. Lv, "Icaps-reslstm: Improved capsule network and residual lstm for eeg emotion recognition," *Biomedical Signal Processing and Control*, vol. 87, p. 105422, 2024.

[15] 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*, 2017, pp. 1103–1114.

[16] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2020. NIH Public Access, 2020, p. 2359.

[17] T. Zhu, L. Li, J. Yang, S. Zhao, H. Liu, and J. Qian, "Multimodal sentiment analysis with image-text interaction network," *IEEE Transactions on Multimedia*, 2022.

[18] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2022.

[19] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang, "Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning," *IEEE Access*, vol. 11, pp. 14 742–14 751, 2023.

[20] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "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*, 2021, pp. 2554–2562.

[21] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L.-N. Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 4730–4738.

[22] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 251–260.

[23] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. B. 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 (Volume 1: Long Papers)*, 2018, pp. 2247–2256.

[24] S. Verma, J. Wang, Z. Ge, R. Shen, F. Jin, Y. Wang, F. Chen, and W. Liu, "Deep-hoseq: Deep higher order sequence fusion for multimodal sentiment analysis," in *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2020, pp. 561–570.

[25] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*. Association for Computational Linguistics (ACL), 2018, pp. 2225–2235.

[26] M. S. Akhtar, D. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 370–379.

[27] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, vol. 2019, 2019, pp. 6558–6569.

[28] W. Han, H. Chen, A. Gelbukh, A. Zadeh, L.-p. Morency, and S. Poria, "Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis," in *Proceedings of the 2021 International Conference on Multimodal Interaction*, 2021, pp. 6–15.

[29] S. Qian, D. Xue, Q. Fang, and C. Xu, "Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[30] J. Yang, Y. Wang, R. Yi, Y. Zhu, A. Rehman, A. Zadeh, S. Poria, and L.-P. Morency, "Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 1009–1021.

[31] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 439–448.

[32] 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 AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 6892–6899.

[33] 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 AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 164–172.

[34] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of The Web Conference 2020*, 2020, pp. 2514–2520.

[35] Y.-H. H. Tsai, M. Q. Ma, M. Yang, R. Salakhutdinov, and L.-P. Morency, "Multimodal routing: Improving local and global interpretability of multimodal language analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2020. NIH Public Access, 2020, p. 1823.

[36] P. P. Liang, Y. Cheng, X. Fan, C. K. Ling, S. Nie, R. Chen, Z. Deng, N. Allen, R. Auerbach, F. Mahmood *et al.*, "Quantifying & modeling multimodal interactions: An information decomposition framework," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[37] Y. Zeng, W. Yan, S. Mai, and H. Hu, "Disentanglement translation network for multimodal sentiment analysis," *Information Fusion*, vol. 102, p. 102031, 2024.

[38] Y. Sun, S. Mai, and H. Hu, "Learning to learn better unimodal representations via adaptive multimodal meta-learning," *IEEE Transactions on Affective Computing*, 2022.

[39] W. Han, H. Chen, and S. Poria, "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*, 2021, pp. 9180–9192.

[40] R. Lin and H. Hu, "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*, 2022, pp. 511–523.

[41] Z. Li, B. Xu, C. Zhu, and T. Zhao, "Clmlf: A contrastive learning and multi-layer fusion method for multimodal sentiment detection," in *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022, pp. 2282–2294.
[42] S. Mai, Y. Zeng, S. Zheng, and H. Hu, "Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, no. 01, pp. 1–1, 2022.
[43] P. P. Liang, Z. Deng, M. Q. Ma, J. Y. Zou, L.-P. Morency, and R. Salakhutdinov, "Factorized contrastive learning: Going beyond multi-view redundancy," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[44] A. Wilf, M. Q. Ma, P. P. Liang, A. Zadeh, and L.-P. Morency, "Face-to-face contrastive learning for social intelligence question-answering," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–7.
[45] S. Becker and G. E. Hinton, "Self-organizing neural network that discovers surfaces in random-dot stereograms," *Nature*, vol. 355, no. 6356, pp. 161–163, 1992.
[46] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.
[47] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.
[48] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[49] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
[50] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
[51] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.
[52] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.
[53] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, pp. 4171–4186.
[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[55] R. Lin and H. Hu, "Dynamically shifting multimodal representations via hybrid-modal attention for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, 2023.
[56] P. Poklukar, M. Vasco, H. Yin, F. S. Melo, A. Paiva, and D. Kragic, "Geometric multimodal contrastive representation learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 782–17 800.
[57] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1122–1131.
[58] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1642–1651.
[59] Y. Li, Y. Wang, and Z. Cui, "Decoupled multimodal distilling for emotion recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6631–6640.
[60] R. Lin and H. Hu, "Multi-task momentum distillation for multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, 2023.
[61] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.
[62] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
[63] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 3718–3727.
[64] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *International Conference on Representation Learning*, 2019.
[65] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
[66] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.
[67] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.
[68] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.
[69] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.
[70] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Cunhang Fan received the PhD degree from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2021, and a BS degree from the Beijing University of Chemical Technology (BUCT), Beijing, China, in 2016. He is currently an associate professor with the School of Computer Science and Technology, Anhui University, Heifei, China. His current research interests include affective computing, speech enhancement, speech recognition and speech processing.

Kang Zhu received a BS degree from Anhui University of Technology (AHUT), Maanshan, China, in 2022. He is currently a first-year graduate student with the School of Computer Science and Technology, Anhui University, Heifei, China. His current research interests include multimodal fusion and affective computing.

Jianhua Tao (Senior Member, IEEE) received an MS degree from Nanjing University, Nanjing, China, in 1996 and a PhD degree from Tsinghua University, Beijing, China, in 2001. He is currently a Professor in the Department of Automation, Tsinghua University, Beijing, China. He has authored or coauthored more than 300 papers in major journals and proceedings, including the IEEE TASLP, IEEE TAFFC, IEEE TIP, IEEE TSMCB, and Information Fusion. His current research interests include speech recognition and synthesis, affective computing, and pattern recognition. He is the Board Member of ISCA, the chairperson of ISCA SIG-CSLP, and the Chair or Program Committee Member for several major conferences, including Interspeech, ICPR, ACII, ICMI, ISCSLP, etc. He was the subject editor for the Speech Communication and is an Associate Editor for Journal on Multimodal User Interface and International Journal on Synthetic Emotions. He was the recipient of several awards from important conferences, including Interspeech and NCMMSC.

Guofeng Yi received BE degree from the School of Computer and Data Engineering, NingboTech University, Ningbo, China, in 2020. He is currently studying for a master's degree at the School of Computer Science and Technology, Anhui University, Hefei, China. His current research interests include multimodal fusion and affective computing.

Jun Xue received a BS degree from Anhui University of Science and Technology (AUST), Bengbu, China, in 2020. He is currently a first-year graduate student with the School of Computer Science and Technology, Anhui University, Heifei, China. His current research interests include fake speech detection and speech signal processing.

Zhao Lv received his PhD degree in Computer Application Technology from Anhui University, Hefei, China, in 2011. He was a visiting scholar at the University of Utah, Salt Lake City, USA, from 2017 to 2018. He is currently a professor (master and doctoral supervisor) in the School of Computer Science and Technology at Anhui University, Hefei, China. His research interests include the broad areas of biomedical signal processing and brain-computer interaction. He has published over 70 scientific articles in prestigious IEEE/Elsevier journals such as IEEE TIM, IEEE TNSRE, and JBHI. He is serving as a Guest Associate Editor for Frontiers in Neuroscience and an Editorial Board Member for Frontiers in Human Neuroscience. He is also the reviewer for more than 20 prestigious IEEE/Elsevier/Springer journals. He received support from the Anhui Province Science Fund for Distinguished Young Scholars in 2022.