Inter-Intra Modal Representation Augmentation With Trimodal Collaborative Disentanglement Network for Multimodal Sentiment Analysis

Chen Chen, Hansheng Hong, Jie Guo, *Member, IEEE*, and Bin Song, *Senior Member, IEEE*

***Abstract*—Recently, Multimodal Sentiment Analysis (MSA) is a challenging research area given its complex nature, and humans express emotional cues across various modalities such as language, facial expressions, and speech. Representation and fusion of features are the most crucial tasks in multimodal sentiment analysis research. However, in the current research, most methods ignore the importance of eliminating potential irrelevant features in the original features of each modality and cross-modal common feature. Moreover, the features extracted from all the modalities contain cluttered background noise and different occlusions noise, which negatively affects feature alignment. Different from these methods, we propose a novel Trimodal Collaborative Disentanglement Network (TCDN) to solve these problems in this paper. This work can obtain effective sentiment results on two aspects: i) Trimodal collaborative uses L1-norm to eliminate irrelevant features and unify the characteristics of the three modals (inter-modal). ii) Disentanglement network introduces an adversary noise by combining the original features of various single modalities and the common representation, alleviating the background noises within each modality (intra-modal). This inter-intra modal feature augmentation method is the first work to obtain the common representation by implementing data augmentation as far as we know. Extensive experiments are completed on two benchmark datasets, including MOSI and MOSEI, demonstrating the superiority of the TCDN model over the state-of-the-art methods.**

***Index Terms*—Multimodal sentiment analysis, multimodal fusion, transformers, data augmentation.**

I. Introduction

IN THE Internet age of information explosion, intensive research interests have been paid for linking vision, language and audio to explore the relationship between them [20], [21], [22], [24], [25], [26], [31], [32]. These two forms have been effectively combined to various tasks, such as multimodal pre-training [19] and multimodal emotion recognition [12]. Multimodal Sentiment Analysis (MSA) is one of the hottest tasks, aiming to use manually aligned complete information, including audio, vision and text to judge speaker sentiments. An example is shown in Fig. 1 to simply describe the process of sentiment analysis. Due to the huge semantic difference between the three modalities, the main challenge of MSA is how to learn the joint representation of the three modalities and obtain an accurate sentiment score.

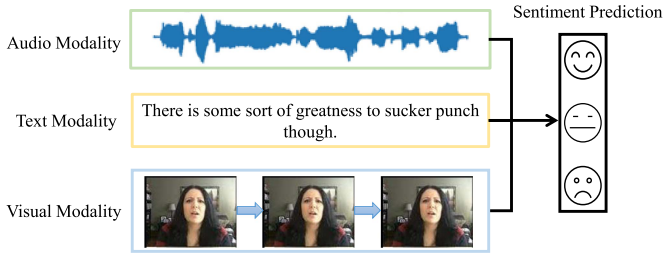


Fig. 1. An example of predicting the user's sentiment through the audio, text and visual modality. The model can effectively predict the overall emotional situation of people during this period by comprehensively considering the performance of the three modals.

To cope with this challenge, the entire original audio, vision and text are mapped into a unimodal representation for most of the early methods [30]. These methods are carefully designed and specialized for the feature extraction of each modality, and are often improved by using pre-trained models obtained from large datasets, e.g., MFCC [28] for audio and BERT encoding for text [6]. However, in recent years, it has been found that the quality of information fusion is the main factor affecting the accuracy of MSA task [18], [45]. These fusion methods mainly include simple operation-based [17], tensor-based [41] and translation-based [13], [16], [34]. The goal of fusion is to learn a modal-invariant embedded space, and then use modal invariant features or combine modal invariant features with modal-specific features to achieve effective prediction.

Although significant improvements have been made through current mechanisms to encode more robust features or capture more accurate comparisons, existing methods only use direct unimodal representations and traditional information fusion for sentiment analysis. Nonetheless, there are still many inter-modal noise and intra-modal noise in the features of video, text, and

Manuscript received 24 June 2022; revised 11 December 2022 and 20 March 2023; accepted 23 March 2023. Date of publication 31 March 2023; date of current version 14 April 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62201419 and 62071354, in part by the Key Research and Development Program of Shaanxi, under Grants 2022ZDLGY05-08 and 2023-YBGY-218, in part by the ISN State Key Laboratory in part by the Fundamental Research Funds for the Central Universities, and in part by the Innovation Fund of Xidian University. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Isabel Barbancho. *(Corresponding author: Bin Song.)*

Chen Chen, Jie Guo, and Bin Song are with the State Key Laboratory of Integrated Services Networks, Xidian University, Xian 710071, China (e-mail: chenchen_123@stu.xidian.edu.cn; jguo@xidian.edu.cn; bsong@mail.xidian.edu.cn).

Hansheng Hong is with the Guangdong OPPO Mobile Telecommunications Corp., Dongguan 523860, China (e-mail: honghansheng@oppo.com).

Digital Object Identifier 10.1109/TASLP.2023.3263801

audio. These noisy features can negatively impact the final accuracy. Previous interaction-based methods directly extract features from a given video, sentence, and audio by using pretrain model [6], [23], [36]. The matching process is often affected by inter-modal noise (features between modalities cannot obtain a unified emotional feature representation) and intra-modal noise (video background noise, sentiment-independent features in text, irrelevant noise in speech), resulting in lower emotional scores.

In this paper, we optimize the process of getting the unimodal representations by using the idea of collaborative learning and disentanglement learning to realize data augmentation for multimodal sentiment analysis. By using the idea of collaborative learning, the features between various modalities are continuously aligned. Therefore, the final features have a more unified representation and the emotions represented by the features are more indicative. Then, the method of disentanglement learning is used for our model, and the potential noise inside each modal is eliminated through the fusion features of three modals, which reduces the influence of unnecessary features on the final sentiment analysis.

Specifically, we propose a new approach, Trimodal Collaborative Disentanglement Network (TCDN) in this paper. In TCDN, we propose to implement the denoising of inter-modalities feature through collaborative learning and denoising intra-modalities feature through disentanglement learning. (i) Trimodal collaborative learning is to learn potential mapping across modals by the architecture based on L1-norm. This module can not only pay attention to the relationship between the various modal features in the latent mapping process, but also ensure the uniformity of the feature representation. (ii) Disentanglement network first uses various modal features which get from trimodal collaborative learning to obtain the common representation. Then, we draw lessons from data augmentation by combining original modality features and multi-domain common representations to alleviate the intra-modal background noise. Through the disentanglement feature augmentation method, the network can obtain the representations which only useful for sentiment analysis.

The main contributions of our method can be summarized as follows:

- We effectively reduce the impact of irrelevant noise feature during the extracted features in multimodal sentiment analysis. The method retains the common features that are conducive to classification.
- We propose a Trimodal Collaborative Disentanglement Network (TCDN) to realize data representation augmentation for multimodal sentiment analysis. TCDN weakens the irrelevant inter-modal features through the inter-modal collaborative learning and removes the intra-modal background features by the intra-modal disentanglement learning.
- Our method is the first work to deal inter-intra modal representation with collaborative learning and disentanglement learning to implement data augmentation for multimodal sentiment analysis as far as we know.
- We demonstrate the effectiveness of TCDN on multiple benchmarks, which demonstrates that TCDN is better than the state-of-the-art methods. We also give a comprehensive set of ablation experiments and analysis to help you understand our method.

II. Related Work

A. Multimodal Sentiment Analysis

The goal of MSA is to predict people's emotions from the video, audio and text of the discourse. The models like MFM [30], MAG-BERT [19], and RAVEN [33] could work on aligned multimodal data, which means that there is a clear correspondence between the auditory and visual frames and the words in the text modal. The MSA models were gradually extended to the field of unaligned multi-schema data input, in order to deal with more practical situations. TFN [41] and LMF [10] used tensor-based method to get joint representation for utterances. MulT [30] first utilized cross-modal transformers to deal with the unaligned multimodal data. MISA [4] learn the invariance and specific representation of each modal to improve the fusion process. However, there is no additional processing of lost multimodal data in these models. MCTN [16] used cyclic translation between modals to generate other modals from only one modal. Therefore, this method could learn robust joint representations. Self-MM [40] introduced a novel weight self-adjusting strategy to balance different task loss constraints.

Our work aims at representation learning based on attention fusion structure. Different from previous studies, we jointly learn inter-modal and intra-modal features with collaborative learning and disentanglement learning. Our method learns common information from trimodal and effectively eliminates the influence of background noise on sentiment analysis.

B. Data Augmentation

Data augmentation is a clear regularization used to learn robustness representation and prevent over-fitting depth models by generating additional data. Data augmentation has achieved great success in various fields such as image classification [5], video analysis [2], target detection [8] and Re-ID [47]. Common data augmentation strategies include cropping, color jitter, flipping, adding noise and rotation. Zhang et al. [46] recommended using pairs of convex combinations as enhancement data. In addition, recent works [3] applied generative adversarial network (GAN) to generate enhanced images. Different from these data augmentation methods, our proposed representation disentanglement augmentation method distills object, keywords and key frame feature from video, text and audio original representation by disentanglement augmenting the features. Inspired by these adversarial methods, we disentangle the three modals representation into parts of objects, keywords, and background parts (cluttered background and varying noise). Then we generate the disentanglement augmented features without the background feature.

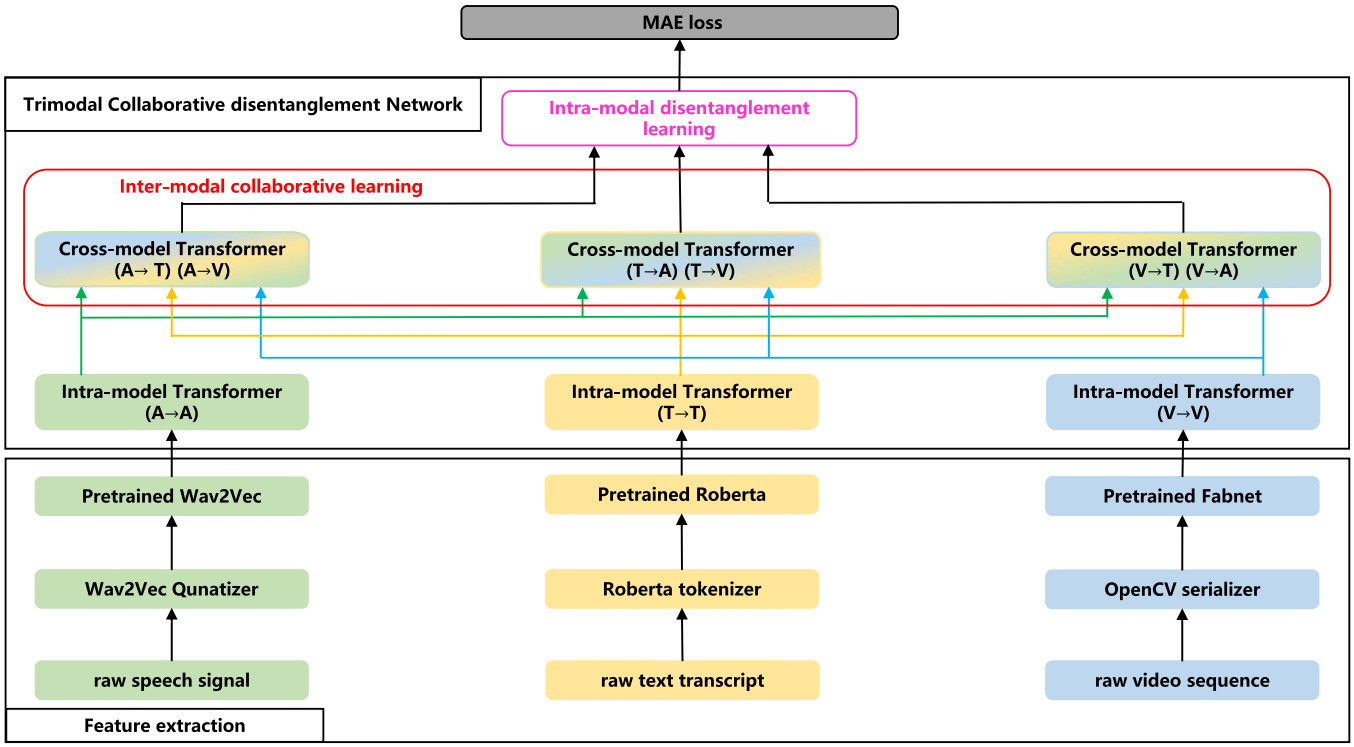


Fig. 2. The overall framework of the TCDN is shown as follows. Firstly, the basic features of every modal are extracted by pre-trained models. Then, the collaborative learning and disentanglement learning modules are used cleverly to denoise the inter- and the intra-modal representation. The collaborative learning weakens the irrelevant inter-modal features and the intra-modal disentanglement learning removes the intra-modal background features. Finally, the collaborative loss and coherence loss constrains the semantic similarities of inter-intra modal embeddings. For more details, the architecture of inter-modal collaborative learning and intra-modal disentanglement learning are shown in Fig. 3.

To the best of our knowledge, the data augmentation method has not been introduced for the multimodal sentiment analysis task to learn robust representation. Different from the existing works, our mechanism is the first to apply the data augmentation method to explore more precise fine-grained features for cross-modal common semantic alignment.

C. Transformer Mechanism

Transformer mainly uses attention mechanism as the basic structure. In natural language processing, the attention mechanism has been used for adaptively filter and aggregate information. It has been intended to discover deeper representation from texts and actively mine the relation between images and texts when it comes to image-text matching. Wu et al. [38] extracted text feature by transformer [6] mechanism and used attention to match latent alignments using words and images. Wei et al. [35] aimed to find the common relationship between images and texts. A loss function is designed to match the internal relations between images and texts by using a transformer structure. In the multimodal pre-train area, many model [7], [11] used improved pre-train vision-language model to obtain the image and text feature. They all mix image and text features into the transformer structure to compute attention-based matching scores sufficiently.

Our work differs from previous transformer matching methods just in the transformer structure. We convert the features to the different modal domains to obtain the multi-domain features effectively and then use the transformer structure. By this way, this structure based on transformer and different input for multi-head attention can better find the features under the most important inverse representation of trimodal common features.

III. The Proposed Method

Figs. 2 and 3 show that the proposed TCDN framework contains two branches that perform efficient data augmentation on intra-modal and inter-modal features, respectively. In this section, we elaborate on the details of these two learning structures.

A. Task Setup

MSA task aims to get the sentiments using multimodal signals by using text (X_t), audio (X_a), and vision (X_v). Generally, MSA can be regarded as either a classification task or a regression task. We regard it as the regression task in this paper. Therefore, TCDN inputs X_t, X_a, and X_v. Then it outputs one sentimental intensity result $\hat{y} \in R$.

B. Feature Extraction

The modality feature extraction module first uses a pre-train layer to processes the modality sequences. Since the great success of the pre-trained language model, we use the pre-trained

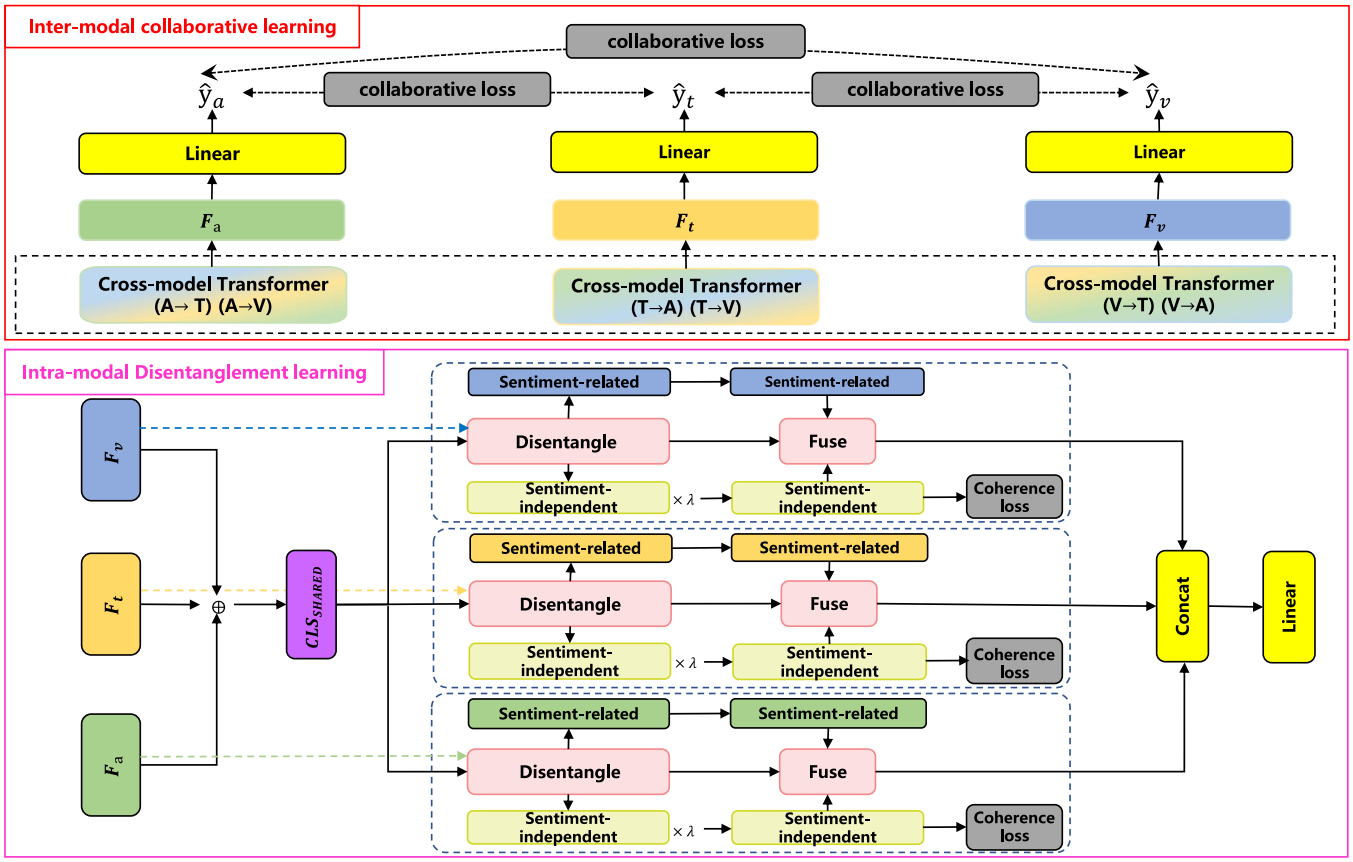


Fig. 3. The overall architecture of the inter-modal collaborative learning and intra-modal disentanglement learning is shown as follows. The inter-modal collaborative learning makes effective use of the relationship between modals and uses L1-norm to shorten the distance between modal features. By calculating the mean value of multimodal features and using the respective modal features, the intra-modal disentanglement learning is used to effectively reduce the influence of unrelated features and noise on sentiment analysis. And through the unique disentanglement enhancement learning method, the feature augmentation in modal is realized.

24-layers RoBERTa [9] to extract sentence representations in the text modality. In the audio modality, we extract the initial vector features from the raw data by pre-trained Wav2vec [23]. In the video modality, the pre-trained Fabnet [36] is used to represent the basic feature of emotions. Then, we put all these original features to an intra-model transformer in order to obtain more flexible embedding.

$$F_{ti} = RoBERTa\left(X_t; \theta_t^{RoBERTa}\right) \in R^{d_t} \quad (1)$$

$$F_{ai} = Wav2vec\left(X_a; \theta_a^{Wav2vec}\right) \in R^{d_a} \quad (2)$$

$$F_{vi} = Fabnet\left(X_v; \theta_v^{Fabnet}\right) \in R^{d_v}, \quad (3)$$

$$F_{t,t} = \textbf{Transformer}\ (F_{ti}, F_{ti}, F_{ti}) \in R^{d_t} \quad (4)$$

$$F_{a,a} = \textbf{Transformer}\ (F_{ai}, F_{ai}, F_{ai}) \in R^{d_a} \quad (5)$$

$$F_{v,v} = \textbf{Transformer}\ (F_{vi}, F_{vi}, F_{vi}) \in R^{d_v} \quad (6)$$

where **Transformer** is the transformer encoder layers. The initial vector are features $X_t \in R^{l_t \times d_t}$, $X_a \in R^{l_a \times d_a}$ and $X_v \in R^{l_v \times d_v}$ from raw data. $F_{ti} \in R^{l_t \times d_t}$, $F_{ai} \in R^{l_a \times d_a}$ and $F_{vi} \in R^{l_v \times d_v}$ are processed from pre-train model. Here, d_t, d_a and d_v are 1024, 512, 256. And then we project them into a lower-dimensional space R^{d_m}, $m \in \{t, a, v\}$.

C. Inter-Modal Collaborative Learning

First of all, we then augment the inter-modal representation, followed by inter-modal transformers to capture modality dynamics for each time-step of the input sequences. Utilizing the attention mechanism to extract information for one sequence $F_{i,i}$ from another sequence $F_{j,j}$, the transformer encoder structure is used for those transformers. Queries, keys, and values are inputs to a transformer encoder. The source of queries is from $F_{i,i}$ while the source of keys and values should be from $F_{j,j}$. So the transformer encoder can be denoted as **Co-Transformer**($F_{i,i}$, $F_{j,j}$, $F_{j,j}$). Here, we use how to get the text inter-modal feature as an example.

$$F_{t,a} = \textbf{Co-Transformer}\ (F_{t,t}, F_{a,a}, F_{a,a}) \in R^{d_t} \quad (7)$$

$$F_{t,v} = \textbf{Co-Transformer}\ (F_{t,t}, F_{v,v}, F_{v,v}) \in R^{d_t} \quad (8)$$

$$F_t = F_{t,a} \odot F_{t,v}. \quad (9)$$

Here we utilize a hadamard product on $F_{t,a}$ and $F_{t,v}$ to get more potential information for text inter-modal features. F_a and F_v can be calculated by eliminating the potential noise in the text in the same way.

In addition, to finally get the same multimodal sample to show consistent sentiment performance after being extracted by the

inter-modal collaborative learning, we designed a collaborative loss function for this purpose.

$$\mathcal{L}_{colla} = \frac{1}{N}\sum_{i}^{N}\left(\mathrm{dist}\left(F_t, F_a\right) + \mathrm{dist}\left(F_t, F_v\right) + \mathrm{dist}\left(F_a, F_v\right)\right), \tag{10}$$

where dist means the predicted value of the pairwise modals is measured by the euclidean distance. N represents the total number of samples.

D. Intra-Modal Disentanglement Learning

First, the following equations take text modal as an example to show the feature intra-modal disentanglement learning process. As illustrated in Fig. 3, the algorithm disentangles the text representation, which is learned by feature representation F_t into the specific object part and irrelevant parts as:

$$F_t = \mathcal{F}_O + \mathcal{F}_B. \tag{11}$$

By the average layer, we can obtain the common representation $\mathcal{F}_O$, and identify it as the specific object feature that is beneficial to the final sentiment analysis through the following equation:

$$\mathcal{F}_O = (FC(F_t) + FC(F_a) + FC(F_v))/3. \tag{12}$$

The prototype $\mathcal{F}_O \in R^d$ represents the consistent part of the feature from the collaborative representation, such as the sentiment information. FC is a linear layer to change all the dimension to 1024 and d here is 1024. Then the irrelevant noise features are disentangled from the text representation as:

$$\mathcal{F}_B = F_t - \mathcal{F}_O. \tag{13}$$

In the above definition, we classify all other irrelevant and redundant features as background features, except for object features related to sentiment. These background features include cluttered background and varying noise. We treat background features as adversarial noise and weak background features to extract sentiment features in this paper.

Through the specific object features and the background features, we have added a special coefficient λ to generate disentanglement features as:

$$F_{t,dis} = \mathcal{F}_O + \lambda\mathcal{F}_B, \tag{14}$$

where $F_{t,dis}$ is the augmented new feature with a lot of background noise. The special coefficient λ randomly varies following a Gaussian distribution $\mathcal{N}$ in the training process $\lambda \sim \mathcal{N}(\gamma = 1, \Sigma)$, where $\gamma = 1$ means that the expected value of the adversarial coefficient is 1. The standard deviation Σ is the hyperparameter that controls the noise amplitude. A larger standard deviation indicates increased noise. In the experiment, the deviation is set as $\Sigma = 0.025$.

$F_{a,dis}$ and $F_{v,dis}$ can be calculated by eliminating the potential noise in the audio and video in the same way. By adding $\mathcal{F}_B$ to the loss function, the noise feature can be smaller.

$$\mathcal{L}_{dis} = \sum_{\mathcal{F}_B^t \in \Omega}\left\|\mathcal{F}_B^t\right\|_2 + \sum_{\mathcal{F}_B^a \in \Omega}\left\|\mathcal{F}_B^a\right\|_2 + \sum_{\mathcal{F}_B^v \in \Omega}\left\|\mathcal{F}_B^v\right\|_2, \tag{15}$$

Algorithm 1: TCDN Algorithm for Retrieving the Target Image or Text by The Multimodal Inputs.

Require:
Training texts, audio and video $\{X_t, X_a, X_v\}$; gaussian noise λ
while not converged **do:**
Sample a mini-batch from $\{X_t, X_a, X_v\}$
Feature embedding learning: F_{ti}, F_{ai} and F_{vi};
for transformer layers = 1,...,n **do**
$F_{t,t} \leftarrow$ **Transformer**(F_{ti}) via (4);
$F_{a,a} \leftarrow$ **Transformer**(F_{ai}) via (5);
$F_{v,v} \leftarrow$ **Transformer**(F_{vi}) via (6);
end for
for Co-Transformer layers = 1,...,n **do**
$F_{t,a} \leftarrow$ **Co-Transformer**$(F_{t,t}, F_{a,a}, F_{a,a})$ via (7);
$F_{t,v} \leftarrow$ **Co-Transformer**$(F_{t,t}, F_{v,v}, F_{v,v})$ via (8);
$F_t \leftarrow dot(F_{t,a}, F_{t,v})$ via (9);
end for
Augment text feature $F_{t,dis} = F_O + \lambda(F_t - F_O)$
Augment audio feature $F_{a,dis} = F_O + \lambda(F_a - F_O)$
Augment video feature $F_{v,dis} = F_O + \lambda(F_v - F_O)$ via (12)-(14);
Obtain F_{all} by $F_{t,dis}, F_{a,dis}$ and $F_{v,dis}$;
Compute $\mathcal{L}$ by $\hat{y}$, L1-norm$(F_t - F_a)$, L1-norm$(F_t - F_v)$ L1-norm$(F_a - F_v)$, L2-norm$(F_t - F_O)$, L2-norm$(F_a - F_O)$ and L2-norm$(F_v - F_O)$ via (10) (15) (17) (18);
Update the network parameters;
end while

where $\|\cdot\|_2$ denotes L2-norm, and Ω is real domain. From another perspective, it is equivalent to applying a mean square error loss to weaken the intra-modal class variance.

E. Model Training

The concatenation of three modality representation is regarded as the fusion results and is fed into a simple classifier to make a final prediction of the sentiment intensity.

$$F_{all} = \text{ Concat } \left[F_{t,dis}, F_{a,dis}, F_{v,dis}\right] \tag{16}$$

$$\hat{y} = W_1 \cdot \text{ LeakyReLU } \left(W_2 \cdot \text{BN}\left(F_{all}\right) + b_2\right) + b_1, \tag{17}$$

where BN is the BatchNorm operation, and LeakyReLu is used as activation.

We take the L1-Loss ($\mathcal{L}_{mae}$) as the basic optimization objective for sentiment intensity prediction. The total loss function can be expressed as:

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N}\left(\left|\hat{y}^i - y^i\right|\right) + \alpha\mathcal{L}_{colla} + \beta\mathcal{L}_{dis}. \tag{18}$$

The complete process of the algorithm is list in Algorithm 1.

IV. Experiments

In this section, some comparison experiments have been conducted in order to evaluate the performance of the proposed TCDN method.

TABLE I
DETAILS OF THE DATASETS STATISTICS

Dataset	Training	Validation	Testing
CMU-MOSEI-seven classes	16326	1871	4659
CMU-MOSE-seven classes	1283	229	686

A. Datasets

In this work, we use two public multimodal sentiment analysis datasets, MOSI [44] and MOSEI [43]. The CMU-MOSI dataset is a human multimodal sentiment analysis dataset consisting of 2,199 short video clips taken from 93 Youtube movie review videos. The CMU-MOSEI dataset expands its data with a higher number of samples over CMU-MOSI. The dataset contains 23,453 annotated video clips from 5,000 videos. For both CMU-MOSI and CMU-MOSEI, each sample is labeled with a sentiment score from -3(strongly negative) to 3(strongly positive). For the number of training, validation, and testing examples that we used for these two datasets are shown in Table I. Using different dataset settings, we can divide the data used for model training into two categories: aligned data and unaligned data. We also marked it in Table II.

B. Implementation Details

We implemented our model using Fairseq [15] toolkit. The networks are trained with Adam optimizer. Warm-up updates and the polynomial decay learning-rate scheduler are used in the training step. The initial learning rate is set to 3e-4. All experiments are trained by using two NVIDIA GeForce RTX 3090.

C. Evaluation Metrics

We report our experimental results in two forms: classification and regression. For classification, we calculate Acc-2 (binary classification accuracy) and F1-score by negative / non-negative (non-exclude zero). Then the Acc-2 and F1 score are computed as follows:

$$\mathrm{ACC}_2 = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{19}$$

$$\mathrm{F1} = \frac{2}{\mathrm{Precision}^{-1} + \mathrm{Recall}^{-1}} = \frac{\mathrm{TP}}{\mathrm{TP}+0.5(\mathrm{FP}+\mathrm{FN})} \tag{20}$$

For regression, we report MEAN Absolute Error (MAE) and Pearson correlation (Corr). For MAE, lower values denote better performance, while others are the opposite.

D. Result Analyses

We select several state-of-the-art baselines to evaluate the effectiveness of our TCDN in Table II. The details of competitors are listed as follows:

TFN [42] obtained a multi-dimensional tensor by calculating the outer product of modal vectors to capture the interaction between single modal, two modals and three modals.

LMF [10] network was an improvement of TFN, which used low-rank multimodal tensor fusion technology to improve the efficiency of modal fusion.

RAVEN [33] introduced multimodal information to effectively solve the problem of insufficient semantic representation of words.

MCTN [16] proposed a modal translation model to learn robust feature representation. The accuracy of modal translation is measured by periodic consistency loss.

MFM [30] decomposed the modal features to generate multimodal discriminant factors and specific modal generation factors. The multimodal discriminant factor was shared in all modals and contains the shared multimodal features needed for sentiment analysis tasks.

MulT [30] used the directional modal cross-attention mechanism, which solves the long-term dependency between modal elements in an end-to-end manner by paying attention to the interaction between multimodal sequences with different time steps, without the need to explicitly align data.

ICCN [27] used canonical correlation network to extract the hidden relationship between sentence embedding, audio sequence and facial features of BERT model.

MAG-BERT [19] employed multimodal adaptive gates and is attached to the BERT model. MAG allows the BERT model to accept multimodal nonverbal data during fine-tuning.

MISA [4] employed modality-invariant and specific representations. It projected each modal into two different subspaces. The first subspace is the modal invariant representation subspace, and the second subspace is the modal specific representation subspace.

TCM-LSTM [14] mainly used acoustics and visual LSTM to enhance the expression of oral language.

CTFN [29] can perform two-way cross-channel cross-correlation in parallel. Furthermore, a hierarchical architecture is established to make use of multiple two-way translations, thus realizing double multi-channel fusion.

TCSP [37] proposed a text-centered multi-channel fusion and sharing private framework. It consists of two parts: cross-modal prediction and emotional regression.

SPT [1] used sampling function to generate sparse attention matrix and captured the interaction between hidden states.

Self-MM [40] designed a tag generation module based on self-supervised learning strategy to generate specific single modal tags. Then, the joint training of multimodal task and single modal task were carried out to learn the consistency and difference between modals.

TPMSA [39] adopted a two-stage multi-channel framework, making full use of the pre-training model and a new multi-task classification learning strategy.

E. Quantitative Results

Table II shows the comparative results on CMU-MOSI and CMU-MOSEI datasets. In our experiments, first, comparing with unaligned models (TFN and LMF), we achieve a significant improvement in all evaluation metrics. Even comparing with aligned models, our method gets competitive results.

TABLE II
SENTIMENT ANALYSIS ON CMU-MOSI AND CMU-MOSEI DATASET

Methods	Sources	CMU-MOSI				CMU-MOSEI			
		regression		classification		regression		classification	
		MAE ↓	Corr ↑	Acc2 ↑	F1 ↑	MAE ↓	Corr ↑	Acc2 ↑	F1 ↑
TFN (unaligned) [44]	EMNLP2017	0.901	0.698	80.8	80.8	0.593	0.700	82.6	82.1
LMF (unaligned) [10]	ACL2018	0.917	0.695	82.5	82.5	0.623	0.677	82.0	82.2
RAVEN (unaligned) [35]	AAAI2019	0.915	0.691	78.0	76.6	0.614	0.662	79.1	79.5
MCTN (aligned) [17]	AAAI2019	0.909	0.676	79.3	79.1	0.609	0.670	79.8	80.6
MFM (aligned) [32]	ACL2019	0.877	0.706	81.7	81.6	0.568	0.717	84.4	84.4
MulT (aligned) [31]	ACL2019	0.871	0.698	83.0	82.8	0.591	0.694	81.6	81.6
ICCN (aligned) [28]	AAAI2020	0.862	0.714	83.1	83.0	0.565	0.713	84.2	84.2
MAG-BERT (aligned) [20]	ACL2020	0.739	0.796	86.1	86.0			84.7	84.5
MISA (aligned) [4]	MM2020	0.783	0.761	83.4	83.6	0.555	0.756	85.5	85.3
TCM-LSTM (unaligned) [14]	TASLP2021	0.903	0.672	81.7	81.8	0.673	0.606	81.4	81.6
CTFN (unaligned) [30]	ACL2021	-	-	82.8	82.9	-	-	-	-
TCSP (unaligned) [39]	ACL2021	0.908	0.71	80.9	81.0	0.576	0.715	82.8	82.7
SPT (unaligned) [1]	EMNLP2021	-	-	82.8	82.9	-	-	82.6	82.8
Self-MM (unaligned) [42]	AAAI2021	0.713	0.798	86.0	86.0	0.530	0.765	85.2	85.3
TPMSA (unaligned) [41]	TASLP2022	0.704	0.799	87.0	87.0	0.542	0.770	85.6	85.6
TCDN (unaligned)		**0.697**	**0.805**	**87.1**	**87.2**	**0.521**	**0.782**	**87.5**	**87.2**

Moreover, we compare our method with the three best performing models Self-MM, MISA and MAG-BERT from all the existed methods. On CMU-MOSI dataset, the indexes of Acc2, F1, MAE and Corr reached 87.1%, 87.2%, 0.697 and 0.805 respectively, which increased by 1.1%, 1.2%, 0.007 and decreased by 0.016 compared with the best results of Self-MM models. On CMU-MOSEI dataset, we find it remarkable that TCDN, despite being a single-level method, outperforms TPMSA by a significant margin, more than 1.9% points for Acc-2 and 1.6% points for F1. The main reason for this phenomenon is that our multi-scale inter-modal feature collaborative network can enhance the emotional features extracted from a single modal, and the co-transformer structure can extract the relationship between modal features. In addition, the multimodal sentiment feature disentanglement learning and the inter-modal emotional feature collaborative learning can guide the model to further enhance and integrate the sentiment features of each modal. We find that our model surpasses them on most of the evaluations. The above results show that our model can be applied to different data scenarios and achieve significant improvements.

F. Ablation Studies

1) Performance Comparison Between Single Modal Input and Multimodal Input: From Table III, in the case of single modal, the text modal has the highest binary classification accuracy and the lowest MAE value, with an Acc2 of 79.2% and a MAE of 0.960; the Acc2 of the video modal is 57.9% and the MAE of the audio modal is 1.441 and the Acc2 of the audio modal is 56.3% and the MAE is 1.440. The results show that the features extracted by the feature enhancement network in the single modal can complete the task of emotion analysis. The final effect of a single text modal is much better than that of a single other modal. The possible reason is that the RoBERTa [9] pre-training model itself can generate CLS tokens, and the generated CLS tokens have rich text features, which do not need to be generated by the feature enhancement module in the training single modal from scratch.

TABLE III
THE IMPACT OF SINGLE MODAL INPUT AND MULTIMODAL INPUT ON CMU-MOSE DATASET

Modal Task	regression		classification	
	MAE ↓	Corr ↑	Acc2 ↑	F1 ↑
A	1.440	0.144	56.3	56.4
V	1.441	0.121	57.9	57.6
T	0.960	0.674	79.0	78.1
A,T	0.724	0.774	86.1	86.2
T,V	0.803	0.781	86.2	86.3
A,V	0.780	0.765	86.0	86.3
A,T,V	**0.697**	**0.805**	**87.1**	**87.2**

In the case of bimodal, thanks to the feature enhancement network within single modal, the feature fusion network between bimodal and auxiliary learning, audio text bimodal, text video bimodal and audio video bimodal all achieve better results than single modal, which proves the effectiveness of bimodal feature fusion network.

In the case of trimodal, a feature fusion network between three modals is added compared with the bimodal model structure, and the optimal results with Acc2 of 87.1%, F1 of 87.2%, MAE of 0.697 and Corr of 0.805 are obtained, which proves the effectiveness of the feature fusion network between three modals.

2) Performance Comparison With or Without Learning Module in Multimodal Scenarios: Through a comparative experiment on whether or not to add learning module tasks to the dual-modal and three-modal fusion networks, the results in Table IV show that the Inter-modal collaborative learning and Intra-modal disentanglement learning as the learning modules improve the performance of both networks. Particularly in the audio and video (AV) dual-modal task, inter-modal collaborative learning and intra-modal disentanglement learning improves the

TABLE IV
THE IMPACT OF WITH OR WITHOUT LEARNING MODULE FOR SENTIMENT ANALYSIS ON CMU-MOSE DATASET

Learning Module Task	regression		classification	
	MAE ↓	Corr ↑	Acc2 ↑	F1 ↑
with learning module AT	0.724	0.774	86.1	86.2
with learning module TV	0.803	0.781	86.2	86.3
with learning module AV	0.78	0.765	86	86.3
with learning module ATV	**0.697**	**0.805**	**87.1**	**87.2**
w/o learning module AT	0.73	0.772	85.2	85.2
w/o learning module TV	0.75	0.781	83.8	84
w/o learning module AV	1.544	0.096	54.4	55.1
w/o learning module ATV	0.803	0.761	84.2	84.5

TABLE V
THE IMPACT OF DIFFERENT LAYERS FOR TRANSFORMER AND CO-TRANSFORMER ON CMU-MOSE DATASET

Number of Layers		regression		classification	
Transformer	Co-Transformer	MAE ↓	Corr ↑	Acc2 ↑	F1 ↑
1	1	**0.697**	**0.805**	**87.1**	**87.2**
1	2	0.803	0.761	84.2	84.5
1	3	0.776	0.768	83.9	83.5
2	1	0.735	0.808	86.2	86.2
2	2	0.743	0.789	85.3	85.6
2	3	0.883	0.725	82.6	82.1
3	1	0.865	0.724	85.0	85.1
3	2	0.765	0.770	86.4	86.4
3	3	0.784	0.759	86.0	86.1

bimodal fusion network's Acc2 by 31.6%, F1 by 31.2%, while MAE decreases by 0.764% and Corr increased by 0.669.

In the three-modals task, adding inter-modal collaborative learning and intra-modal disentanglement learning increases the trimodal fusion network's Acc2 by 2.9%, F1 by 2.7%, while MAE decreases by 0.106 and Corr increases by 0.044.

In cases where the dual-modal fusion network performed poorly, the addition of learning modules greatly improve its performance. The trimodal fusion network already performed well, adding learning modules still significantly improve its performance.

The experimental results demonstrate that intra-modal disentanglement learning suppresses emotionally unrelated parts of features while inter-modal collaborative learning integrates and complements sentiment features across modals to ultimately improve model accuracy.

In conclusion, text modal plays a vital role in emotion recognition but our two learning modules can still enhance accuracy even without text.

3) Configuration of the Number of Transformer Layers and Co-Transformer Layers: Table V illustrates the impact of different numbers in Transformer layers and Co-Transformer layers. It can be seen that the result of our model increases at first and then decreases when the layer's number becomes larger. With the increase of the layers, we get the best performance when the layer numbers are set to 1 and 1 for transformer and co-transformer on CMU-MOSE datasets. The consequences indicate that larger layers do not always contribute to better performance. Because of the larger parameters, the model may be difficult to get a good result when using multi layers. So it is necessary to use a suitable-sized layer for TCDN model.

TABLE VI
THE IMPACT OF DIFFERENT FUSION MODELS

Fusion Mode	regression		classification	
	MAE ↓	Corr ↑	Acc2 ↑	F1 ↑
concat	**0.697**	**0.805**	**87.1**	**87.2**
dot	0.958	0.658	79.8	79.5
add	0.779	0.766	85.2	85

TABLE VII
THE IMPACT OF DIFFERENT LOSS FUNCTION FOR TCDN MODULE ON MOSI TEST SET AND MOSEI TEST SET

Datasets	$\mathcal{L}_{mae}$	$\mathcal{L}_{colla}$	$\mathcal{L}_{dis}$	regression		classification	
				MAE ↓	Corr ↑	Acc2 ↑	F1 ↑
MOSI	✓			0.803	0.761	84.2	84.5
	✓	✓		0.73	0.772	85.2	85.2
	✓		✓	0.702	0.801	86.9	86.9
	✓	✓	✓	**0.697**	**0.805**	**87.1**	**87.2**
MOSEI	✓			0.529	0.761	86.3	86.3
	✓	✓		0.525	0.774	87.1	86.8
	✓		✓	0.523	0.780	87.2	87.1
	✓	✓	✓	**0.521**	**0.782**	**87.5**	**87.2**

4) Effect of Different Fusion Models: Table VI shows the fusion models of TCDN model with three different functions in (16). The results also show that both add and dot make the final feature lose some semantic information, which leads to the decrease of accuracy. Therefore, we can find that the information concat of the three modals can ensure that each modal plays a role in the final score as much as possible.

5) Effect of Orthogonal Vector-Decomposed Disentanglement: Table VII shows the consequences of TCDN model with three different loss functions which have represented in (18).

We only use $\mathcal{L}_{mae}$ as the total loss function. At this time, the preliminary features' representation (F_t, F_a, F_v) is used for analysis. Compared with the following loss function, its retrieval accuracy is the lowest.

On the basis of the former, we introduce the inter-modal representation obtained by the collaborative learning network to construct a new loss function. This loss effectively ensures a consistent emotional expression after the fusion. From Table VII, we find that the improvement of accuracy is conspicuous.

Finally, we add the representation obtained by intra-modal disentanglement network to construct the loss function ($\mathcal{L}_{mae}$, $\mathcal{L}_{reg}$, $\mathcal{L}_{dis}$). The loss function performs better than the two variants above on two datasets, indicating that the network can effectively weaken the cluttered background features and different occlusions features from intra-modal. The goal of this loss function is to limit the irrelevant noise features and redundant features effectively.

6) Impact of Loss Factor α, β and λ: For CMU-MOSE dataset and CMU-MOSEI dataset, Fig. 4 shows the impact of α from Eq.(18). Left of the image, we can find that when α is 0, the model trains without inter-modal collaborative learning and does not use inter-modal collaborative loss. So it cannot get a good performance on two datasets at this time. As the value

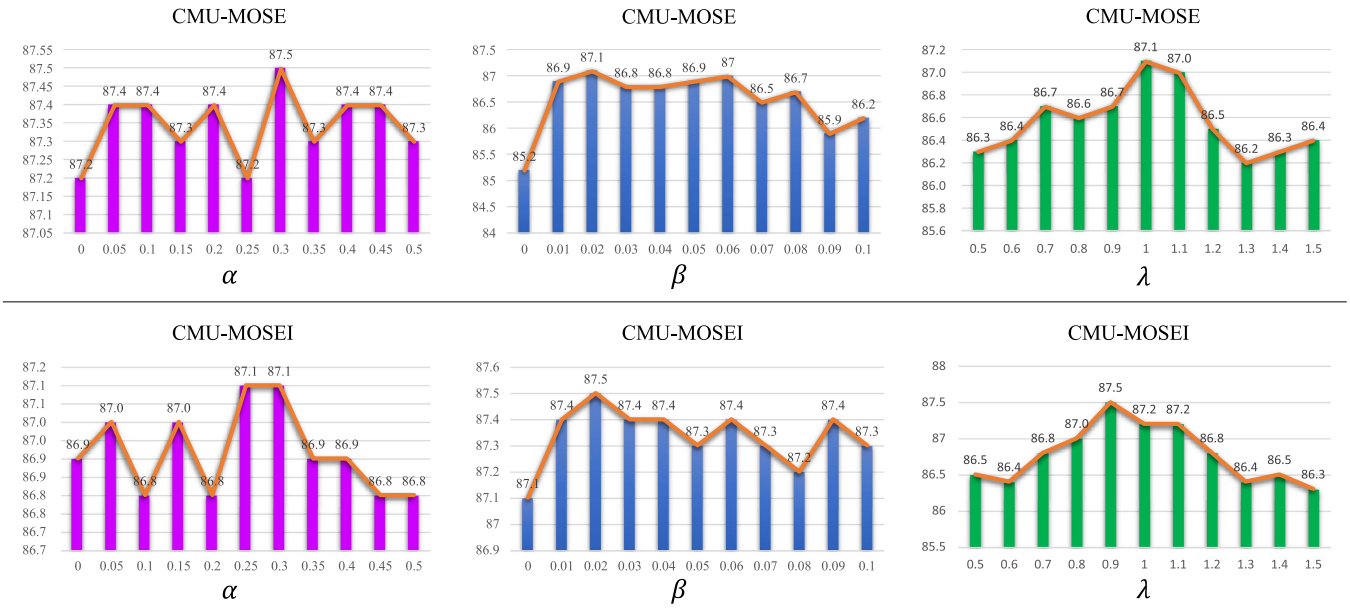


Fig. 4. Sentiment analysis consequences in on CMU-MOSE dataset and CMU-MOSEI dataset with different α, β and λ value.

of α increases, the accuracy increases slightly on two datasets. The accuracy reaches the highest point when α is 0.3. The results show that no matter what the rate α of the accuracy other result is always better than the result of without inter-modal collaborative training.

In the middle of Fig. 4, it shows the impact of β in (18) on CMU-MOSE dataset and CMU-MOSEI dataset. We can find that when β is 0, the model trains without intra-modal disentanglement learning and the features have many unrelated contents here. So it cannot get a good performance on the dataset at this time. As the value of β increases, the accuracy increases slightly on two datasets. At this time, the redundant and irrelevant features have been removed. The accuracy reaches the highest point when β is 0.02. The results show that no matter what the rate β of the accuracy is other results always better than the result of without intra-modal disentanglement training.

In the right of Fig. 4, it shows the impact of λ in (14) on CMU-MOSE dataset and CMU-MOSEI dataset. The size adjustment of λ is mainly adjusted by γ. Increasing γ means that the adversarial coefficient of Gaussian noise is getting larger and larger. We can find that when λ is large, the noise added in the process of model training is too large, so the accuracy is not high. So it cannot get a good performance on the dataset at this time. As the value of λ is small, the accuracy increases slightly on two datasets. At this time, The valid part of the background is preserved and the invalid part is removed. The accuracy reaches the highest point when λ is 1 and 0.9 on CMU-MOSE dataset and CMU-MOSEI dataset.

G. Visualization Results

1) Examples of Visualization Results: To prove that our network can have good results in the field of sentiment analysis, we visualize the input of text, voice and video, as well as the score of each modal feature and the final sentiment analysis score in Fig. 5. Red represents negative emotion, green represents positive emotion, and white represents no emotion.

For the first example, we cannot read any emotional orientation from the text. However, the real label of this sample is -0.4, which shows the limitation of text modality, and there are some situations where words are not enough to express emotion. And the tone of the audio modal is calm and emotionally biased. We believe that the real tag is presented by considering the visual modal that accompanies the video, and our model has a score of -0.93 in the visual modal. This explains that our model ends up with negative predictions.

For the second example, we found a positive word "*LAUGHED*" in the text. However, if we carefully observe the voice and video, we can find that the overall voice is only a simple statement, and the whole process of the video is expressionless. The emotional bias of the text modal is obvious, which is a positive emotion. TCDN makes up for the defect of emotional expression of audio and video modal to some extent by using text modal.

For the third example, we can see the obvious emotional polarization from the acoustic point of view. However, acoustic does not always express positive emotions. In other words, the comment is ambiguous in terms of the feelings expressed in the audio. We observe that the real label of this sample is 0.6 and the prediction of our model is 0.65, which indicates a weak positive emotion. However, the prediction result with only speech modal is -0.31, which is a weak negative emotion. TCDN effectively avoids this problem through collaborative learning and antagonistic learning. The final score results prove the effectiveness of visual and text patterns in acoustic disambiguation.

2) Visualize the Transformers' Attention Values: To prove that our network can extract useful fine-grained cross-modal information, we visualize the words in the text corresponding to the frame in the input video and the words in the text corresponding to the frame in the input audio in Figs. 6 and

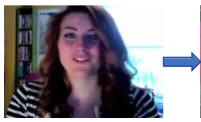
(a)

#	words, acoustic and visual behaviors	Modal predictive value	Our model TCDN	Ground Truth
words	UM AND NOW THE PERFORMANCES ARE LIKE SOMETHING LIKE TO WRITE HOME ABOUT	0.03	-0.53	-0.4
acoustic	Unemotional tone	0.04		
visual	 Scratch ears Helpless	-0.93		

(b)

#	words, acoustic and visual behaviors	Modal predictive value	Our model TCDN	Ground Truth
words	I LAUGHED AT QUITE A BIT OF THE DIALOGUE	1.59	1.29	1.2
acoustic	indicative mood Calm tone	0.03		
visual	 Calm face	-0.10		

(c)

#	words, acoustic and visual behaviors	Modal predictive value	Our model TCDN	Ground Truth
words	UM SO IF YOU LIKE TO HEAR A UM LIKE MORE POSITIVE REVIEW CHECK OUT HER REVIEW IT IS VERY GOOD	1.01	0.65	0.6
acoustic	Calm tone Falling tone	-0.31		
visual	 Smile expression	1.15		

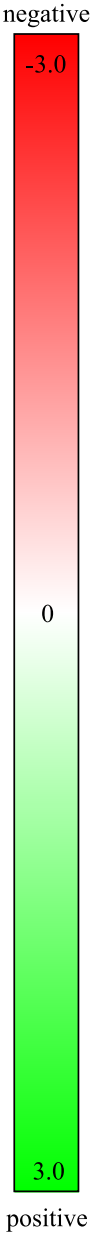


Fig. 5. Three examples for the visualization figures of the sentiment analysis, as well as the trimodal scores extracted by our TCDN on CMU-MOSE dataset. The color box reflects the positive and negative of the sentiment. Red represents negative emotion, green represents positive emotion, and white represents no emotion.

7. We use the features changed by cross-modal transformer to calculate the attention weight after the multi-head attention mechanism in the transformer and finally obtain the number with the highest weight corresponding to each box. The darker the yellow, the higher the weight of the attention values. The ellipsis part indicates the padding operation to be carried out uniformly during training. The dimension 0 represented by the first line is the classification (CLS) position in transformer, which is expressed as the weight relationship of classification features.

For example, in the case of the video and text on the left and the audio and text on the right in Fig. 6, the returned text has the corresponding words "G," "ALMOST" and "HAD," where the relevant areas and words are highlighted. The emotional picture shown by the person in the video when she says the word G is pessimistic and the sound is accentuated, so she gets the highest weight in this position from V-T co-Transformers' attention values. She is also depressed when she says the words "ALMOST" and "HAD," so she gets a relatively high weight in this position from A-T co-Transformers' attention values. The contents of the three modals are consistent with the attention value we got. The results show that our method can obtain the cross-modal features from co-transformer layers for MSA task.

In Fig. 7, to explore the effectiveness of cross-modal transformer for video and audio in our model, we show exemplary visualization results. The emotional focus in the audio is consistent with the emotional performance point of the task in the video. When the tone changes, attention value is also relatively high. In the whole audio sequence, a relatively high weight is

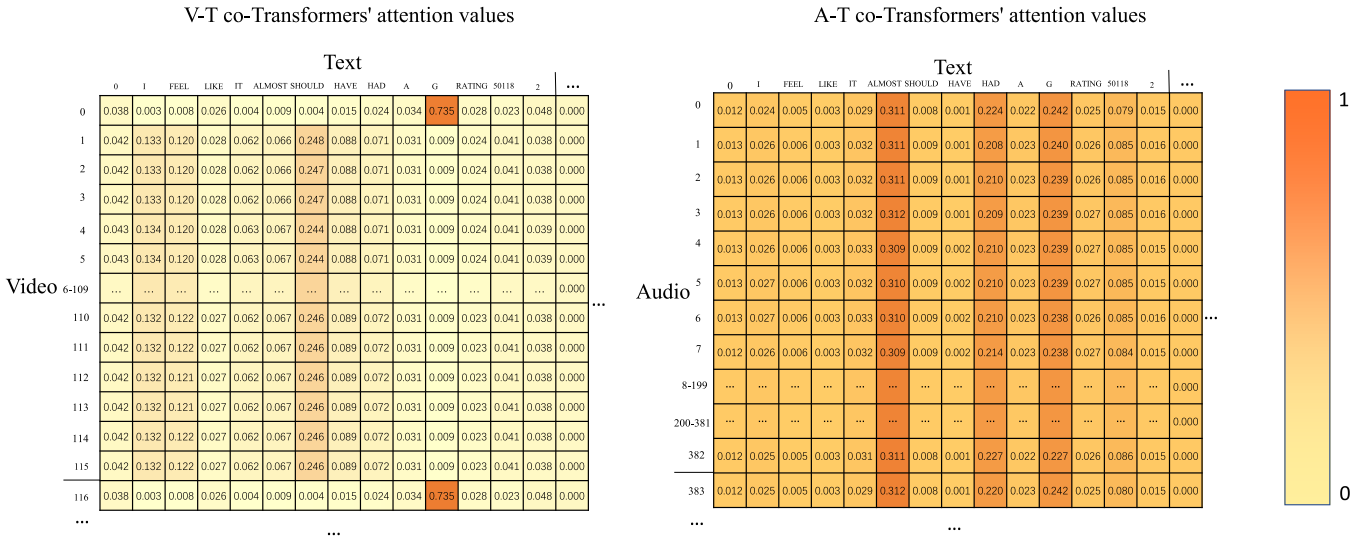


Fig. 6. Sentiment analysis consequences in the cross-modal Transformers' Attention Values on V-T and A-T.

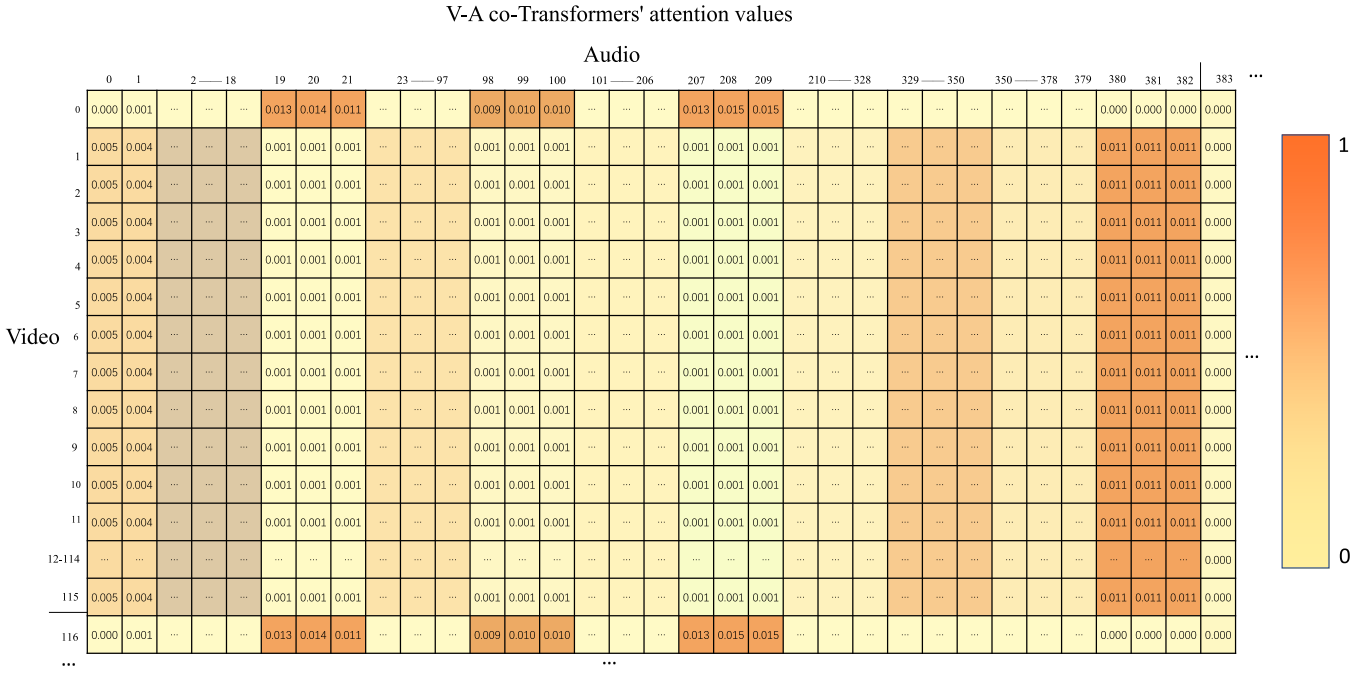


Fig. 7. Sentiment analysis consequences in the cross-modal Transformers' Attention Values on V-A.

obtained in the first few frames, the first few frames, and the last few frames, which is consistent with the subjective perception of the audio and video modals observed in the experiment.

3) Visualize the t-SNE Plot: In Fig. 8, through the t-SNE experiment, we can find that the model can separate the characteristics of different emotions and cluster the content similar to emotion. From the clustering results, the results obtained by the model are similar to the tasks. When the tasks are classification tasks, there is an obvious dividing line between the clustering results. The red on the left is negative emotion, and the green on the right is positive emotion.

By comparing box 1 and box 2 in Fig. 8, our model can better cluster emotion-like features together than baseline, so our model is more accurate on Acc2.

By comparing box 3 and box 4 in Fig. 8, according to the results of Acc7 and Acc2 classification, we can find that our model can cluster stronger emotions in the tail, and the color decreases gradually according to the gradual decrease of emotion. It shows that our model also has a great advantage in the accuracy of subdivision. Therefore, through collaborative learning and disentanglement learning, our model extracts more emotional features and achieves higher accuracy.

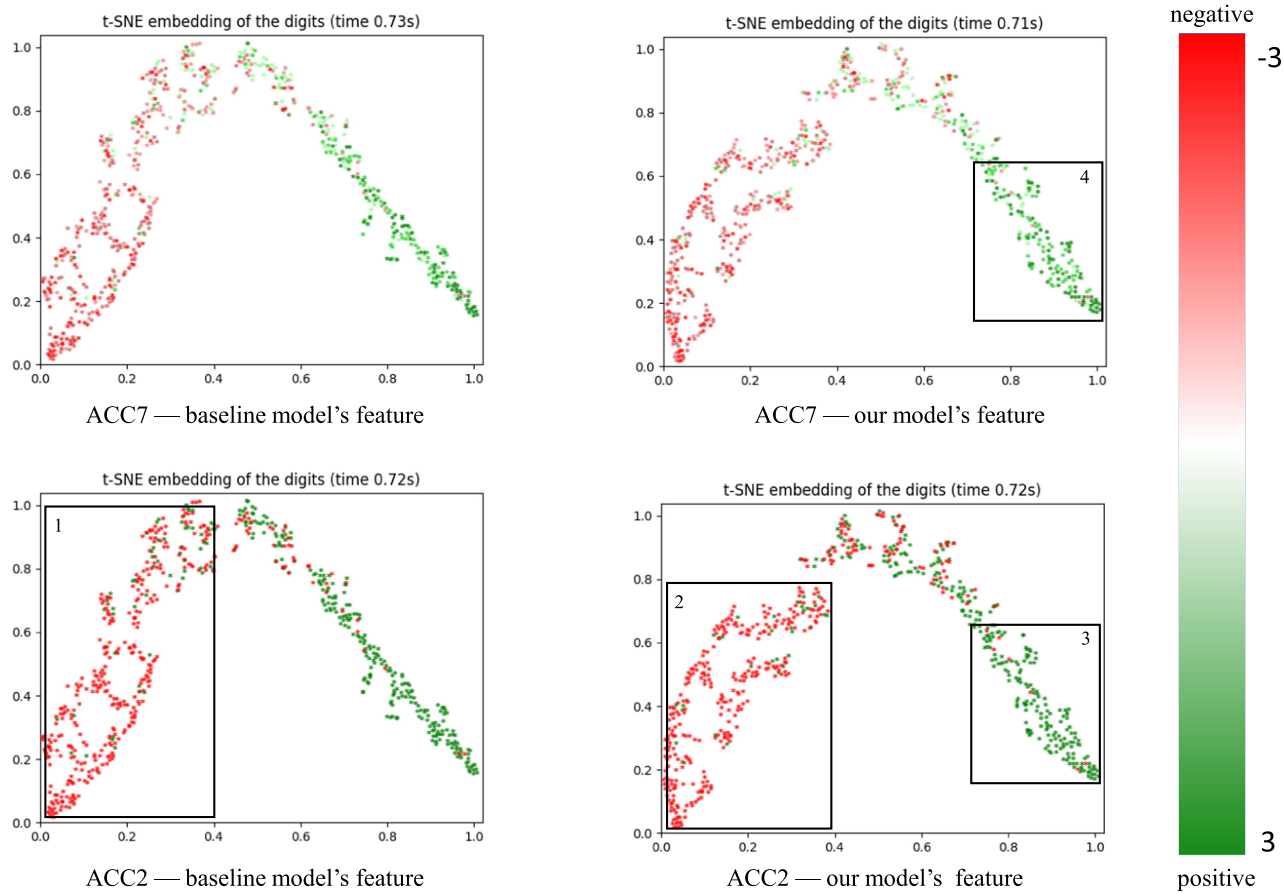


Fig. 8. Sentiment analysis consequences on CMU-MOSEI dataset with t-SNE embedding.

V. Conclusion

In this paper, we study the problem of sentiment analysis, and propose a novel inter-intra modal representation augmentation method based on Trimodal Collaborative Disentanglement Network (TCDN). This method significantly improves the multimodal sentiment analysis performance. Through collaborative learning to control the relationship between modals, we effectively reduce the impact of inter-modal noise features on accuracy. Combined with the common representation and each modalities' feature, the method of disentanglement learning is adopted to effectively reduce the influence of intra-modal noise feature on accuracy. Comprehensive experiments show that the performance of this method on two benchmark datasets is obviously superior to the other traditional methods.

In the future, under the premise of ensuring the accuracy, more consideration should be given to the lightweight of the model to effectively reduce the model parameters for the MSA task.

References

[1] J. Cheng, I. Fostiropoulos, B. Boehm, and M. Soleymani, "Multimodal phased transformer for sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2447–2458.

[2] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.

[3] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, 2018, pp. 289–293.

[4] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 1122–1131.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[7] X. Li et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 121–137.

[8] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[9] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.

[10] Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2247–2256.

[11] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13–23.

[12] H. Luo, L. Ji, Y. Huang, B. Wang, S. Ji, and T. Li, "Scalevlad: Improving multimodal sentiment analysis via multi-scale fusion of locally descriptors," 2021, *arXiv:2112.01368*.

[13] S. Mai, H. Hu, and S. Xing, "Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 164–172.

[14] S. Mai, S. Xing, and H. Hu, "Analyzing multimodal sentiment via acoustic- and visual-LSTM with channel-aware temporal convolution network," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 1424–1437, 2021.

[15] M. Ott et al., "fairseq: A fast, extensible toolkit for sequence modeling," *Proc. Conf. North Amer. Chapter Assoc. Computat. Linguistics (Demonstrations)*, 2019, pp. 48–53.
[16] 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 *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6892–6899.
[17] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 439–448.
[18] 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 Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132, Jan.–Mar. 2023.
[19] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. Conf. Assoc. Comput. Linguistics Meeting*, 2020, Art. no. 2359.
[20] S. U. Rehman, S. Tu, Y. Huang, and O. U. Rehman, "A benchmark dataset and learning high-level semantic embeddings of multimedia for cross-media retrieval," *IEEE Access*, vol. 6, pp. 67176–67188, 2018.
[21] S. U. Rehman, S. Tu, Y. Huang, and Z. Yang, "Face recognition: A novel un-supervised convolutional neural network method," in *Proc. IEEE Int. Conf. Online Anal. Comput. Sci.*, 2016, pp. 139–144.
[22] S. U. Rehman, S. Tu, O. Ur Rehman, Y. Huang, C. M. S. Magurawalage, and C. C. Chang, "Optimization of CNN through novel training strategy for visual classification problems," *Entropy*, vol. 20, no. 4, 2018, Art. no. 290.
[23] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. Interspeech*, 2019, pp. 3465–3469.
[24] Y. Su, K. Fan, N. Bach, C. C. J. Kuo, and F. Huang, "Unsupervised multimodal neural machine translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10482–10491.
[25] Y. Su and C.-C. J. Kuo, "On extended long short-term memory and dependent bidirectional recurrent neural network," *Neurocomputing*, vol. 356, pp. 151–161, 2019.
[26] Y. Su et al., "Recurrent neural networks and their memory behavior: A survey," *APSIPA Trans. Signal Inf. Process.*, vol. 11, no. 1, 2022.
[27] Z. Sun, P. Sarma, W. Sethares, and Y. Liang, "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 8992–8999.
[28] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, 2018.
[29] J. Tang, K. Li, X. Jin, A. Cichocki, Q. Zhao, and W. Kong, "CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5301–5311.
[30] Y. H. Tsai, S. Bai, P. Pu Liang, J. Z. Kolter, L. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics. Meeting*, 2019, Art. no. 6558.
[31] S. Tu et al., "CSFL: A novel unsupervised convolution neural network approach for visual pattern classification," *AI Commun.*, vol. 30, no. 5, pp. 311–324, 2017.
[32] S. ur Rehman, et al., "Unsupervised pre-trained filter learning approach for efficient convolution neural network," *Neurocomputing*, vol. 365, pp. 171–190, 2019.
[33] Y. Wang, Y. Shen, Z. Liu, P. P. Liang, A. Zadeh, and L. P. Morency, "Words can shift: Dynamically adjusting word representations using nonverbal behaviors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 7216–7223.
[34] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proc. Web Conf.*, 2020, pp. 2514–2520 .
[35] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10941–10950.
[36] A. S. Koepke, O. Wiles, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," in *Proc. Brit. Mach. Vis. Conf.*, 2018, p. 302.
[37] Y. Wu, Z. Lin, Y. Zhao, B. Qin, and L. Zhu, "A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis," in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 4730–4738.
[38] Y. Wu, S. Wang, G. Song, and Q. Huang, "Learning fragment self-attention embeddings for image-text matching," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2088–2096.
[39] B. Yang, L. Wu, J. Zhu, B. Shao, X. Lin, and T. Y. Liu, "Multimodal sentiment analysis with two-phase multi-task learning," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 2015—2024, 2022.
[40] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10790–10797.
[41] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
[42] A. Zadeh and M. Chen, S. Poria, E. Cambria, L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1103–1114.
[43] A. Zadeh and P. Pu, "Multimodal language analysis in the wild: CMU-mosei dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2236–2246.
[44] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," 2016, *arXiv:1606.06259*.
[45] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.
[46] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Representations*, 2018.
[47] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13001–13008.

Chen Chen received the B.E. degree in communication engineering from the Lanzhou University of Technology, Lanzhou, China, in 2019. He is currently working toward the Ph.D. degree in information and telecommunication engineering with the School of Telecommunications Engineering, Xidian University, Xi'an, China. His research interests include deep learning, recommendation systems, and cross-modal retrieval.

Hansheng Hong received the master's degree in computer engineering from the National University of Singapore, Singapore, in 2010. He is currently the Chief Software Technical Expert and Chief Software Architect of OPPO. He has very rich practical experience in software operating systems, and has been responsible for the planning and design of flyme, coloros, pantanal, and other software systems.

Jie Guo (Member, IEEE) received the B.S. degree in communication engineering from Zhengzhou University, Zhengzhou, China, in 2011, and the Ph.D. degree from Xidian University, Xian, China, in 2017. She is currently an Associate Professor with Xidian University. From 2015 to 2016, she got the State Scholarship Fund from China Scholarship Council to be an exchange Ph.D. Student with Carleton University, Ottawa, ON, Canada. Her research interests include information fusion, deep learning, recommendation systems, and cross-modal retrieval.

Bin Song (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 1996, 1999, and 2002, respectively. He is currently a Professor of information and telecommunication engineering with the Xidian University. He has authored more than 100 journal papers or conference papers and 50 patents. His research interests and areas of publication include multimedia communication, multimodal data fusion, content-based image recognition and machine learning, reinforcement learning, Internet of Things, Big Data, and recommendation systems.