International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02304-3

Noise-Resistant Multimodal Transformer for Emotion Recognition

Yuanyuan Liu[1] · Haoyu Zhang[1,2] · Yibing Zhan[4] · Zijing Chen[5,6] · Guanghao Yin[1] · Lin Wei[1] · Zhe Chen[3,6]

Received: 19 January 2023 / Accepted: 12 November 2024

Abstract
Multimodal emotion recognition identifies human emotions from various data modalities like video, text, and audio. However, we found that this task can be easily affected by noisy information that does not contain useful semantics and may occur at different locations of a multimodal input sequence. To this end, we present a novel paradigm that attempts to extract noise-resistant features in its pipeline and introduces a noise-aware learning scheme to effectively improve the robustness of multimodal emotion understanding against noisy information. Our new pipeline, namely Noise-Resistant Multimodal Transformer (NORM-TR), mainly introduces a Noise-Resistant Generic Feature (NRGF) extractor and a multimodal fusion Transformer for the multimodal emotion recognition task. In particular, we make the NRGF extractor learn to provide a generic and disturbance-insensitive representation so that consistent and meaningful semantics can be obtained. Furthermore, we apply a multimodal fusion Transformer to incorporate Multimodal Features (MFs) of multimodal inputs (serving as the key and value) based on their relations to the NRGF (serving as the query). Therefore, the possible insensitive but useful information of NRGF could be complemented by MFs that contain more details, achieving more accurate emotion understanding while maintaining robustness against noises. To train the NORM-TR properly, our proposed noise-aware learning scheme complements normal emotion recognition losses by enhancing the learning against noises. Our learning scheme explicitly adds noises to either all the modalities or a specific modality at random locations of a multimodal input sequence. We correspondingly introduce two adversarial losses to encourage the NRGF extractor to learn to extract the NRGFs invariant to the added noises, thus facilitating the NORM-TR to achieve more favorable multimodal emotion recognition performance. In practice, extensive experiments can demonstrate the effectiveness of the NORM-TR and the noise-aware learning scheme for dealing with both explicitly added noisy information and the normal multimodal sequence with implicit noises. On several popular multimodal datasets (e.g., MOSI, MOSEI, IEMOCAP, and RML), our NORM-TR achieves state-of-the-art performance and outperforms existing methods by a large margin, which demonstrates that the ability to resist noisy information in multimodal input is important for effective emotion recognition.

Keywords Multimodal · Emotion recognition · Transformer · Noise-resistant generic feature · Noise-aware learning scheme

1 Introduction

An accurate understanding of human emotions is beneficial for several applications, such as multimedia analysis, digital entertainment, health monitoring, human-computer interaction, *etc* (Shen et al., 2009; Beale & Peter, 2008; Qian et al., 2019; D'Mello & Kory, 2015). Compared with traditional emotion recognition, which only uses a unimodal data source, multimodal emotion recognition that exploits and explores different data sources, such as visual, audio, and text, has shown significant advantages in improving the understanding of emotions (Zadeh et al., 2017; Tsai et al., 2019; Lv et al., 2021; Hazarika et al., 2020; Yuan et al., 2021), including happiness, anger, disgust, fear, sadness, neutral, and surprise.

Recently, most existing multimodal emotion recognition methods mainly focus on multimodal data fusion, including tensor-based fusion methods (Liu et al., 2018; Zadeh et al., 2017; Sahay et al., 2020; Yuan et al., 2021) and attention-based fusion methods (Zhao et al., 2020; Huang et al., 2020; Zhou et al., 2021). The tensor-based fusion methods aim to obtain a joint representation of data with different modalities via multilinear function calculation. For example, TFN (Liu et al., 2018) used Cartesian product operation

Communicated by Vittorio Murino.

Extended author information available on the last page of the article

Published online: 22 December 2024

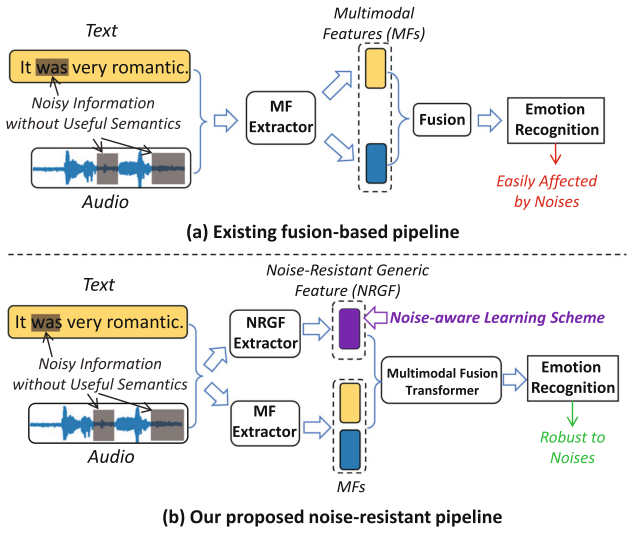

Text
It was very romantic.
Noisy Information without Useful Semantics
Audio
MF Extractor
Multimodal Features (MFs)
Fusion
Emotion Recognition
Easily Affected by Noises
(a) Existing fusion-based pipeline
Text
It was very romantic.
Noisy Information without Useful Semantics
Audio
NRGF Extractor
MF Extractor
Noise-Resistant Generic Feature (NRGF)
Noise-aware Learning Scheme
MFs
Multimodal Fusion Transformer
Emotion Recognition
Robust to Noises
(b) Our proposed noise-resistant pipeline

Fig. 1 Our proposed multimodal emotion recognition methodology, i.e., Noise-Resistant Modality Transformer (NORM-TR) (as shown in **b**), compared to common multimodal fusion methodology (as shown in **a**). Using multimodal input that may contain noisy information with irrelevant useless semantics, existing multimodal emotion recognition method that directly fuses Multimodal Features (MFs) can be easily affected by the noises. Alternatively, we make our NORM-TR learn to extract Noise-Resistant Generic Feature (NRGF) with the help of a novel noise-aware learning scheme. Then, by using a multimodal fusion Transformer to make MFs complement the NRGF, we obtain much more robust multimodal emotion recognition results with our NORM-TR

to calculate the relationship between different modalities to obtain satisfactory performance. Since the computational complexity of the Cartesian product increases dramatically with the feature dimension and the number of modalities, its performance easily suffers from degradation if not using overwhelmingly large model capacities. LMF (Zadeh et al., 2017) introduced low-rank matrix factorization operation to reduce the computational cost. However, LMF tends to reduce useful information, resulting in a decrease in model performance. The attention-based fusion methods mainly employ attention learning mechanisms to make multimodal information interact with each other. For instance, Zhou et al. (2021) introduced attention learning to automatically calculate the importance weights of audio and video modalities so that obtaining effective emotion-related information. Zhao et al. (2020) proposed a new attention-based VAANET that integrated spatial, channel-wise, and temporal attentions for audio-video emotion recognition. Overall, although attention-based methods achieve progresses by weighting the importance of modalities for effective fusion, they may be still affected by noises inherent within each modality without explicitly depicting noisy information.

Despite current progress in fusion strategies, we argue that alleviating the negative impacts of noisy information is also important. More specifically, we observe that, in a multimodal sequence, there could be plenty of information that shows little relevance to emotion understanding, which can be viewed as noisy information. For example, the background sounds in audio data are irrelevant to the human who smiles in the corresponding video. As a result, modeling the trivial information of these background sounds would likely affect the multimodal fusion and the final understanding performance. In our experiments, we can show that noisy information greatly degrades emotion recognition accuracy, which further implies that being insensitive to noises can be beneficial for accurate emotion understanding. However, to the best of our knowledge, current literature on multimodal emotion understanding lacks sufficient study on noisy information, thus still obtaining sub-optimal performance.

In light of the above issue, we propose a novel Noise-Resistant Multimodal Transformer (NORM-TR) to address the adverse effects of noisy information on multimodal emotion recognition. The motivation of NORM-TR and the comparison to existing fusion methods are shown in Fig. 1. In general, we make the NORM-TR learn to extract a Noise-Resistant Generic Feature (NRGF) and then apply a Transformer (Vaswani et al., 2017) to incorporate Multimodal Features (MFs) extracted from the multimodal input according to their relations to the NRGF, thus obtaining more robust and more accurate emotion understanding results against noises. More specifically, we tend to formulate the NRGF to be generic and insensitive to the disturbances caused by noises. To obtain the NRGF, we employ an NRGF extractor and make it learn to summarize meaningful semantics from multimodal data. The extracted NRGF can provide a robust representation against noisy information, which also runs the risk of being insensitive to useful details for accurate emotion recognition. Therefore, we further introduce a multimodal fusion Transformer with NRGF serving as query and MFs serving as key and value; therefore, the relations between MFs and NRGF are reasoned, and the MFs can complement NRGF, achieving robust and accurate emotion recognition predictions.

To train the NORM-TR effectively, we first apply normal emotion recognition losses to make it learn to estimate human emotions. Meanwhile, we further apply our proposed noise-aware learning scheme to help make our model become robust against noises. Our noise-aware learning scheme explicitly adds noises to the input and applies adversarial losses to train the NRGF extractor in NORM-TR. Adding explicit noises can provide definite information about when noisy information occurs in the input, which can facilitate the discriminators of the related adversarial losses to be able to distinguish whether a feature contains noises. Specifically, two manners of adding noisy information are involved: (1) we make the added noisy information appear in all the modalities randomly, and (2) we add the noisy information to only a specific modality at some random periods. Two adversarial losses are introduced regarding both types of added noises, respectively. By fooling the discriminator that distinguishes the first type of added noises, we can make the NRGF extractor focus on more generic features against noises in all the multimodal data. Similarly, fooling the discriminator on the second type of added noises can make the extracted NRGF more robust against noises in each specific modality. Together with the emotion recognition loss, our proposed noise-aware learning scheme helps obtain robust NRGF and facilitates our NORM-TR model to achieve favorable multimodal emotion recognition performance.

In summary, the major contributions of the paper can be described as:

- We present a novel comprehensive study on noisy information for the multimodal emotion understanding task. To achieve more robust emotion understanding performance, we introduce the Noise-Resistant Multimodal Translator (NORM-TR) to extract Noise-Resistant Generic Features (NRGF) and significantly reduce the negative impacts of noise in the multimodal data.
- Based on the NRGF, we devise a novel Transformer-based end-to-end pipeline for multimodal emotion recognition. Besides, a novel noise-aware learning scheme is further designed to help optimize the NORM-TR appropriately.
- In practice, we demonstrate that the NORM-TR is effective in obtaining noise-invariant representations. Furthermore, our extensive experimental analysis of different popular datasets also illustrates that the NORM-TR significantly improves emotion recognition accuracy and achieves state-of-the-art performance by using the NRGF to alleviate the adverse impacts brought by noisy information, indicating the importance of handling noisy information.

2 Related Work

2.1 Multimodal Emotion Recognition

Multimodal emotion recognition aims to predict human emotion from multiple modalities, such as video, audio, and text. Most existing methods (Hazarika et al., 2020; Zhao et al., 2020; Tsai et al., 2019; Yang et al., 2022; Sun et al., 2020; Tsai et al., 2019) mainly focus on how to learn and fuse multimodal emotion representations from data of different modalities by considering the difference and consistency of different modalities. For instance, Hazarika et al. (2020) proposed a multimodal representation learning method for modality-invariant and -specific subspace projection. Moreover, to learn the interactive information between different modalities, recent increasing work has concentrated on multimodal fusion mechanisms, where many elaborate multimodal fusion methods have been proposed. Tsai et al. (2019) introduced Multimodal Factorization Mode (MFM) to explore intra-modal and cross-modal interactions by decomposing the modality representation into two independent sets of factors. Sun et al. (2020) proposed Interaction Canonical Correlation Network (ICCN) that used Canonical Correlation Analysis (CCA) to model the relationship between audio-text and video-text modalities. Lv et al. (2021) proposed the Progressive Modality Reinforcement (PMR) approach to conduct multimodal fusion by considering the three-way interactions across all the involved modalities.

Although progress, current methods rarely focus on the noise problem in multimodal emotion recognition. Yuan et al. (2021) also pointed out that multimodal data contain a large amount of noise, such as missing data in some modal sequences, which can greatly degrade the results of multimodal fusion methods. How to effectively reduce the noise effect on multimodal data remains an open problem.

2.2 Adversarial Learning

Adversarial learning is widely used in domain adaptation learning (Ganin & Lempitsky, 2015; Pei et al., 2018; Wang et al., 2020; He et al., 2022) and cross-modal retrieval (Wang et al., 2017; Li et al., 2018), etc. Recently, to improve the effectiveness of fusion, adversarial learning is increasingly used in multimodal emotion recognition by learning common subspace representations. Yang et al. (2022) employed FDMER with adversarial learning to mine the commonality and diversity of different modalities, achieving an impressive performance for multimodal emotion recognition. Despite the progress, FDMER did not consider the side effect of noise on the robustness of the model. To address this limitation, we propose the NORM-TR with a novel noise-aware adversarial learning to extract Noise-Resistant Generic Features (NRGF), thereby greatly reducing the negative impact of noisy information and improving the robustness of the multimodal fusion.

2.3 Transformer

Transformer is an attention-based building block for machine translation introduced by Vaswani et al. (2017). By aggregating data from the whole sequence, Transformer can learn the relationships between tokens scanned over time, replacing RNNs for a variety of tasks, such as natural language processing (Kenton & Toutanova, 2019; Ding et al., 2021), computer vision (Zhang et al., 2022; Liu et al., 2021), as well multimodal emotion recognition (Hazarika et al., 2020; Liang et al., 2020; Tsai et al., 2019; Huang et al., 2020; Yuan et al., 2021; Liu et al., 2022). Tsai et al. (2019) introduced the Multimodal Transformer (MulT) to address modal data misalignment and long-distance dependencies. Huang et al. (2020) utilized the Transformer to fuse audio-visual information on the model level, showing the superiority of model-level fusion over other layers of fusion strategies. Yuan et al. (2021) proposed a Transformer-based feature reconstruction network to achieve more robust multimodal emotion recognition. Despite the progress, existing Transformer-based methods mainly consider the feature from a specific modality as the query, which could introduce unnecessary noises or trivial information related to the modality.

2.4 Multi-task Learning

Multi-task learning seeks to enhance the ability of a model to generalize across various interconnected tasks by harnessing the collective knowledge gathered from the ensemble of tasks (Zhang & Yang, 2022). This approach is gaining traction in various fields (Bousmalis et al., 2016; Niu et al., 2020; Liu et al., 2023). For example, Bousmalis et al. (2016) proposed an novel approach by designing several tasks to disentangle feature representations of source and target domains, thereby enhancing the model's generalization ability across different domains. Niu et al. (2020) introduces multi-tasks for remote physiological measurement, employing a cross-verified feature disentangling strategy to simultaneously estimate multiple physiological signals. In the field of multimodal emotion recognition, increasing number of works (Liu et al., 2018; Hazarika et al., 2020; Wang et al., 2022) introduce multi-task learning to achieve more robust emotion recognition performance. For instance, Liu et al. (2018) designed self-learning based uni-modal tasks to learn the consistency and difference between each modality. Wang et al. (2022) proposed a multi-task learning framework MT-TCCT, which enhances the performance of modality-private and modality-shared representations by leveraging the interdependence of sub-tasks. Overall, existing works are almost no specific design of subtasks for emotion-irrelevant noisy, which may lead to the model's performance degradation in the face of different levels of noise effects.

In this work, we introduce a noise-aware learning scheme based on multi-task learning and adversarial learning, which has facilitated the model's ability to perceive emotion-irrelevant noise, thus improving the robustness of the model in recognizing emotions.

3 Method

3.1 Overview

The overall processing pipeline of the proposed Noise-Resistant Multimodal Transformer (NORM-TR) for the robust emotion recognition is shown in Fig. 2. We make the NORM-TR first extract Noise-Resistant Generic Features (NRGFs) and Multimodal Features (MFs) from the input. Then, a multimodal fusion Transformer is employed to integrate the MFs according to their relations to the NRGFs, thus obtaining an end-to-end noise-resistant model for emotion understanding. To train the NORM-TR properly, we introduce a noise-aware learning scheme. In our learning scheme, we manually erase some certain periods of information in the multimodal input to implement the explicit inclusion of noisy information, which does not contain any useful semantics. By explicitly adding noisy information to either all the multimodal inputs or the input data from a specific modality, we devise two adversarial learning objectives to make the NORM-TR robust against both types of added noisy information.

Formally, our NORM-TR employs an NRGF extractor, denoted as $\mathcal{F}_{NR}$, and an MF extractor, denoted as $\mathcal{F}_M$, to extract detailed NRGFs and MFs from its input U', respectively. Then, a multimodal fusion Transformer, $Trans(\cdot)$, translates the MFs to the desired outputs $\hat{y}_E$ for emotion recognition according to the NRGFs. Therefore, our overall pipeline can be described by:

$$\hat{y}_E = Trans\Big(\mathcal{F}_{NR}(U'), \mathcal{F}_M(U')\Big), \tag{1}$$

where $\hat{y}_E$ is the emotion recognition output, and the U' is the noise-corrupted multimodal input. It is worth mentioning that our NORM-TR, as described in Eq. 1, also works for the normal multimodal input without explicitly added noises.

3.2 Noise-Corrupted Multimodal Input

Regarding the normal multimodal emotion recognition, we use the symbol U to represent the provided input multimodal information of a sequence. The U can represent the input of audio, video, text, etc. In the rest of paper, we use the U_a, U_v, and U_t to represent the audio, video, and text, respectively. In the related literature (Tsai et al., 2019; Hazarika et al., 2020; Mao et al., 2022), *pre-computed features* rather than raw data of different modalities are commonly used, thus, to be fair, the U in our paper represents the pre-computed feature vectors. For example, rather than using 2D images of a video, we can have as input the pre-computed features $U_v \in \mathbb{R}^{T \times N_v}$ where T represents the length of the video, and N_v represents the length of the feature vector. Correspondingly, for audio and text modalities, we can also have as input the pre-computed features $U_a \in \mathbb{R}^{T \times N_a}$ and $U_t \in \mathbb{R}^{T \times N_t}$, respectively, where N_a and N_t are the lengths of the corresponding feature vectors, and T is the unified the length of the feature vector of each modality. We would like to mention that using pre-computed features is widely accepted in the literature on multimodal emotion recognition,

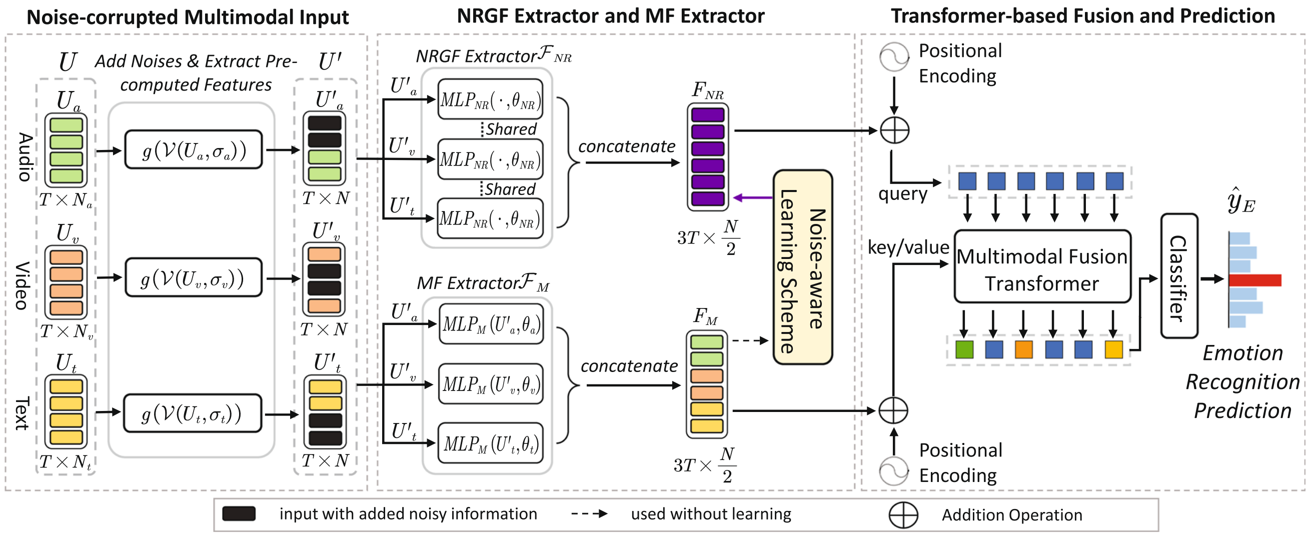

Noise-corrupted Multimodal Input
NRGF Extractor and MF Extractor
Transformer-based Fusion and Prediction
U
Add Noises & Extract Pre-computed Features
U'
Audio
Video
Text
U_a
U_v
U_t
$T \times N_a$
$T \times N_v$
$T \times N_t$
$g(\mathcal{V}(U_a,\sigma_a))$
$g(\mathcal{V}(U_v,\sigma_v))$
$g(\mathcal{V}(U_t,\sigma_t))$
U'_a
U'_v
U'_t
$T \times N$
NRGF Extractor $\mathcal{F}_{NR}$
$MLP_{NR}(\cdot,\theta_{NR})$
Shared
MF Extractor $\mathcal{F}_M$
$MLP_M(U'_a,\theta_a)$
$MLP_M(U'_v,\theta_v)$
$MLP_M(U'_t,\theta_t)$
concatenate
F_{NR}
F_M
$3T \times \frac{N}{2}$
Noise-aware Learning Scheme
Positional Encoding
query
key/value
Multimodal Fusion Transformer
Classifier
$\hat{y}_E$
Emotion Recognition Prediction
input with added noisy information
used without learning
Addition Operation

Fig. 2 Processing pipeline of the proposed NORM-TR for multimodal emotion recognition. With the noise-corrupted multimodal input U', we first apply a noise-resistant generic feature (NRGF) extractor to extract a generic and disturbance-insensitive representation. Then, we employ a multimodal fusion Transformer to improve NRGFs based on the more detailed multimodal features (MFs). Moreover, we introduce a novel noise-aware learning scheme to properly train the NORM-TR, thus obtaining an end-to-end noise-resistant model for emotion understanding

and the research on more appropriate features to describe multimodal raw data is beyond the scope of this study. With the pre-computed features $U = \{U_a, U_v, U_t\}$, we employ three fully-connected layers to unify the lengths of different feature vectors, respectively. We use the symbol g to represent these fully-connected layers. After g, each modality of the obtained $U' = \{U'_a, U'_v, U'_t\}$ would have a dimension of $T \times N$, where N is the unified length of feature vectors. In the meantime, we use the $\mathcal{V}(U, \sigma)$ to represent the process that explicitly adds the noisy information σ to the input U. Therefore, the noise-corrupted multimodal input to our NORM-TR can be formulated as:

$$U' = g(\mathcal{V}(U, \sigma)). \tag{2}$$

The detailed formulation of our proposed NORM-TR, such as the NRGF extractor, the MF extractor, the multimodal fusion Transformer, and the noise-aware learning scheme, will be described in the following sections, subsequently.

3.3 NRGF Extractor and MF Extractor

3.3.1 NRGF Extractor

With multimodal input U', we introduce an NRGF extractor to obtain a generic feature that is insensitive and invariant to noisy information. We achieve this by simply employing a multi-layer perceptron:

$$F_{NR} = \mathcal{F}_{NR}(U') = MLP_{NR}(U', \theta_{NR}), \tag{3}$$

where F_{NR} is the obtained NRGF, MLP_{NR} represents multi-layer perceptrons with shared parameters for extracting NRGF. In practice, MLP_{NR} is designed as a 2-layer fully connected network, with each layer followed by Leaky-ReLU (Maas et al., 2013) as the activation function. Using the MLP_{NR}, we reduce the feature dimension of U'. More specifically, if considering 3 modalities, the U' obtained by Eq. 2 has the dimension of $3T \times N$, and the F_{NR} will have a dimension of $3T \times \frac{N}{2}$. In fact, we found that reducing the feature dimension can save parameters and achieve higher efficiency without sacrificing performance.

To make the NRGFs effective for extracting noise-insensitive features, we introduce the noise-aware learning scheme, which will be described in detail in the Sect. 3.5.

3.3.2 MF Extractor

In addition to the NRGF extractor, we also introduce the MF extractor to extract and exploit detailed multimodal features from the input. This is because we found that the NRGF may lose some details that are beneficial for accurate emotion understanding. To implement the MF extractor, we still use a multi-layer perceptron, thus we have:

$$F_M = \mathcal{F}_M(U') = MLP_M(U', \theta_*). \tag{4}$$

MLP_M represents three separated multi-layer perceptrons for MF extraction, where each multi-layer perceptron architecture in MLP_M is the same as MLP_{NR}. $* \in \{a, v, t\}$. Therefore, given 3 modalities as input, the F_M will have the size of $3T \times \frac{N}{2}$. We would like to mention that the input feature U' is shared with both NRGF extractor and MF extractor. In addition, we do not make the $\mathcal{F}_M$ learn to become consistent with noise-caused changes, making it more sensitive to the variations in the multimodal input.

3.4 Transformer Structure and Emotion Recognition Output

With the obtained NRGFs and MFs, we employ a multimodal fusion Transformer to achieve effective multimodal emotion recognition. The Transformer can model the relations between NRGFs and MFs, which can make MFs more appropriately complement the NRGFs based on their relevance to the NRGFs. We found that this relation modeling is important because the MFs could be more affected by noisy information, and directly fusing them with the NRGFs would introduce the negative impacts of noises (with more details in 4.7.1).

3.4.1 Multimodal Fusion Transformer

According to the definition of Vaswani et al. (2017), a Transformer takes as input the query, key, and value tensors. Then, it uses the query tensor as a reference and transforms the value tensor into desired output based on the relations between the query tensor and key tensor. When transforming, multihead attention mechanisms are performed to achieve the relation modeling and data fusion. For more details, we refer readers to Tsai et al. (2019). In NOMR-TR, NRGFs is specifically designed to minimize the impact of emotion-irrelevant noisy, and NRGFs also need the more detailed information from MFs to capture modality-sensitive information and achieve more promising performance. To reduce the possibility of introducing noise to the fused feature again, we apply a Transformer decoder structure which makes NRGF as query and MFs as key/value data. Using this formulation, the MFs will be fused *w.r.t.* NRGF according to attentional weights. Since NRGF is noise-resistant, low attentional correlation between NRGF and MFs would indicate irrelevant information, i.e., noisy information defined in our paper. As a result, our Transformer-based fusion method would minimize the chance of introducing noisy information again during the fusion.

In our study, using the extracted NRGFs as query and MFs as key and value, we follow the typical formulation of the Transformer structure for more effective multimodal fusion, and implement the $Trans(\cdot)$ as:

$$Trans(\cdot) = Trans(q = F_{NR}, k/v = F_M), \tag{5}$$

where q, k, and v represent the query, key, and value tensors in a Transformer, respectively. More specifically, the query, i.e., NRGFs extracted by $\mathcal{F}_{NR}$, have a shape of $3T \times \frac{N}{2}$. The key and value tensors share the same multimodal feature obtained by $\mathcal{F}_M$, which have a shape of $3T \times \frac{N}{2}$. The $Trans(\cdot)$ contains eight attention heads and attention blocks of different depths on different datasets, e.g., 2 levels of depth for MOSI (Zadeh et al., 2016) and RML (Wang & Guan, 2008) datasets, and 4 levels of depth for MOSEI (Zadeh et al., 2018) and IEMOCAP (Busso et al., 2008) datasets. This is because using more depths for smaller datasets like MOSI and RML can easily result in over-fitting. In addition, following Dosovitskiy et al. (2021), we perform a learnable positional encoding of timestamp of the sequence and add it to the input of the Transformer. It is worth mentioning that we have tried more complicated structures for fusion, but it does not improve performance quite much (with more details in 4.7.4). We found that deep Transformer structures can also overfit the dataset easily.

3.4.2 Emotion Recognition Output

After the multimodal fusion Transformer, we obtain the final emotion recognition output $\hat{y}_E$ by applying an emotion classification layer on the outputs of Transformer. Specifically, the Transformer outputs a tensor of shape $3T \times \frac{N}{2}$ after relation modeling and tensor fusion. Then, we apply an average pooling operation on the obtained tensor to reduce its dimension from $3T \times \frac{N}{2}$ to $1 \times \frac{N}{2}$ for relieving the computational burden. The pooled tensor is later fed into a fully connected layer for emotion classification. Lastly, we have the $\hat{y}_E$ of a shape $1 \times C$, where C represents the number of categories. Each element of the $\hat{y}_E$ represents a specific emotion like happiness and angry. In general, with the help of the NRGF extractor and the Transformer, we obtain an accurate estimation $\hat{y}_E$ that is much less affected by noisy information.

3.5 Noise-Aware Learning Scheme

By devising the NORM-TR, it is essential to apply appropriate learning objectives to help our model learn to resist noisy information. Therefore, we introduce a novel noise-aware learning scheme for training the NORM-TR and making it robust to noises. Our novel learning scheme explicitly adds two types of noisy information to corrupt raw multimodal data and then introduces two corresponding adversarial losses to encourage the NORM-TR to provide an NRGF invariant to the added noisy information.

3.5.1 Explicit Noisy Information

We explicitly add noisy information as described in Eq. 2 because it is extremely difficult to properly define what patterns should belong to noisy information and what should not. Without explicit noisy information, we would not know when a model should be insensitive to the changes of patterns

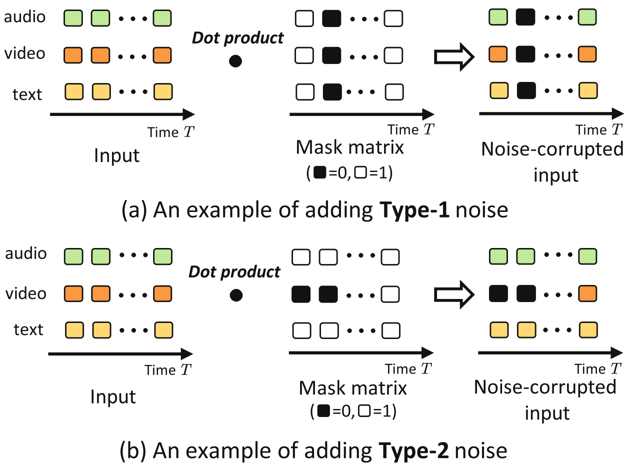

audio
video
text
Time T
Input
Dot product
Mask matrix
(■=0,□=1)
Noise-corrupted input
(a) An example of adding **Type-1** noise
(b) An example of adding **Type-2** noise

Fig. 3 Examples of adding noises of two types. **a** Adding **Type-1** noise, **b** adding **Type-2** noise

in the input sequence. Therefore, we attempt to explicitly add noisy information to the multimodal input. Regarding this, we randomly erase some certain periods of input multimodal representation to remove any potential semantics in the input data. In this study, we mainly consider two types of noisy information: (**Type-1**) all the input multimodal data contains random noisy information (see Fig. 3a); and (**Type-2**) only the input data of a specific modality contains random noisy information (see Fig. 3b). Training on the first case can help the NORM-TR summarize generic and globally consistent semantics, and the second case can help the NORM-TR focus on improving its robustness against noisy information to each specific modality. Our experimental results can validate the importance of adding both types of noisy information. In practice, we masking-out input information to implement the Eq. 2. For example, when adding the Type-1 noise, we generate a mask for the pre-computed multimodal feature vector of each modality. In each mask, we randomly sample a time window whose length ranges from 0 to $\frac{T}{2}$ time steps, and then we set the values within this time window to 0 with the remaining values of this mask being 1. Thus, each modality has its individual mask for processing. Then, we multiply these masks with the pre-computed multimodal feature vectors U, erasing the semantics contained in the period corresponding to the sampled time window of this mask. In addition, for adding the Type-2 noise, the semantic erasing operations are similar to the procedure of adding the Type-1 noise, but the mask generation is different. In this case, we only generate one mask, in which a time window containing 0 s is sampled randomly. We multiply this mask to the pre-comupted feature of a random modality, thus adding the noise to this modality only.

As shown in Fig. 4, after adding noisy information, we use two adversarial loss functions to define learning objectives regarding both types of added noisy information, respectively. We apply adversarial loss functions because they are powerful at making a model produce the high-dimensional output with some desired patterns, e.g., producing a fake 2D image that has very similar image patterns to the desired real 2D images. To implement the adversarial-based learning, a discriminator is employed to learn to distinguish whether the generated output fails to obtain the desired patterns. The unsuccessfully generated output will be easily identified by the discriminator, while the successfully generated output will confuse the discriminator. The detailed generative adversarial learning scheme can be found in Goodfellow et al. (2014).

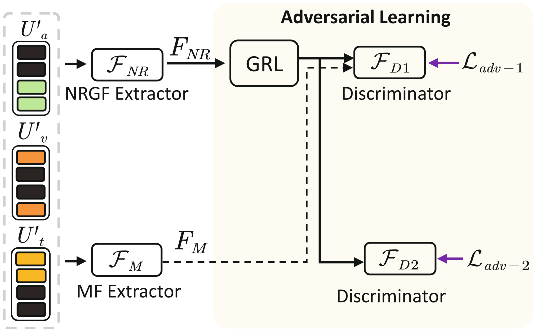


Fig. 4 The pipeline of the noise-aware learning scheme

3.5.2 Adversarial Learning-Based Learning Scheme

Here, we attempt to use adversarial losses to make our NORM-TR learn to produce the NRGF that is invariant to noises. By resisting the impacts of the added noisy information via adversarial learning, the NRGF extractor would be more effective at extracting semantically meaningful features from the input data, which can further improve the robustness of emotion recognition. Formally, we use the symbols $\mathcal{F}_{D1}$, $\mathcal{F}_{D2}$ to represent the discriminators of Type-1 and Type-2 added noises, respectively. Figure 4 shows the detailed adversarial learning structure of our noise-aware learning scheme.

More specifically, since the Type-1 noises are randomly added to all the multimodal input, the NRGF extractor is supposed to find generic and universal patterns from the noise-corrupted multimodal input, so that the extracted NRGF can be invariant to the Type-1 noises. To achieve this, we employ a discriminator $\mathcal{F}_{D1}$ and make it learn to classify whether its input feature contains modality-dependant patterns: if the input to $\mathcal{F}_{D1}$ is a modality-specific feature from the MFs F_M, we make this discriminator learn to predict which modality this feature represents; if the input to $\mathcal{F}_{D1}$ is the NRGF, we make the NRGF extractor confuse this discriminator. Therefore, when the NRGF extractor fails to provide generic and consistent features, it would not confuse the discriminator effectively. By confusing the discriminator, the extracted NRGF would be representative for the entire multimodal input rather than a specific modality that can be more likely to be affected by the added Type-1 noise. To sum up, suppose the model parameter of NRGF extractor is θ_{NR}, and the model parameters of the discriminator $\mathcal{F}_{D1}$ are θ_{D1}. Then, the first objective of our noise-aware learning scheme can be described by the optimization *w.r.t.* the adversarial losses $\mathcal{L}_{adv-1}$ for Type-1 added noisy information:

$$\min_{\theta_{D1}} \max_{\theta_{NR}} \mathcal{L}_{adv-1} = -\frac{1}{N_b} \sum_{i=0}^{N_b} y_M^i \cdot log\ \mathcal{F}_{D1}(F_{NR}/F_M; \theta_{D1}), \tag{6}$$

where N_b is the number of samples in the training set, y_M^i represents the label indicating which modality the F_{NR} or F_M comes from, and F_{NR} is the NRGF extracted according to the parameter θ_{NR}. To make the NRGF compatible to $\mathcal{F}_{D1}$ that identifies the modality-dependent features, we add 3 extra MLPs to explain the F_{NR} of $3T \times \frac{N}{2}$ into the features related to 3 input modalities, respectively, with a shape of $3 \times (T \cdot \frac{N}{2})$.

Regarding the Type-2 noises, we also employ a discriminator $\mathcal{F}_{D2}$ to help train the NRGF extractor. Different from the Type-1 noises, we add noises to only one random modality. Learning against the Type-2 noises can help the NORM-TR works effectively on the input without adding explicit noises while still maintaining its capability of being robust against noises existing in a random modality. Regarding the Type-2 noises, we simply make the discriminator $\mathcal{F}_{D2}$ identify whether a modality contains noises: if the input to $\mathcal{F}_{D2}$ is related to the input without added noises, we make this discriminator predict a negative label; otherwise, we make this discriminator predict a positive label. Therefore, if the NRGF fails to be invariant to the noises only added to a specific modality, the NRGF extractor would also fail to confuse the discriminator and the $\mathcal{F}_{D2}$ can easily identify the noises by predicting a positive label. To this end, suppose the model parameters of the $\mathcal{F}_{D2}$ are θ_{D2}. We have the second objective of our noise-aware learning scheme as the optimization *w.r.t.* the adversarial losses $\mathcal{L}_{adv-2}$ for Type-2 added noisy information:

$$\min_{\theta_{D2}} \max_{\theta_{NR}} \mathcal{L}_{adv-2} = -\frac{1}{N_b} \sum_{i=0}^{N_b} y_N^i \cdot log\ \mathcal{F}_{D2}(F_{NR}; \theta_{D2}), \tag{7}$$

where y_N^i is the label indicating which modality is corrupted by added noises.

In practice, we use the fully-connected layer to implement the two discriminators, each of which consists of a fully-connected layer. In addition, we apply the gradient reversal layer (GRL) (Ganin & Lempitsky, 2015) to implement the adversarial learning *w.r.t.* NRGF extractor that is supposed to confuse the two discriminators.

3.6 Overall Learning Objectives

To sum up, our method involves three learning objectives, including two adversarial loss functions $\mathcal{L}_{adv-1}$ and $\mathcal{L}_{adv-2}$, and one final emotion learning loss $\mathcal{L}_{er}$. Considering that the emotion labels on different datasets are different, for example, the labels on the RML (Wang & Guan, 2008) and IEMOCAP (Busso et al., 2008) datasets are discrete, while the labels on the MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2018) datasets are continuous. Therefore, we introduce the cross-entropy loss as the emotion learning loss $\mathcal{L}_{er}$ for classification on the RML and IEMOCAP datasets, and the mean squared error (MSE) as $\mathcal{L}_{er}$ for regression on the MOSI and MOSEI datasets. The emotion learning loss $\mathcal{L}_{er}$ can be written as:

$$\mathcal{L}_{er} = \begin{cases} -\frac{1}{N_b} \sum_{i=0}^{N_b} y^i \cdot \log \hat{y}_E^i & \text{for classification} \\ \frac{1}{N_b} \sum_{i=0}^{N_b} \left\| y^i - \hat{y}_E^i \right\|_2^2 & \text{for regression} \end{cases} \tag{8}$$

where y^i is the emotion label of the i-th sample. $\hat{y}_E^i$ is the prediction of NORM-TR. The overall objective function $\mathcal{L}$ is the sum of $\mathcal{L}_{adv-1}$, $\mathcal{L}_{adv-2}$, and $\mathcal{L}_{er}$. Mathematically, the $\mathcal{L}$ can be written as:

$$\mathcal{L} = \alpha \mathcal{L}_{adv-1} + \beta \mathcal{L}_{adv-2} + \gamma \mathcal{L}_{er}. \tag{9}$$

In our experiments, on MOSI and MOSEI datasets, $\alpha = 0.01$, $\beta = 0.01$ and $\gamma = 1$; on RML and IEMOCAP datasets, $\alpha = 0.005$, $\beta = 0.005$ and $\gamma = 1$.

4 Experiment

4.1 Datasets

We conducted extensive experiments on three trimodal datasets (MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018), and IEMOCAP (Busso et al., 2008)), as well as a bimodal emotion dataset (RML (Wang & Guan, 2008)). These datasets cover different languages and various scenarios like natural and laboratory scenes, dialogue, and solo presentations.

MOSI The MOSI dataset consists of 2,199 multimodal sequence samples with video, audio, and text modalities. The training, validation, and testing sets of MOSI contain 1,284 samples, 229 samples, and 686 samples, respectively. Each multimodal sample has a uniform label that ranges from -3 to 3. The -3 and +3 represent strongly negative and strongly positive emotions, respectively.

MOSEI The MOSEI dataset comprises 22,851 video clips collected from YouTube with spontaneous expressions, head poses, occlusions, illuminations, and so on. This dataset is divided into 16,326 training samples, 1,871 validation samples, and 4,659 test samples in speaker-independent settings. Each sample is manually annotated with a sentiment score from -3 (strongly negative) to 3 (strongly positive).

IEMOCAP The IEMOCAP includes video, audio, and text transcriptions and 12 h of video recordings of situational dialogues. The data is divided into five sessions with a total of 10,039 samples and 9 emotion categories. Following the comparison methods (Tsai et al., 2019; Lv et al., 2021), we used the four emotion categories, i.e., happiness, anger, sadness, and neutral. The data is partitioned into 2717 training samples, 798 validation samples, and 938 testing samples.

RML The RML is an audio-visual emotion dataset, containing 8 topics with 720 video samples, with six different languages (i.e., English, Mandarin, Urdu, Punjabi, Persian, and Italian). Each data was labeled as one of six emotions, i.e., anger, disgust, fear, happiness, sadness, and surprise. The training and testing sets were in a ratio of about 8:2 for cross-validation with speaker-independent settings, which ensures that the speakers in the training set were not in the corresponding test set.

4.2 Implementation Details

We used PyTorch to implement our method. The experiments were conducted on a PC with Intel(R) Xeon(R) Gold 6240C CPU at 2.60 GHz and 128 GB memory and NVIDIA GeForce RTX 3090. The key training parameters include initial learning rate (0.0001), cosine annealing schedule to adjust the learning rate, mini-batch size (16), and warm up.

For the sequence length setting, we unified the length of the sequences to 8 on the RML and IEMOCAP datasets (i.e., $T = 8$), and 50 on the MOSI and MOSEI datasets (i.e., $T = 50$). The generation of video modality is different from audio and text modalities. More specifically, for the video modality, we divided the input video into T segments and randomly sampled one frame from each of the segment to form a video sequence of length T. For the audio and text modalities, we directly truncated the first T frames of the data as the input sequence.

4.3 Pre-computed Feature Extraction

Video features U_v: For the RML and IEMOCAP datasets, following the existing method (Zhao et al., 2020), we employed a ResNet-18 (He et al., 2016) to extract the last global averaging pooling output of the ResNet-18 as the pre-computed video features. For the MOSI and MOSEI datasets, referring to the existing methods (Tsai et al., 2019; Hazarika et al., 2020; Mao et al., 2022), we used the features provided in the dataset, which had been extracted by the OpenFace (Baltrusaitis et al., 2016).

Audio features U_a: For RML and IEMOCAP, we first used Librosa to compute the log mel-spectrogram and its first and second-order differentials of each sample, and then employed a ResNet-18 (He et al., 2016) to extract features. Finally, we stacked all feature vectors and obtained the pre-computed audio features. For MOSI and MOSEI, we used the features provided by the dataset, which were extracted by the Librosa.[1]

Text features U_t: For IEMOCAP, we used a pre-trained BERT (Kenton & Toutanova, 2019) as the feature extractor to encode the transcribed word sequences into the pre-computed text features. For MOSI and MOSEI datasets, we also used the text features provided by the dataset, which were extracted by BERT.

4.4 Evaluation Metrics

On the RML dataset, we chose two widely-used evaluation metrics, i.e., six classification accuracy (Acc-6) and weighted F-Score (F1) to evaluate the performance. On the IEMOCAP dataset, we followed previous works (Lv et al., 2021) to report the binary classification accuracy (Acc-2) and weighted F1 for each emotion category.

On MOSI and MOSEI, referring to prior works (Yu et al., 2020), we used six widely-used evaluation metrics: Acc-2, weighted F1, seven classification accuracy (Acc-7), mean absolute error (MAE), and the correlation of the model's prediction with human annotations (Corr). Specifically, following prior works (Hazarika et al., 2020; Yu et al., 2021), we calculated Acc-2 and F1 in two ways: negative/non-negative and negative/positive on MOSI and MOSEI datasets, respectively. Additionally, it should be noted that some works Franceschini et al. (2022); Mittal et al. (2020a) use MOSEI to study multi-label multimodal emotion recognition with six discrete basic expressions: Happiness, Sadness, Fear, Surprise, Disgust, and Anger. For a more comprehensive evaluation, although our work primarily focuses on Positive/Negative single-label emotion classification, we also evaluate our method using a multi-label emotion recognition metric on the MOSEI dataset (see Sect. 4.7.14).

To validate the robustness of our method to different intensities of noise, referring to previous work (Yuan et al., 2021), we introduced Area Under Indicators Line Chart (AUILC) to evaluate the noise robustness of our method on the test set. The AUILC can be written as:

$$\mathrm{AUILC} = \sum_{t} \frac{(x_t + x_{t+1})}{2} \cdot (r_{t+1} - r_t) \tag{10}$$

[1] https://librosa.org.

Table 1 Comparison results on MOSI dataset

Methods	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)
TFN (Zadeh et al., 2017)	34.9	–/80.8	–/80.7	0.901	0.698
LMF (Liu et al., 2018)	33.2	–/82.5	–/82.4	0.917	0.695
MFM (Tsai et al., 2019)	35.4	–/81.7	–/81.6	0.877	0.706
ICCN (Sun et al., 2020)	39.0	–/83.0	–/83.0	0.877	0.706
MuLT (Tsai et al., 2019)	40.0	–/83.0	–/82.8	0.871	0.698
MISA (Hazarika et al., 2020)	42.3	81.8/83.4	81.7/83.6	0.783	0.761
PMR (Lv et al., 2021)	40.6	–/83.6	–/83.4	–	–
Self-MM (Yu et al., 2021)	45.8	84.0/86.0	84.4/86.0	0.713	0.798
FDMER (Yang et al., 2022)	44.1	–/84.6	–/84.7	0.724	0.788
Our NORM-TR	**48.5**	**84.3/86.1**	**84.4/86.2**	**0.698**	**0.808**

For each evaluation metric, ↑ indicates the bigger the better while ↓ indicates the smaller the better. The best result is highlighted in bold

Table 2 Comparison results on MOSEI dataset

Method	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	MAE (↓)	Corr (↑)
TFN (Zadeh et al., 2017)	50.2	–/82.5	–/82.1	0.593	0.700
LMF (Liu et al., 2018)	48.0	–/82.0	–/82.1	0.623	0.677
MFM (Tsai et al., 2019)	51.3	–/84.4	–/84.3	0.568	0.717
ICCN (Sun et al., 2020)	51.6	–/84.2	–/84.2	0.565	0.713
MuLT (Tsai et al., 2019)	51.8	–/82.5	–/82.3	0.580	0.703
MISA (Hazarika et al., 2020)	52.2	83.6/85.5	83.8/85.3	0.555	0.756
PMR (Lv et al., 2021)	52.5	–/83.3	–/82.8	–	–
Self-MM (Yu et al., 2021)	53.5	82.8/85.2	82.5/85.3	0.530	0.765
FDMER (Yang et al., 2022)	54.1	–/86.1	–/85.8	0.536	0.773
Our NORM-TR	**54.6**	**84.3/86.6**	**84.5/86.6**	**0.529**	**0.778**

For each evaluation metric, ↑ indicates the bigger the better while ↓ indicates the smaller the better. The best result is highlighted in bold

where x_t and x_{t+1} represent the t-th and $t+1$-th evaluation results under masking percentages of r_t and r_{t+1}, respectively.

4.5 Overall Performance

4.5.1 Experiments on the MOSI Dataset

Table 1 lists the comparison results of our proposed method and state-of-the-art methods on the MOSI dataset. As shown in the table, the proposed NORM-TR achieved an improvement of 2.7% on the Acc-7 compared to the second best result obtained by Self-MM (Yu et al., 2021). Compared to the other Transformer-based method FDMER (Yuan et al., 2021), our method gained a relative improvement of 9.98% on the Acc-7. Moreover, we also achieved state-of-the-art performance on all other metrics, especially on the more difficult seven classification task. We attribute such a large improvement to the fact that the extracted noise-resistant features can help our NORM-TR suppress useless noisy information during the fusion process, thus improving multimodal emotion recognition.

4.5.2 Experiments on the MOSEI Dataset

Table 2 reports the comparison results of our method and state-of-the-art methods on the MOSEI dataset. Our NORM-TR achieved significant improvement at the Acc-2, F1, Acc-3, Acc-5, MAE, and Corr. Compared to these Transformer-based methods, namely FDMER (Yang et al., 2022), and MulT (Tsai et al., 2019), we achieved relative improvements in all metrics, e.g., with 0.92% on Acc-7 and 0.93% on F1, respectively. Achieving such superior performance on large-scale datasets with more complex scenarios demonstrates the ability of our NORM-TR to extract effective emotion information from various scenarios.

4.5.3 Experiments on the IEMOCAP Dataset

Table 3 shows the comparison results of our method and state-of-the-art methods, including MulT (Tsai et al., 2019),

Table 3 Comparison results on IEMOCAP dataset

Methods	Happiness		Sadness		Anger		Neutral		Average	
	Acc-2	F1	Acc-2	F1	Acc-2	F1	Acc-2	F1	Acc-2	F1
EF-LSTM	76.2	75.7	70.2	70.5	72.7	67.1	58.1	57.4	69.3	67.7
LF-LSTM	72.5	71.8	72.9	70.4	68.6	67.9	59.6	56.2	68.4	66.6
RAVEN (Wang et al., 2019)	77.0	76.8	67.6	65.6	65.0	64.1	62.0	59.5	67.9	66.5
MCTN (Pham et al., 2019)	80.5	77.5	72.0	71.7	64.9	65.6	49.4	49.3	66.7	66.0
MulT (Tsai et al., 2019)	84.8	81.9	77.7	74.1	73.9	70.2	62.5	59.7	74.7	71.5
PMR (Lv et al., 2021)	86.4	83.3	78.5	75.3	75.0	71.3	63.7	60.9	75.9	72.7
ScaleVLAD(Luo et al., 2021)	86.7	85.9	84.8	84.6	86.8	86.9	72.1	72.1	82.6	82.4
Our NORM-TR	**87.7**	**88.5**	**86.2**	**86.4**	**88.6**	**88.6**	**74.8**	**74.3**	**84.3**	**84.5**

Note: the best result is highlighted in bold

PMR (Lv et al., 2021), and ScaleVLAD (Luo et al., 2021), on the IEMOCAP dataset. It is observed that our proposed NORM-TR achieved the best performance, which demonstrates the superiority of our NORM-TR. Compared with the state-of-the-art method ScaleVLAD, our NORM-TR achieved a relative 2.06% and 2.55% improvements on the averaged Acc and F1, respectively. In addition, we also achieved best performance for all four categories on the binary accuracy corresponding F1.

4.5.4 Experiments on the RML Dataset

Table 4 reports the comparison results of our method and state-of-the-art methods on the RML dataset. Compared to the second best result obtained by MulT (Tsai et al., 2019), our NORM-TR achieved a relative 3.23% boost on the averaged accuracy. Compared to the Census-Transform proposed by Cornejo and Pedrini (2019), our NORM-TR achieved a greater relative improvement of 7.75%. It shows that our method effectively addresses the effect of noise information to improve the performance of Transformer.

4.6 Robustness Evaluation for Noisy Data

To further verify the robustness of our method to noisy data, we evaluated our method with test data under different masking percentages r_t. More specifically, we first obtain the performance of NORM-TR corresponding to varying values of $r_t \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, and based on this range of r_t and its corresponding performance x_t, we calculated the results using the AUILC formula as shown in Eq. 10. Table 5 reports the AUILC results of our method and state-of-the-art methods on the MOSI and MOSEI datasets, respectively. It is obvious that our NORM-TR achieved better performance on almost all metrics on the MOSI dataset, *i,e.,* 31.3% on Acc-7, 68.6% on Acc-2, 1.093 on MAE, and 0.506 on Corr, respectively. On the MOSEI dataset, our method also obtained significant improvements on all metrics. For example, compared to TFR-Net (Yuan et al., 2021), our method obtained a relative improvement of 1.72% on the Acc-7, 2.05%/2.41% on the Acc-2, 1.90% on the MAE, and 6.60% on the Corr, respectively. It demonstrates the great robustness of our method in the face of noise disturbances.

In addition, Fig. 5 shows the metric curves with the test data under various mask percentages on MOSI dataset. As shown in the figure, NORM-TR outperforms the other methods on almost all evaluation metrics at various mask percentages $r_t \in \{0, 0.1, \cdots, 0.8\}$, indicating that our NORM-TR achieves greater robustness to noisy.

4.7 Ablation Study and Analysis

4.7.1 Effects of Different Components

To better study the influence of each component in the proposed NORM-TR, Table 6 reports the ablation results of the subtraction of each component from the NORM-TR framework on the MOSI and MOSEI datasets, respectively. It is worth noting that the proposed NORM-TR has achieved state-of-the-art performance. As shown in the table, subtracting the NRGF or MF extractor decreases the accuracy to suboptimal performance, demonstrating the importance of both extractors for robust multimodal emotion recognition.

Specifically, although only removing the MF extractor (see the third row) results in a relatively minor performance drop on the MOSI dataset (only 0.4%), it should be emphasized that this small decrease highlights the dataset's size limitations. The MOSI dataset, with about 1,284 training instances, poses a significant challenge for capturing fine-grained patterns. In contrast, on the larger MOSEI dataset with about 16,326 training samples, the inclusion of MFs leads to a more noticeable improvement, underscoring their

Table 4 Comparison results on RML

Methods	Acc-6	F1
(El-Madany et al., 2016)	75.00	–
(Zhang et al., 2018)	80.36	–
(Ma et al., 2019)	80.46	–
(Cornejo & Pedrini, 2019)	82.50	–
TFN* (Zadeh et al., 2017)	83.19	83.22
MulT* (Tsai et al., 2019)	86.11	85.87
MISA* (Hazarika et al., 2020)	81.11	80.78
Our NORM-TR	**88.89**	**88.81**

The best result is highlighted in bold and * indicates that the result is reproduced by authors

Table 5 Model robustness comparison on MOSI and MOSEI datasets

Method	MOSI				MOSEI			
	Acc-7 (↑)	Acc-2 (↑)	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	MAE (↓)	Corr (↑)
TFN (Zadeh et al., 2017)	–	–/60.4	1.327	0.300	–	–/–	–	–
MulT (Tsai et al., 2019)	–	–/61.8	1.288	0.334	–	–/–	–	–
MISA (Hazarika et al., 2020)	–	–/63.2	1.209	0.403	–	–/–	-	-
TFR-Net (Yuan et al., 2021)	–	–/**69.0**	1.155	0.467	–	–/–	–	–
TFN* (Zadeh et al., 2017)	22.0	62.6/62.6	1.281	0.342	45.1	70.8/72.6	0.726	0.409
MulT* (Tsai et al., 2019)	22.7	63.5/64.0	1.265	0.339	46.4	74.5/75.5	0.692	0.504
MISA* (Hazarika et al., 2020)	27.0	65.4/65.4	1.181	0.412	44.5	73.3/74.5	0.720	0.421
TFR-Net* (Yuan et al., 2021)	26.8	67.8/68.2	1.175	0.445	46.5	73.3/74.7	0.686	0.515
Our NORM-TR	**31.3**	**68.6**/68.6	**1.093**	**0.506**	**47.3**	**74.8/76.5**	**0.673**	**0.549**

For each evaluation metric, ↑ indicates the bigger the better while ↓ indicates the smaller the better. The best result is highlighted in bold, * indicates that the result is reproduced by authors

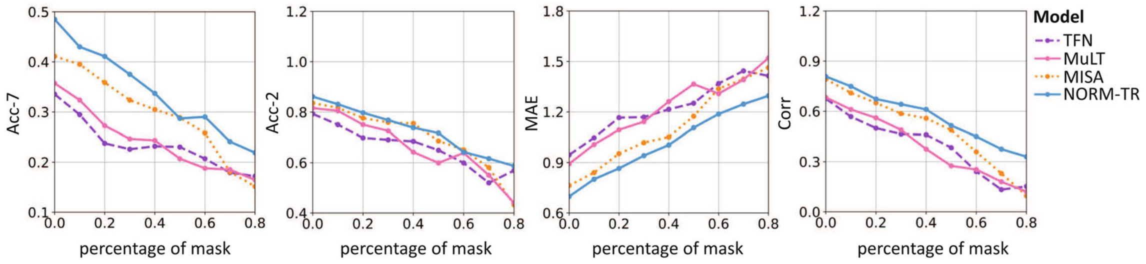


Fig. 5 Visualization of the metrics curves for test data under various mask percentages on the MOSI dataset

Table 6 Ablation study of the proposed NORM-TR

Method	MOSI		MOSEI	
	Acc-7 (↑)	MAE (↓)	Acc-7 (↑)	MAE (↓)
NORM-TR	**48.5**	**0.698**	**54.6**	**0.529**
w/o only NRGF extractor	46.8	0.726	52.3	0.547
w/o only MF extractor	48.1	0.706	53.6	0.534
w/o Multimodal fusion transformer	45.6	0.734	53.8	0.536
w/o Noise-aware learning scheme	46.9	0.721	53.2	0.544
Standard transformer (baseline)	***21.0***	***0.900***	***40.3***	***0.831***

For each evaluation metric, ↑ indicates the bigger the better while ↓ indicates the smaller the better. The best result is highlighted in bold
Each line '*w/o*' indicates the effect of subtracting this component in NORM-TR on the MOSI and MOSEI dataset, respectively

importance in leveraging detailed information for better performance.

The removal of the multimodal fusion Transformer results in another performance drop, highlighting its effectiveness in modeling the relations between NRGFs and MFs. Finally, the performance drops significantly when the noise-aware learning scheme is removed, especially for Acc-7, indicating that the noise-aware learning scheme helps the model learn more useful emotion semantics from the multimodal data. Additionally, we observed that the performance degradation is more significant when the NRGF extractor is removed compared to the removal of the MF extractor. This could be because MFs contain more noisy information.

Notably, with the removal all components proposed in NORM-TR framework (see the last row), we conducted a standard Transformer as our baseline. When using the standard Transformer, we simply concatenate the three input modalities without leveraging the NRGF and MF extractors for learning, and then fuse them using the Transformer. As expected, this method yields the lowest accuracy. For example, the Acc-7 on the MOSI dataset is only 21%, showing a dramatic drop. However, we found that on the larger dataset MOSEI, the decline is not as large as on the MOSI dataset, although the model Acc-7 accuracy was also low. We believe this is due to two main reasons: (1) the unprocessed multimodal inputs have significantly different distributions, increasing the difficulty of fusion for the Transformer. (2) The training data is not sufficient for the standard Transformer to learn to bridge the gaps, which can be partially supported by recent studies (Lian et al., 2024; Akbari et al., 2021; Kim et al., 2021)

4.7.2 Effects of Different Modalities

To discuss the effect of each modality on performance, Table 7 presents the ablation results of different modality settings on MOSI and MOSEI datasets, respectively. We observe that the combination of video, audio, and text information provided the best performance, suggesting that our model can learn the effective multimodal emotion representation for robust emotion recognition. On both datasets, the performance sharply dropped when the text modality was removed, indicating that the text modality plays an important role in multimodal emotion recognition.

In addition, we tried to include different levels of Type-2 noise in each modality respectively on MOSI dataset, to discuss the sensitivity of each modality to noise. We found that the performance degradation was more significant when adding the noise to the text modality (e.g., decreased by 13% on Acc-7 at 50% mask percentage) than the audio and video modalities (e.g., decreased by 0.6% and 1.1% on Acc-7 separately at 50% mask percentage). It shows that the text modality is more sensitive to noise than the other modalities.

4.7.3 Performance of Unimodality with Masked Input for Recognition

In order to explore the advantages of multimodality under the mask setting, we conducted classification experiments on the unimodal data with mask input. The relevant result is shown in Table 8. More specifically, we evaluated on the test set of MOSI and MOSEI under different masking percentages $r_t \in \{0, 0.1, ..., 1.0\}$, and used AUILC (see Eq. 10) to evaluate the performance. The results show that the model achieve the best performance under the multimodal setting. In contrast, when only a unimodality is used for emotion recognition, the performance decreases significantly. This phenomenon demonstrate that multimodal data can effectively enhance the robustness of the model. Moreover, it is worth noting that the accuracy is relatively low when either video or audio modality alone is used for emotion recognition. This may be due to the fact that text modality usually provides a greater contribution in emotion recognition, as also observed and discussed in other studies (Zhang et al., 2023). Therefore, we believe that pay more attention to suppressing emotion-irrelevant noise in each modality is necessary.

4.7.4 Effects of the Hyper-Parameter Settings in Transformer

Figure 6a presents the accuracy of emotion recognition on the RML dataset, which is effected by the number of attention blocks in the Transformer architecture. As shown from the results, the accuracy achieved the highest 88.89% when we set the depth to 2. Besides, we also observe that different Transformer depths only resulted in minor performance variations, indicating that the parameter has a small impact on our method.

In addition, we tried using more complex Transformer models instead of our multimodal fusion Transformer for fusion on the RML dataset, e.g., (1) using pairs of Transformers similar to MulT (Tsai et al., 2019); (2) concatenating the NRGFs and MFs and using a deeper ViT (Dosovitskiy et al., 2021) for fusion. We found that these complex models did not improve the performance, i.e., obtaining 88.75% and 88.19% of Acc-6 on the RML dataset by MulT and ViT, repectively. Meanwhile, these complex models require more parameters for training and additional computational cost (about 1 MACs). Our simple but effective multimodal fusion Transformer is able to complement the potentially insensitive but useful information of NRGFs by MFs containing more details, achieving more accurate emotion understanding (88.89% of Acc-6).

Table 7 Effects of different modalities

Method	MOSI		MOSEI	
	Acc-7 (↑)	MAE (↓)	Acc-7 (↑)	MAE (↓)
NORM-TR	**48.5**	**0.698**	**54.6**	**0.529**
w/o Audio	43.6	0.761	50.6	0.655
w/o Video	43.3	0.765	51.8	0.592
w/o Text	18.1	1.410	41.2	0.831

For each evaluation metric, ↑ indicates the bigger the better while ↓ indicates the smaller the better. The best result is highlighted in bold

Table 8 Performance of unimodality with masked input for recognition

Modality	MOSI		MOSEI	
	Acc-7 (↑)	MAE (↓)	Acc-7 (↑)	MAE (↓)
Text (T)	30.7	1.104	45.9	0.679
Audio (A)	16.6	1.406	40.5	0.853
Video (V)	17.0	1.415	41.6	0.822
T+A+V	**31.3**	**1.093**	**47.3**	**0.673**

For each evaluation metric, ↑ indicates the bigger the better while ↓ indicates the smaller the better. The best result is highlighted in bold

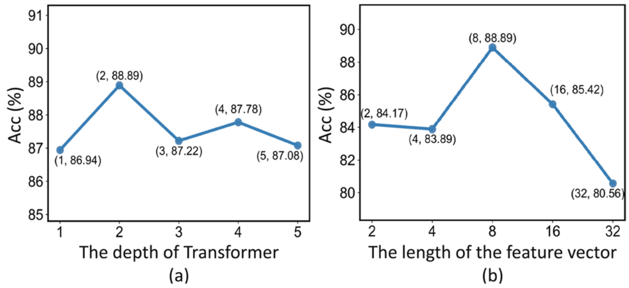


Fig. 6 The impact of important parameter settings in Transformer on the RML dataset. **a** The effect of attention blocks in Transformer, and **b** the effect of the number of sampled frames in per sequence

4.7.5 Effects of the Input Feature Vector Length

In Fig 6b, we present the accuracy curves, which are effected by the length of the input feature vector, i.e., the number of frames sampled from an original sequence. As shown in the figure, the accuracy achieved the highest 88.89% when we set the length to 8. Hence, in this study, we set the length of the input feature vector to 8. Moreover, the performance sharply dropped when the length was set to 32, indicating that too many frames sampled from data tend to introduce additional noise and lead to poor results.

4.7.6 Different Query, Key, and Value Settings in Transformer

Table 9 presents the experimental results of different query, key, and value settings in Transformer on the MOSI and MOSEI datasets, respectively. We find that the best performance of the model was obtained when using the NRGFs as the query and the MFs as the key and value. This demonstrates that our NORM-TR has the ability to extract a more generic and useful query for the Transformer to help achieve better emotion recognition performance.

Table 9 Effect of different Query, Key, and Value Setting in Transformer

Q	K & V	MOSI		MOSEI	
		Acc-7 (↑)	MAE (↓)	Acc-7 (↑)	MAE (↓)
MFs	NRGFs	47.1	0.706	53.2	0.539
NRGFs	**MFs**	**48.5**	**0.698**	**54.6**	**0.529**

Note: for each evaluation metric, ↑ indicates the bigger the better while ↓ indicates the smaller the better. The best result is highlighted in bold

4.7.7 Effects of the Weights for Regularization

To discuss the effect of the regularization of NORM-TR, we show the ablation result of α and β settings on MOSI and SIMS dataset in Table 10. More specifically, in our experiments, we chose different values of α and β for a series of evaluations. The experimental results show that the overall performance of the model on the MOSI dataset is best when the values of both α and β are set to 0.01. It is worth noting that when α and β are 0.005, the model performs poorly in Acc-7, even though it achieves the best MAE on the MOSI dataset. Therefore, based on these observations, we empirically fix the values of α and β to 0.01.

Table 10 Effects of the weights for regularization

α	β	γ	MOSI		MOSEI	
			Acc-7 (↑)	MAE (↓)	Acc-7 (↑)	MAE (↓)
0.001	0.001	1.0	44.8	0.701	52.9	0.543
0.005	0.005	1.0	47.4	**0.697**	52.3	0.547
0.01	0.01	1.0	**48.5**	0.698	**54.6**	**0.529**
0.05	0.05	1.0	46.5	0.715	53.2	0.546
0.1	0.1	1.0	44.9	0.728	52.0	0.552
0.5	0.5	1.0	46.5	0.713	52.5	0.546
1.0	1.0	1.0	44.6	0.716	52.3	0.550

Note: for each evaluation metric, ↑ indicates the bigger the better while ↓ indicates the smaller the better. The best result is highlighted in bold

Table 11 Effects of the different backbone

Backbone	Acc-6	F1
ResNet-18	**88.89**	**88.81**
ResNet-34	85.28	85.09
ResNet-50	84.44	84.25
ViT-B/16	79.03	78.91

4.7.8 Effects of the Different Backbone

To discuss the impact of the different backbone on model performance. As shown in Table 11, we selected ResNet and Vision Transformer (ViT) (Dosovitskiy et al., 2021) as backbone for our experiments on the RML dataset. Obviously, the NORM-TR achieve the best performance on the RML dataset when ResNet-18 is set as backbone, with a 6-classification accuracy and F1 score of 88.89% and 88.81%, respectively. In contrast, when backbone is replaced with ResNet-34, ResNet-50, and ViT, the performance of the model drops significantly. We observed that the models under all four Backbone configurations showed overfitting due to the small sample size of the dataset (e.g., the RML dataset has only 720 training samples). A similar situation was also observed on the IEMOCAP dataset. Therefore, we chose to use ResNet-18 as the backbone of the NORM-TR to extract features from these datasets more efficiently.

4.7.9 Effects of Different Fusion Techniques

To discuss the effects of different fusion techniques, we use a different technology for feature fusion on MOSI and MOSEI datasets. The details are shown in Table 12. Obviously, NOMR-TR perform best when using Transformer for feature fusion, demonstrating that using NRGFs as query, MFs as Key/Value can obtain a more complementary feature for emotion recognition.

Table 12 Effects of different fusion techniques

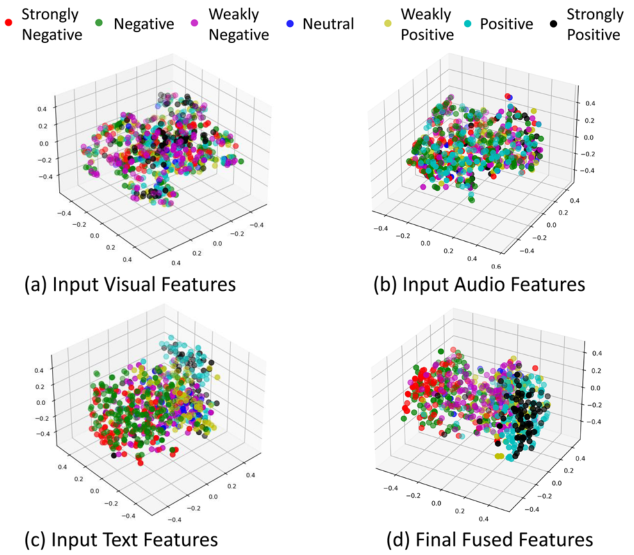
Fusion technique	MOSI		MOSEI	
	Acc-7 (↑)	MAE (↓)	Acc-7 (↑)	MAE (↓)
Concatenation	45.6	0.734	53.8	0.536
LSTM	44.5	0.754	53.2	0.544
GRU	43.2	0.768	52.7	0.538
Tensor Fusion (TFN)	44.9	0.759	53.2	0.543
Low-rank Fusion (LMF)	45.5	0.744	53.0	0.532
Ours	**48.5**	**0.698**	**54.6**	**0.529**

4.7.10 Visualization of the Input Unimodal Features and Final Fused Features

As shown in Fig. 7, we use t-sne (Van der Maaten & Hinton, 2008) to visualize the input unimodal feature (i.e., U_a', U_v' *and* U_t') and the final fused features. The visualizations shows that while unimodal features provide overlapping feature distributions for emotion classification, the final fused features exhibit more distinct clustering, especially for strongly positive. It demonstrates that multimodal fusion enhances feature separation and potentially improves classification accuracy by integrating the diverse and complementary information from each modality.

4.7.11 Visualization of the NRGFs and MFs

From the Fig. 8a, we can observe that the NRGFs extracted from the text modality are closer to the NRGFs extracted from the video and audio modalities in the feature space, which implies that the model prefers NRGFs that are complementary to the text modality when extracting features from the video and audio modalities.This phenomenon is in line with the results of our previous ablation experiments (see Sects. 4.7.2 and 4.7.3) and previous works (Zhang et al., 2023), which show that the model performance shows a significant decrease when the text modality is removed. This



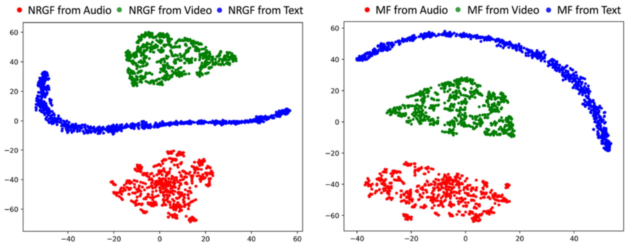
(a) Input Visual Features

(b) Input Audio Features

(c) Input Text Features

(d) Final Fused Features

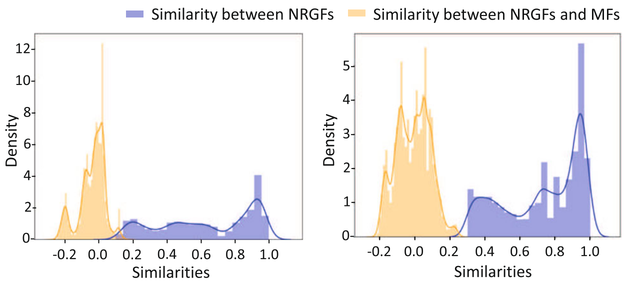
Fig. 7 Visualization of the input unimodal features ($U'_a/U'_v/U'_t$) and final fused features on MOSI dataset



(a) NRGFs Extracted from each modality

(b) MFs Extracted from each modality

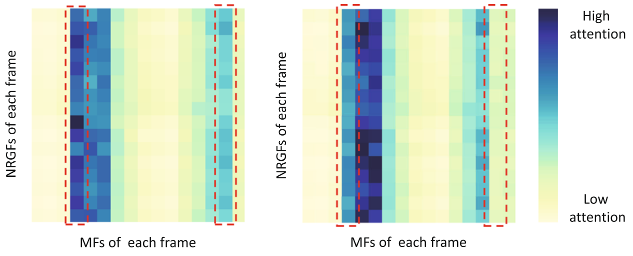
Fig. 8 Visualization of the NRGFs and MFs on MOSI dataset



(a) w/o Noise-aware Learning Scheme

(b) with Noise-aware Learning Scheme

Fig. 9 Visualization of NRGFs and MFs distributions on the MOSI dataset, with and without using the proposed noise-aware learning scheme, respectively. **a** Similarities distributions without using the noise-aware learning scheme on MOSI dataset; **b** Similarities distributions with using the noise-aware learning scheme on MOSI dataset



(a) without noise

(b) with noise

Fig. 10 Visualization of the attention weights learned by the multi-modal fusion Transformer for a randomly selected sample without and with a random mask on the RML dataset. **a** The attention weights without the mask noise, **b** the attention weights with the mask noise. Note: darker colors indicate higher attention weights for learning and the red dashed boxes represent the attention distribution of two randomly selected frames in the sequence

result confirms that the text modality plays a more important role in emotion recognition. On the contrary, we can see from the Fig. 8b that the difference in the distribution of MF features between modalities is more significant, indicating that the model adopts a different strategy from NRGF in extracting MF features, thus acquiring a more comprehensive modal representation for emotion recognition.

Furthermore, despite the moderating effect of $\mathcal{F}_{D_1}$ in adversarial learning, the distribution of NRGF did not completely converge to a same distribution. We hypothesize that this reflects the complexity of the emotion features themselves, i.e., the model does not tend to forcefully align NRGFs extracted in different modalities to the same distribution. Instead, the model seems to adaptively find a balance between maintaining performance optimization and feature distribution consistency.

4.7.12 Visualization of Noise-Resistant and Multimodal Feature Distributions

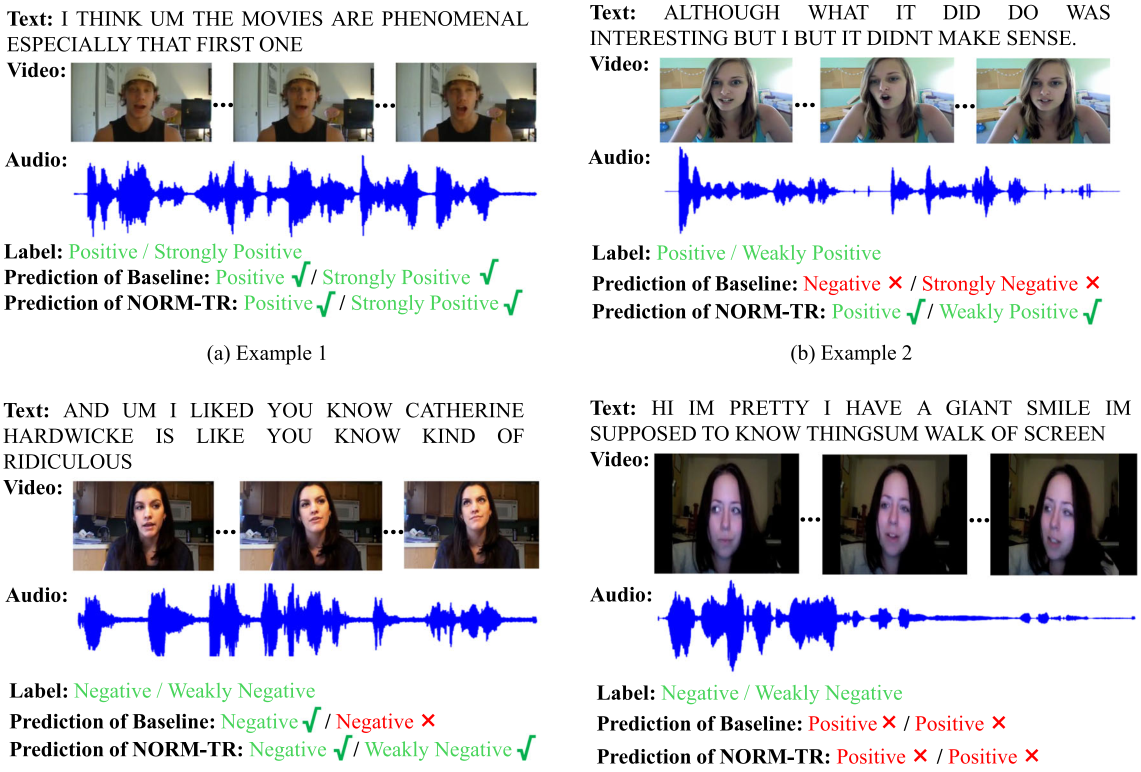
In Fig. 9, we visualized the similarity distributions of the extracted Noise-Resistant Generic Features (NRGFs) and Multimodal Features (MFs) on the MOSI dataset, with and without using the proposed noise-aware learning scheme, respectively. We applied the Kernel Density Estimation (KDE) and cosine similarity to describe the similarity distribution of the two types of features. As shown in Fig. 9b, with a noise-aware learning scheme, most of the NRGFs are similar (similarities close to 1) while most of the NRGFs and MFs are different (similarities close to 0), demonstrating that the NRGFs are more generic and noise-consistent while MFs contain more modality-specific characteristics. On the contrary, without using the noise-aware learning scheme, the similarities between the NRGFs are significantly increased (see Fig. 9a), indicating that they retain more noisy information from different modalities.

4.7.13 Visualization of Attention Weights Learned in Transformer

Figure 10a and b show the attention weight matrixes learned by the Transformer for a randomly selected sample without and with a random mask on the RML dataset, respectively. In

Table 13 Performance comparison with the state-of-the-art method for multi-label emotion classification, where the results in the first segment are from the original paper, and the results in the second segment are reproduced by the authors in the same settings. Note: w-Acc is weighted binary classification accuracy of each category. ⋆ means the results are independently reproduced by the authors due to the unavailability of their original open-source code. Bold indicates the best results in each segment

Methods	Happiness		Sadness		Anger		Suprise		Disgust		Fear		Overall	
	w-Acc	F1	w-Acc	F1	w-Acc	F1	w-Acc	F1	w-Acc	F1	w-Acc	F1	w-Acc	F1
MESM	64.10	72.30	**63.00**	46.60	**66.80**	49.30	**65.70**	27.20	**75.60**	56.40	**65.80**	28.90	**66.80**	46.80
Graph-MFN	**66.30**	66.30	60.40	66.90	62.60	72.80	53.70	85.50	69.10	76.60	62.20	89.90	62.35	76.33
CIA	51.90	71.30	61.80	72.90	64.70	74.70	58.20	86.00	74.10	81.80	63.90	87.80	62.88	79.08
M3ER⋆	**62.91**	**62.92**	55.28	70.11	**58.96**	**73.29**	**56.19**	83.01	67.65	**80.97**	**52.49**	85.30	**58.91**	75.93
Ours (NORM-TR)	60.79	60.88	**57.08**	**70.88**	58.31	73.00	51.30	**86.41**	**70.66**	80.24	50.91	**87.92**	58.17	**76.56**


Text: I THINK UM THE MOVIES ARE PHENOMENAL ESPECIALLY THAT FIRST ONE
Video:
Audio:
Label: Positive / Strongly Positive
Prediction of Baseline: Positive √ / Strongly Positive √
Prediction of NORM-TR: Positive √ / Strongly Positive √
Text: ALTHOUGH WHAT IT DID DO WAS INTERESTING BUT I BUT IT DIDNT MAKE SENSE.
Video:
Audio:
Label: Positive / Weakly Positive
Prediction of Baseline: Negative ✕ / Strongly Negative ✕
Prediction of NORM-TR: Positive √ / Weakly Positive √
Text: AND UM I LIKED YOU KNOW CATHERINE HARDWICKE IS LIKE YOU KNOW KIND OF RIDICULOUS
Video:
Audio:
Label: Negative / Weakly Negative
Prediction of Baseline: Negative √ / Negative ✕
Prediction of NORM-TR: Negative √ / Weakly Negative √
Text: HI IM PRETTY I HAVE A GIANT SMILE IM SUPPOSED TO KNOW THINGSUM WALK OF SCREEN
Video:
Audio:
Label: Negative / Weakly Negative
Prediction of Baseline: Positive ✕ / Positive ✕
Prediction of NORM-TR: Positive ✕ / Positive ✕

(a) Example 1

(b) Example 2

(c) Example 3

(d) Example 4

Fig. 11 Visualization of successful use cases and failures. Note: The left side of the /' represents the label/prediction for Acc-2, while the right side of the /' represents the label/prediction for Acc-7. The baseline model used for comparison is the standard Transformer

practice, for audio and video data, we set the pre-computed feature vectors U_a, U_v of the mask frame to 0. Since the NRGFs and MFs are shaped as $2T \times \frac{N}{2}$, where $T = 8$, the size of the weight matrix is 16×16. Each square in the matrix represents the attention weight score learned by the Transformer between the corresponding frames of the NRGFs and MFs. The red dashed boxes show the attention distributions of two randomly selected frames in the matrix. Compared with the attention weights of the unmasked frames (obtaining darker squares in Fig. 10a), the attention weights of the masked frames decrease significantly (obtaining lighter squares in Fig. 10b). This indicates that the Transformer prefers to translate the information of the frame without the mask by extracting NRGFs as the query and MFs as the key and value, rather than focusing on noisy frames with the mask, so that suppress the side influence of noise information.

4.7.14 Performance Comparison for Multi-label Classification

Since the MOSEI dataset also provides multi-label annotations, we explored the performance of NORM-TR in multi-label emotion classification. Specifically, we applied the Sigmoid function and BCE loss function to NORM-TR to achieve multi-label prediction. Table 13 shows the comparison results with other methods (i.e., M3ER M3ER (Mittal et al., 2020b), MESM (Dai et al., 2021), Graph-MFN (Zadeh et al., 2018), and CIA (Chauhan et al., 2019)). To ensure a fair comparison, we replicated the most SOTA method, M3ER, because of the differences in data processing methods and experimental settings across all methods, and compared it with NORM-TR under the same setup. Consistent with previous works (Franceschini et al., 2022; Mittal et al., 2020b), we report weighted binary classification accuracy (w-Acc) and weighted F1 (F1). Notably, although our approach does not specifically focus on improving multi-label classification, it still achieves very competitive performance compared to the SOTA method in the same setting. This demonstrates the effectiveness of NORM-TR for multi-label emotion classification.

4.7.15 Case Visualization and Analysis

As shown in Fig. 11, we analyzed some examples from the datasets used in our study and compared the NORM-TR with the baseline model (namely standard Transformer). We can see that the NORM-TR can correctly predict most samples, demonstrating the robustness of our approach. In addition, it should be noticed that the example 2 and example 4 are representative hard samples and contains conflicting/noise information. From the example 2, we can see that the first frame seems to express a positive emotion, while the second and third frames seem to express a negative and neutral emotion. These potential conflicts can be viewed as noise and our model correctly predicted the sample as positive, while the baseline model makes an incorrect prediction. This demonstrates that the model can effectively capture the emotion cues despite contradictory phrases. In contrast, the Sample 4 is incorrectly predicted as positive due to the strong positive phrase “GIANT SMILE." In this case, certain video frames showed exaggerated smiling expressions that acted as noise, misleading both the NORM-TR and baseline model towards a positive prediction. This indicates a potential limitation in handling ambiguous or noisy emotional expressions, where visual or audio noise can override the textual sentiment. In our follow-up work (Zhang et al., 2023), we have tried to solve this problem by using language to guide other modal representations. Additionally, other examples show that our method can correctly judge most samples, demonstrating the robustness of our approach.

According to these examples, we also highlight the following strengths of the NORM-TR model. Firstly, it effectively integrates text, audio, and video data to capture complex emotional cues in noisy samples due to its noise-aware learning scheme. Secondly, it performs well in complex environments, indicating strong real-world applicability. Lastly, the model demonstrates high robustness in handling various emotional expressions across different contexts.

5 Conclusion

This paper proposed a novel Noise-Resistant Multimodal Transformer (NORM-TR) approach for multimodal emotion recognition. The NORM-TR consists of a Noise-Resistant Generic Feature (NRGF) extractor, a Multimodal Feature (MF) extractor, and a multimodal fusion Transformer to fuse NRGFs and MFs, thus significantly reducing the negative impacts of noise in the multimodal data. To this end, a novel noise-aware learning scheme is further designed to help optimize the NORM-TR appropriately to obtain noise-invariant emotion representations. Extensive experiments on several multimodal datasets, including MOSI, MOSEI, IEMOCAP, and RML, show that our method outperforms other approaches, demonstrating the importance of handling noisy information as well as the effectiveness of our method. Despite the effectiveness of our method, we found that our method does not capture the problem of the inconsistency and absence of emotional labels. In the future, we will introduce more advanced semi-supervised or self-supervised learning mechanisms into our method to learn from unlabeled data, thus obtaining a more robust emotion understanding.

Acknowledgements This study was partially supported by National Natural Science Foundation of China Grant (62076227, 62002090), Natural Science Foundation of Hubei Province Grant (2023AFB572) and Major Science and Technology Projects in Yunnan Province (202202AD080007).

Data Availability All datasets used in this study are publicly available.

Declaration

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

Akbari, H., Yuan, L., Qian, R., Chuang, W., Chang, S., Cui, Y., & Gong, B. (2021). VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, neurips 2021, December 6–14, 2021, virtual* (pp. 24206–24221).

Baltrusaitis, T., Robinson, P., & Morency, L. (2016). Openface: An open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (wacv)* (pp. 1–10). IEEE Computer Society.

Beale, R., & Peter, C. (2008). The role of affect and emotion in HCI. In *Affect and emotion in human–computer interaction* (Vol. 4868, pp. 1–11). Springer.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., & Erhan, D. (2016). Domain separation networks. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., & Garnett, R. (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation, 42*(4), 335–359.

Chauhan, D. S., Akhtar, M. S., Ekbal, A., & Bhattacharyya, P. (2019). Context-aware interactive attention for multi-modal sentiment and emotion analysis. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019* (pp. 5646–5656). Association for Computational Linguistics.

Cornejo, J. Y. R., & Pedrini, H. (2019). Audio-visual emotion recognition using a hybrid deep convolutional neural network based on census transform. In *2019 IEEE international conference on systems, man and cybernetics* (pp. 3396–3402). IEEE.

Dai, W., Cahyawijaya, S., Liu, Z., & Fung, P. (2021). Multimodal end-to-end sparse model for emotion recognition. In *Proceedings of the 2021 conference of the north American chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2021, online, June 6–11, 2021* (pp. 5305–5316). Association for Computational Linguistics.

Ding, L., Wang, L., Liu, X., Wong, D. F., Tao, D., & Tu, Z. (2021). Understanding and improving lexical choice in non-autoregressive translation. In *International conference on learning representations (iclr).*

D'Mello, S. K., & Kory, J. M. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys (CsUR), 47*(3), 1–36.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

El-Madany, N. E., He, Y., & Guan, L. (2016). Multiview emotion recognition via multi-set locality preserving canonical correlation analysis. In *IEEE international symposium on circuits and systems(iscas)* (pp. 590–593). IEEE.

Franceschini, R., Fini, E., Beyan, C., Conti, A., Arrigoni, F., & Ricci, E. (2022). Multimodal emotion recognition with modality-pairwise unsupervised contrastive loss. In *26th international conference on pattern recognition, ICPR 2022, Montreal, QC, Canada, August 21–25, 2022* (pp. 2589–2596). IEEE.

Ganin, Y., & Lempitsky, V. S. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning* (Vol. 37, pp. 1180–1189). JMLR.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (Vol. 27, pp. 2672–2680).

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

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (cvpr)* (pp. 770–778).

He, Z., Zhang, L., Gao, X., & Zhang, D. (2022). Multi-adversarial faster-rcnn with paradigm teacher for unrestricted object detection. *International Journal of Computer Vision*, 1–21.

Huang, J., Tao, J., Liu, B., Lian, Z., & Niu, M. (2020). Multimodal transformer fusion for continuous emotion recognition. In *2020 IEEE international conference on acoustics, speech and signal processing (icassp)* (pp. 3507–3511). IEEE.

Kenton, J. D. M-W. C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl-hlt* (pp. 4171–4186).

Kim, W., Son, B., & Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th international conference on machine learning, ICML 2021, 18–24 July 2021, virtual event* (Vol. 139, pp. 5583–5594). PMLR.

Li, C., Deng, C., Li, N., Liu, W., Gao, X., & Tao, D. (2018). Self-supervised adversarial hashing networks for cross-modal retrieval. In *2018 IEEE/cvf conference on computer vision and pattern recognition* (pp. 4242–4251).

Lian, Z., Sun, L., Sun, H., Chen, K., Wen, Z., Gu, H., & Tao, J. (2024). Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion, 108*, 102367.

Liang, J., Li, R., & Jin, Q. (2020). Semi-supervised multi-modal emotion recognition with cross-modal distribution matching. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2852–2861). ACM.

Liu, Y., Dai, W., Feng, C., Wang, W., Yin, G., Zeng, J., & Shan, S. (2022). *MAFW: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild* (pp. 24–32). ACM.

Liu, Y., Wang, W., Feng, C., Zhang, H., Chen, Z., & Zhan, Y. (2023). Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition, 138*, 109368. https://doi.org/10.1016/J.PATCOG.2023.109368

Liu, Y., Wang, W., Zhan, Y., Feng, S., Liu, K., & Chen, Z. (2023). Pose-disentangled contrastive learning for self-supervised facial representation. In *Proceedings of the IEEE/C conference on computer vision and pattern recognition (cvpr)* (pp. 9717–9728).

Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. (2018). Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, volume 1: Long papers* (pp. 2247–2256). Association for Computational Linguistics.

Luo, H., Ji, L., Huang, Y., Wang, B., Ji, S., & Li, T. (2021). Scalevlad: Improving multimodal sentiment analysis via multi-scale fusion of locally descriptors. arXiv:2112.01368

Lv, F., Chen, X., Huang, Y., Duan, L., & Lin, G. (2021). Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *2021 IEEE/CVF conference on computer vision and pattern recognition (cvpr)* (pp. 2554–2562). Computer Vision Foundation/IEEE.

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Kosir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion, 46*, 184–192.

Maas, A. L., Hannun, A. Y., & Ng, A. Y., et al. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the ICML* (Vol. 30, p. 3).

Mao, H., Yuan, Z., Xu, H., Yu, W., Liu, Y., & Gao, K. (2022). M-SENA: An integrated platform for multimodal sentiment analysis. *Proceedings of the 60th annual meeting of the association for computational linguistics* (pp. 204–213). Association for Computational Linguistics.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020a). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 1359–1367).

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020b). M3ER: multiplicative multimodal emotion recognition using facial, textual, and speech cues. In *The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, February 7–12, 2020* (pp. 1359–1367). AAAI Press.

Niu, X., Yu, Z., Han, H., Li, X., Shan, S., & Zhao, G. (2020). Video-based remote physiological measurement via cross-verified feature disentangling. In *Computer vision-ECCV 2020—16th European conference, Glasgow, August 23–28, 2020, proceedings, part II* (Vol. 12347, pp. 295–310). Springer.

Pei, Z., Cao, Z., Long, M., & Wang, J. (2018). Multi-adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, pp. 3934–3941). AAAI Press.

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

Qian, Y., Zhang, Y., Ma, X., Yu, H., & Peng, L. (2019). EARS: Emotion-aware recommender system based on hybrid information fusion. *Information Fusion, 46*, 141–146.

Sahay, S., Okur, E., Kumar, S.H., & Nachman, L. (2020). Low rank fusion based transformers for multimodal sequences. arXiv:2007.02038

Shen, L., Wang, M., & Shen, R. (2009). Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Journal of Educational Technology and Society, 12*(2), 176–189.

Sun, Z., Sarma, P.K., Sethares, W.A., & Liang, Y. (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 8992–8999). AAAI Press.

Tsai, Y. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L., & Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019, Florence, July 28–August 2, 2019, volume 1: Long papers* (pp. 6558–6569). Association for Computational Linguistics.

Tsai, Y. H., Liang, P. P., Zadeh, A., Morency, L., & Salakhutdinov, R. (2019). Learning factorized multimodal representations. In *ICLR*.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. In *Journal of machine learning research*, *9*(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30, pp. 5998–6008).

Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017). Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on multimedia* (pp. 154–162). ACM.

Wang, Y., Chen, Z., Chen, S., & Zhu, Y. (2022). MT-TCCT: Multi-task learning for multimodal emotion recognition. In *Artificial neural networks and machine learning-ICANN 2022—31st international conference on artificial neural networks, Bristol, September 6–9, 2022, proceedings, part III* (Vol. 13531, pp. 429–442). Springer.

Wang, Y., & Guan, L. (2008). Recognizing human emotional state from audiovisual signals. *IEEE Transactions on Multimedia, 10*(5), 936–946.

Wang, Y., Herranz, L., & van de Weijer, J. (2020). Mix and match networks: Cross-modal alignment for zero-pair image-to-image translation. *International Journal of Computer Vision, 128*(12), 2849–2872.

Wang, Y., Shen, Y., Liu, Z., Liang, P. P., Zadeh, A., & Morency, L. (2019). Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 7216–7223). AAAI Press.

Yang, D., Huang, S., Kuang, H., Du, Y., & Zhang, L. (2022). Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM international conference on multimedia* (pp. 1642–1651). ACM.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., & Yang, K. (2020). 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* (pp. 3718–3727). Association for Computational Linguistics.

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

Yuan, Z., Li, W., Xu, H., & Yu, W. (2021). Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 4400–4407). ACM.

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1103–1114). Association for Computational Linguistics.

Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. (2018). 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* (pp. 2236–2246).

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

Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., & Yu, T. (2023). Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 conference on empirical methods in natural language processing, EMNLP 2023, Singapore, December 6–10, 2023* (pp. 756–767). Association for Computational Linguistics.

Zhang, Q., Xu, Y., Zhang, J., & Tao, D. (2022). Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *abs/2202.10108*

Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2018). Learning affective features with a hybrid deep model for audio-visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology, 28*(10), 3030–3043.

Zhang, Y., & Yang, Q. (2022). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering, 34*(12), 5586–5609.

Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., & Keutzer, K. (2020). An end-to-end visual-audio attention network for emotion recognition in user-generated videos. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, pp. 303–311). AAAI Press.

Zhou, H., Du, J., Zhang, Y., Wang, Q., Liu, Q., & Lee, C. (2021). Information fusion in attention networks using adaptive and multi-level factorized bilinear pooling for audio-visual emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 2617–2629.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Yuanyuan Liu[1] · **Haoyu Zhang**[1,2] · **Yibing Zhan**[4] · **Zijing Chen**[5,6] · **Guanghao Yin**[1] · **Lin Wei**[1] · **Zhe Chen**[3,6]

Yuanyuan Liu
liuyy@cug.edu.cn

Yibing Zhan
zhanyibing@jd.com

Zijing Chen
zijing.chen@acu.edu.au

Guanghao Yin
ygh2@cug.edu.cn

Lin Wei
linw@cug.edu.cn

Zhe Chen
zhe.chen1@sydney.edu.au

[1] School of Computer Science, China University of Geosciences (Wuhan), Wuhan, China

[2] National Engineering Research Center for Geographic Information System, China University of Geosciences (Wuhan), Wuhan, China

[3] School of Computer Science, The University of Sydney, Sydney, Australia

[4] JD Explore Academy, JD.com, Beijing, China

[5] Peter Faber Business School, Australian Catholic University, Sydney, Australia

[6] Cisco-La Trobe Centre for AI and IoT, La Trobe University, Melbourne, Australia