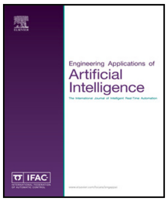
Engineering Applications of Artificial Intelligence 133 (2024) 108348

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Token-disentangling Mutual Transformer for multimodal emotion recognition

Guanghao Yin[a], Yuanyuan Liu[a,b,*], Tengfei Liu[a], Haoyu Zhang[a], Fang Fang[a], Chang Tang[a], Liangxiao Jiang[a]

[a] School of Computer Science, China University of Geosciences (Wuhan), Wuhan, China
[b] School of Computer Science and Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Multimodal emotion recognition
Multimodal emotion Token disentanglement
Token mutual Transformer
Token separation learning
Bi-directional query learning

ABSTRACT

Multimodal emotion recognition presents a complex challenge, as it involves the identification of human emotions using various modalities such as video, text, and audio. Existing methods focus mainly on the fusion information from multimodal data, but ignore the interaction of the modality-specific heterogeneity features that contribute differently to emotions, leading to sub-optimal results. To tackle this challenge, we propose a novel Token-disentangling Mutual Transformer (TMT) for robust multimodal emotion recognition, by effectively disentangling and interacting inter-modality emotion consistency features and intra-modality emotion heterogeneity features. Specifically, the TMT consists of two main modules: multimodal emotion Token disentanglement and Token mutual Transformer. In the multimodal emotion Token disentanglement, we introduce a Token separation encoder with an elaborated Token disentanglement regularization, which effectively disentangle the inter-modality emotion consistency feature Token from each intra-modality emotion heterogeneity feature Token; consequently, the emotion-related consistency and heterogeneity information can be performed independently and comprehensively. Furthermore, we devise the Token mutual Transformer with two cross-modal encoders to interact and fuse the disentangled feature Tokens by using bi-directional query learning, which delivers more comprehensive and complementary multimodal emotion representations for multimodal emotion recognition. We evaluate our model on three popular three-modality emotion datasets, namely CMU-MOSI, CMU-MOSEI, and CH-SIMS, and the experimental results affirm the superior performance of our model compared to state-of-the-art methods, achieving state-of-the-art recognition performance. Evaluation Codes and models are released at https://github.com/cug-ygh/TMT.

1. Introduction

Emotion recognition is a research hotspot in artificial intelligence and affective computing (Sun et al., 2020a; An and Wan Zainon, 2023; Singh and Kapoor, 2023). Compared with traditional unimodal emotion recognition, which only focuses on the recognition of emotions from a single modality (Chen and Joo, 2021), multimodal emotion recognition that exploits different data sources, such as video, audio, and text, has shown significant advantages in improving the understanding of human emotions and is more in line with real-world emotion interaction and applications (Zadeh et al., 2017a; Tsai et al., 2019a; Lv et al., 2021; Hazarika et al., 2020a; Yuan et al., 2021a; Yan et al., 2023a,b). Sentiment information obtained from videos, posts and comments on the Internet can be used for many purposes (Liu et al., 2023d; Zhong et al., 2022). For example, governments can use these information to predict how people want to vote. Movie producers can predict the final box office direction of a movie based on comments. Companies can improve their products based on user feedback (Zeng et al., 2024).

To meet the emotion interaction requirements of real application scenarios, the multimodal emotion recognition has received increasing attention from researchers by extracting and fusing emotion information from different modalities. For instance, Sun et al. (2020b) introduced Deep Canonical Correlation Analysis (DCCA) to capture correlation features between modalities. Liang et al. (2021) developed a model that addresses distribution differences in the modality invariant space, thereby reducing inter-modality heterogeneity. Zadeh et al. (2017b) fused the eigenvectors of multiple modalities into tensors and dynamically modeled the interplay between modalities, successfully overcoming the challenge of fusion in multimodal scenarios. Despite

[*] Corresponding author at: School of Computer Science, China University of Geosciences (Wuhan), Wuhan, China.
E-mail addresses: ygh2@cug.edu.cn (G. Yin), liuyy@cug.edu.cn, scse-yyliu@ntu.edu.sg (Y. Liu), tf66366@cug.edu.cn (T. Liu), zhanghaoyu@cug.edu.cn (H. Zhang), fangfang@cug.edu.cn (F. Fang), tangchang@cug.edu.cn (C. Tang), ljiang@cug.edu.cn (L. Jiang).

https://doi.org/10.1016/j.engappai.2024.108348
Received 28 October 2023; Received in revised form 21 February 2024; Accepted 26 March 2024
Available online 6 April 2024

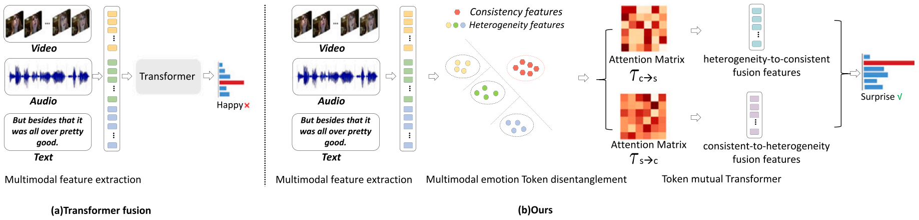

Video
Audio
But besides that it was all over pretty good.
Text
Transformer
Happy ×
Multimodal feature extraction
(a)Transformer fusion
Video
Audio
But besides that it was all over pretty good.
Text
Consistency features
Heterogeneity features
Attention Matrix
$\mathcal{T}_{c\to s}$
heterogeneity-to-consistent fusion features
Attention Matrix
$\mathcal{T}_{s\to c}$
consistent-to-heterogeneity fusion features
Surprise √
Multimodal feature extraction
Multimodal emotion Token disentanglement
Token mutual Transformer
(b)Ours

Fig. 1. The comparison between the existing Transformer fusion-based method and the proposed TMT-based method for multimodal emotion recognition. (a) Transformer fusion-based multimodal emotion recognition, (b) Our TMT-based multimodal emotion recognition. The TMT achieves more comprehensive and complementary multimodal emotion representations for robust emotion recognition through joint multimodal emotion Token disentanglement and Token mutual Transformer.

the progress achieved by the aforementioned methods, they mainly focus on modeling the common emotion information through multimodal fusion techniques while ignoring the unique information of different modalities. This makes it difficult for these methods to effectively model emotion relations between different modalities, thus affecting the performance of multimodal emotion recognition.

Recently, the Transformer has been extensively utilized in the fields of computer vision, natural language processing, and multimodal recognition due to its relation learning and modeling capabilities for sequence data (Dosovitskiy et al.; Liu et al., 2021; Liang et al., 2021; Lin et al., 2021; Zhang et al., 2024). For instance, Delbrouck et al. (2020) introduced a novel encoding architecture based on the Transformer model, utilizing a modular attention mechanism for encoding the relations of multiple modalities. Han et al. (2021) employed different modalities as input sources and target sources of the Transformer translation model to achieve modality fusion by modeling their relations to emotions. Wang et al. (2020) introduced a text–audio fusion module and a text–video fusion module, respectively, enhancing the output results of both modules using the Transformer's gating mechanism. Although leveraging Transformer for multimodal feature fusion is effective in modeling affective relationships between different modalities, they still do not take into account the subtle heterogeneous features of a specific modality to emotion interactions, resulting in the sub-optimal performance (see Fig. 1).

To achieve this, one strategy is to introduce adversarial learning (Yang et al., 2022b) to extract consistency information of multiple modalities and heterogeneity information of each modality, respectively. Park and Im (2016) applied adversarial learning to multimodal sentiment representation learning, using only category information for multimodal embedding. He et al. (2023) proposed the adversarial invariant-specific representations fusion model to achieve modality-invariant representations by narrowing the distribution gaps among different modalities. Despite the progress, the interaction of subtle heterogeneity between different modalities still confuses us and could easily comprise to learn comprehensive multimodal representations. For example, the audio modality contains unique intonation emotion information, which is difficult to match in other modalities (*e.g.*, image and text). As a result, current adversarial learning-based methods that directly use a normal Transformer (Vaswani et al., 2017) or Multilayer Perceptron (Tolstikhin et al., 2021) to perform the fusion of heterogeneous and consistent features are difficult to capture the subtle heterogeneity emotion interactions between them. That is, the subtle but meaningful modality information would still be ignored. Moreover, the adversarial learning-based methods (Yang et al., 2022b; Park and Im, 2016; He et al., 2023; Liu et al., 2023b) also require additional well-designed network modules and a large amount of training data for proper training. This can result in an overwhelmingly large model capacity, making it challenging to achieve robust and efficient multimodal emotion recognition.

To address the above issue, we propose a novel **T**oken-disentangling **M**utual **T**ransformer(TMT) for multimodal emotion recognition. The TMT can effectively disentangle inter-modality emotion consistency features and intra-modality emotion heterogeneity features and mutually fuse them for more comprehensive multimodal emotion representations by introducing two primary modules, namely multimodal emotion Token disentanglement and Token mutual Transformer. The motivation of TMT and the comparison to existing fusion methods are shown in Fig. 1. Specifically, the multimodal emotion Token disentanglement module first thoroughly disentangles the inter-modality emotion consistency feature Token from each intra-modality emotion heterogeneity feature Token via a novel Token separation encoder and its Token disentanglement regularization. Then, to completely explore the emotion interactions between the disentangled feature Tokens, we further devise the Token mutual Transformer to integrate the disentangled features for the more comprehensive multimodal emotion representations by conducting two bi-directional query learning in two cross-modal encoders. Together with the two modules, our TMT can achieve state-of-the-art multimodal emotion recognition performance. In summary, the significant contributions of this paper can be summarized as follows:

- We propose a novel TMT for robust multimodal emotion recognition. The TMT introduces a multimodal emotion Token disentanglement module and a Token mutual Transformer to effectively mine and integrate multimodal emotion information to achieve robust multimodal emotion recognition. Our method outperforms existing state-of-the-art approaches in multimodal emotion recognition, as demonstrated by experiments on three widely-used datasets (CMU-MOSI, CMU-MOSEI, CH-SIMS).
- We propose a novel and easy-to-implement multimodal emotion Token disentanglement module to disentangle the inter-modality emotion consistency feature Token from each intra-modality emotion heterogeneity feature Token effectively. To achieve this, we introduce a Token separation encoder with its Token disentanglement regularization into the module to help the Transformer separate four groups of features without taking additional parameters and computational complexity.
- We devise a Token mutual Transformer with two bi-directional query learning to fully interact and integrate the emotion consistency and heterogeneity information by exploring their mutual contribution in emotion interactions, resulting in more comprehensive and complementary multimodal emotion representations.

2. Related work

In this section, we provide an overview of related work, including multimodal emotion recognition and multimodal Transformer.

2.1. Multimodal emotion recognition

Most existing methods for multimodal emotion recognition can be broadly categorized into two main approaches: representation learning-based methods and multimodal fusion-based methods.

Representation learning-based methods focus on learning modality representations by considering the difference and consistency of different modalities, thus improving multimodal emotion recognition. For instance, Yang et al. (2022a) employed encoders and discriminators to learn consistent and heterogeneity features across multiple modalities using adversarial learning. Hazarika et al. (2020a) used metric learning to learn modality-specific and modality-invariant representations for multimodal emotion recognition tasks. Han et al. (2021b) proposed a framework named MMIM that improves multimodal representation with hierarchical mutual information maximization. Zhao et al. (2020) proposed a new attention-based VAANET that integrates spatial, channel, and temporal attention for audio–video emotion recognition. Lv et al. (2021) introduces a message hub to exchange information with each modality by sending common messages to each modality and reinforcing their features. Zeng et al. (2024) proposes a feature-based restoration dynamic interaction network for multimodal sentiment analysis. Recently, the feature disentanglement methods have been applied to emotion recognition for emotion-related feature alignment. For instance, the MISA approach, as presented by Hazarika et al. (2020b), involves projecting each modality into two distinct subspaces through the application of carefully designed constraints and encoders. Similarly, Yang et al. (2022a) have introduced the Feature Separation Multimodal Emotion Recognition (FDMER) method, addressing modality heterogeneity by mapping each modality into both a modality-invariant subspace and a modality-specific subspace. Through the incorporation of adversarial learning strategies, they refine both the public and private representations of feature separation. Another notable contribution is the Multimodal Feature Separation Approach (MFSA) presented by Yang et al. (2022b), which proposes an approach for acquiring effective multimodal representations in asynchronous sequences, with a specific emphasis on achieving feature disentanglement.

However, these methods mainly concentrate on the emotion representation of different modalities, neglecting the modality fusion for comprehensive feature learning, resulting in sub-optimal performance.

Multimodal fusion-based methods primarily aim to reduce heterogeneity among modalities and obtain more comprehensive multimodal emotion features. Zadeh et al. (2017b) proposed a tensor fusion method (TFN) to model the relationships between different modalities by computing the cartesian product. Liang et al. (2018) proposed a recursive multi-level fusion method that addressed the fusion problem by dividing it into multiple stages and continuously fusing subsets of multimodal signals to accomplish the final fusion task. Lv et al. (2021) proposed the Progressive Modality Reinforcement (PMR) approach, which introduces a modal reinforcement unit to learn the directional paired attention between cross-modal elements for modal asynchronous fusion. Despite the advancements made, many current methods primarily emphasize the integration of multiple features, rarely delve into the interaction of these features, and often overlook subtle heterogeneous features.

2.2. Multimodal Transformer

Transformer is an attention-based block for machine translation introduced by Vaswani et al. (2017). It learns the relationships between Tokens by aggregating data from the entire sequence, showing an excellent modeling ability in various tasks, such as speech processing, natural language processing, and computer vision, etc. (Kenton and Toutanova, 2019; Carion et al., 2020; Chen et al., 2022; Liu et al., 2023a; Tang et al., 2023, 2022). Dosovitskiy et al. proposed an approach that involves adding an additional learnable token to the sequence in order to capture classification information. Consequently, it is utilized in multimodal emotion recognition to facilitate fusion between multimodal sequences. For example, Yuan et al. (2021b) proposed a Transformer-based feature reconstruction network for handling randomly missing multimodal datasets. They employed a cross-modal attention mechanism for fusion at the front end and incorporated a missing reconstruction module to generate missing features. Tsai et al. (2019b) introduced an end-to-end network called MulT, which leverages a cross-modality attention mechanism for information interaction among different modalities, enabling the potential flow of information across modalities. Although progress has been made, existing transformer-based methods mainly learn modality relationships and still do not take into account the subtle heterogeneous features of a specific modality to emotion interactions, resulting in sub-optimal performance. To address this limitation, our proposed TMT can obtain multimodal emotion information more effectively by first decoupling and then interacting with multiple features.

Table 1
The symbols defined in the proposed method.

Symbol	Meaning
U_a, U_v, U_t	Pre-computed feature vectors of audio, video, and text
H_{c_0}	Initialized inter-modality consistency feature Token
$H_{a_0}, H_{t_0}, H_{v_0}$	Initialized intra-modality heterogeneity feature Token
H_c	Disentangled inter-modality consistency feature Token
H_a, H_v, H_t	Disentangled intra-modality heterogeneity feature Token
H_s	Heterogeneity feature tensor spliced by H_a, H_v, H_t
H_{f1}	Heterogeneity-to-consistent fusion features
H_{f2}	Consistent-to-heterogeneity fusion features
Y_o	Final comprehensive multimodal emotion representation

3. Methodology

Rather than the normal Transformer that focuses on feature fusion for multimodal emotion recognition while neglecting cross-modal interaction, this paper proposes a novel and effective Token-disentangling mutual Transformer (*i.e,* TMT) for robust multimodal emotion recognition by fully disentangling and interacting the emotion consistency and heterogeneity information from multimodal data. Fig. 2 shows the training pipeline of TMT for multimodal emotion recognition, which is composed of two main modules, namely the multimodal emotion Token disentanglement and Token mutual Transformer. Specifically, with the multimodal features extracted by three multi-layer perceptrons (MLP) from multimodal data, the multimodal emotion Token disentanglement module first employs a parameter-sharing Token separation encoder with its Token disentanglement regularization (including the intra-modality similarity loss, inter-modality orthogonal loss, and multimodal disentanglement loss), to respectively disentangle the inter-modality emotion consistency feature Token from each intra-modality emotion heterogeneity feature Token. Then, to completely explore the contribution of the disentangled Tokens in terms of the emotion interactions, we further devise the Token mutual Transformer with two cross-modal encoders for emotion feature fusion by using bi-directional emotion query learning, so as to obtain more comprehensive multimodal emotion representations for robust multimodal emotion recognition.

To help users better understand the proposed method, we summarize all the symbols defined in the proposed method in Table 1. We will describe the multimodal emotion Token disentanglement and Token mutual Transformer in the following sections.

3.1. Multimodal feature extraction

Regarding normal multimodal emotion recognition, we use the symbols U_a, U_v, and U_t to represent the input multimodal information of audio, video, and text, respectively. In the related literature (Mao et al., 2022), pre-computed features rather than raw data of different modalities are commonly used. Thus, to be fair, the U_a, U_v, and U_t in our paper represent the pre-computed feature vectors. For example, rather than using 2D images of a video, we can have as video input the pre-computed features $U_v \in \mathbb{R}^{T_o \times d_v}$ for the video modality, where T_o represents the length of the video, and d_v represents the length of each video feature vector. Therefore, we also have as input the pre-computed features for the text and audio modalities as $U_t \in \mathbb{R}^{T_o \times d_t}$

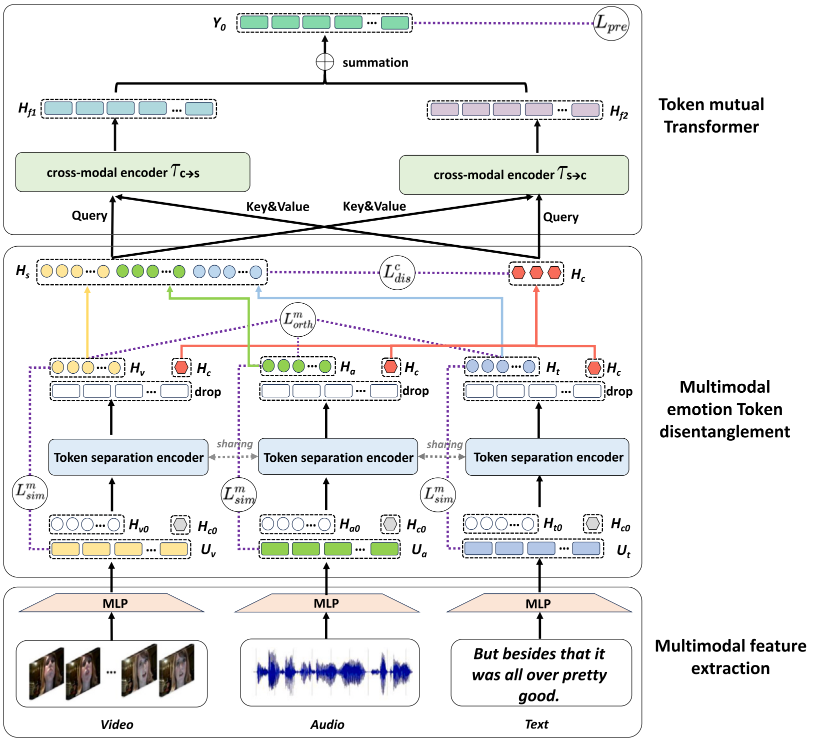


Fig. 2. The training pipeline of the proposed Token-disentangling mutual Transformer(TMT) for multimodal emotion recognition. With the extracted multimodal features, the TMT first uses the multimodal emotion Token disentanglement to separate the inter-modality emotion consistency feature Token from each intra-modality emotion heterogeneity feature Token and then employ the Token mutual Transformer to perform feature interaction and fusion via bi-directional query learning in two cross-modal encoders, thus obtaining more comprehensive multimodal emotion representations.

and $U_a \in \mathbb{R}^{T_o \times d_a}$, respectively, where d_t, d_a represent the lengths of the corresponding text and audio feature vectors, respectively. The use of pre-computed features has been widely accepted in the literature on multimodal emotion recognition. With the pre-computed features U_a, U_v, U_t, we employ three parallel Multilayer Perceptrons (represented as $MLP()$), namely fully connected layers, to normalize the sequence dimension of each feature vector to $d = 256$. With these three normalized feature vectors, we then implement the concatenate operation as $U_0 = con(MLP(U_a), MLP(U_v), MLP(U_t))$, thus obtaining the multimodal feature vector U_0 with a dimension of $3T_o \times d$.

3.2. Multimodal emotion Token disentanglement

3.2.1. Token separation encoder

Using the obtained multimodal features U_0, the multimodal emotion Token disentanglement module employs a parameter-sharing Transformer encoder as the Token separation encoder to disentangle the emotion-related inter-modality emotion consistency feature Token from each intra-modality emotion heterogeneity feature Token, so that obtain four group emotion feature Tokens. The separation learning procedure is shown in Fig. 2.

Formally, instead of the single Token vector used in the normal Transformer, we randomly initialize four different Token vectors, including an inter-modality consistency feature Token represented as H_{c_0} and three separated intra-modality heterogeneity feature Tokens represented as $H_{a_0}, H_{t_0}, H_{v_0}$, respectively. Among them, H_{c_0} is used to learn the emotion-related consistency features among the three modalities with the dimension of $6 \times d$, while $H_{a_0}, H_{t_0}, H_{v_0}$ are used to learn the heterogeneity features of audio, text and video modality, respectively, in each dimension of $2 \times d$. Then, the initialized inter-modality consistency feature Token H_{c_0}, three intra-modality heterogeneity feature Tokens H_{a_0}, H_{v_0}, H_{t_0}, as well as the multimodal features U_0 are concatenated row by row to form a combined splicing Token tensor $con(H_{c_0}, H_{a_0}, H_{v_0}, H_{t_0}, U_0)$, where $con()$ represents the concatenation operation. The combined splicing Token tensor is fed into the Token separation encoder, denoted as $\mathcal{T}()$. Through properly training, the $\mathcal{T}()$ aims to extract four separated Tokens, namely H_c, H_a, H_v, H_t, from the combined tensor. The process can be written as:

$$H_c, H_a, H_v, H_t = \mathcal{T}(q/k/v = con(H_{c_0}, H_{a_0}, H_{v_0}, H_{t_0}, U_0)), \tag{1}$$

where q, k, v represent query, key, and value tensors in the Transformer encoder, respectively. Specifically, the query, key, and value tensors share the same initialized Token and multimodal features, which have a shape of $(3T_o + 12) \times d$.

Following a typical Transformer encoder structure (Vaswani et al., 2017), the $\mathcal{T}()$ consists of two layers of multi-head self-attention, normalization layers, and MLP layers, as illustrated in Fig. 3(a). In addition, following Dosovitskiy et al., we perform a learnable positional encoding of the timestamp of the sequence and add it to the input of the Transformer. We use Tokens as additional learnable information to learn decoupling features. During training, the $\mathcal{T}()$ translates the initialized T_k and U_0 into the separated inter-modality consistency feature Token and intra-modality heterogeneity feature Tokens that can be defined explicitly. To help the Token separation encoder satisfy the disentanglement, we devise a Token disentanglement regularization including three joint multimodal emotion Token disentanglement losses (see Section 3.2.2 for details).

3.2.2. Token disentanglement regularization

In order to make the $\mathcal{T}()$ satisfy the above Token separation, we devise a Token disentanglement regularization with three multimodal emotion Token disentanglement losses, represented as the intra-modality similarity loss $\mathcal{L}^m_{sim}$, inter-modality orthogonal loss $\mathcal{L}^m_{orth}$, and the multimodal disentanglement loss $\mathcal{L}^c_{dis}$, to guide the Token disentanglement. More specifically, the intra-modality similarity loss $\mathcal{L}^m_{sim}$ is used to make the intra-modality heterogeneity feature Tokens H_a, H_v, H_t as consistent as possible with the corresponding input multimodal features U_a, U_v, U_t, respectively, such that H_a,

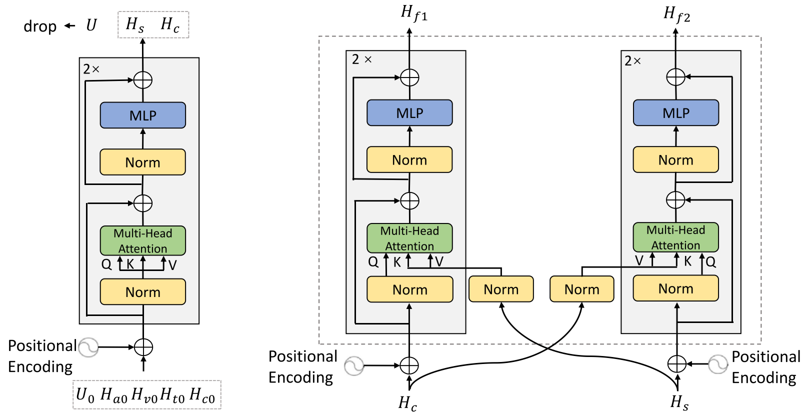


Fig. 3. The Transformer structure in the TMT. (a) The detailed structure of the Token separation encoder, (b) The architecture of the cross-modal encoder in Token mutual Transformer.

H_v, and H_t contain more modality-specific information. The inter-modality orthogonal loss $\mathcal{L}^m_{orth}$ is further introduced to make the H_v, H_a, H_t independent of each other. Both the $\mathcal{L}^m_{orth}$ and $\mathcal{L}^m_{sim}$ optimize the disentangled emotion-related intra-modality heterogeneity feature Tokens. Moreover, the multimodal disentanglement loss $\mathcal{L}^c_{dis}$ is used to separate the H_c from H_a, H_v, H_t by pulling the distribution of H_c away from H_a, H_v, and H_t as much as possible. By co-learning these three loss terms, the emotion-related inter-modality consistency feature Token H_c can contain as many common features as possible, while the intra-modality heterogeneity feature Tokens H_a, H_v, H_t contain unique emotion information within a specific modality, respectively. The detailed learning process for the three loss terms is as follows.

Intra-modality similarity loss. To make the intra-modality heterogeneity feature Tokens H_a, H_v, H_t and the corresponding original input features U_a, U_v, U_t as similar as possible, respectively, we employ the intra-modality similarity loss $\mathcal{L}^m_{sim}$ to learn more modality-specific information for the heterogeneity feature Tokens. To achieve this, $\mathcal{L}^m_{sim}$ introduces the Maximum Mean Discrepancy (MMD) (Borgwardt et al., 2006) to calculate and constrain the distribution similarity between the Token and its corresponding modality features, which can be described as:

$$\mathcal{L}^m_{sim} = \frac{1}{N}\sum_{i=1}^{N}(\mathrm{MMD}(U_t, H_t) + \mathrm{MMD}(U_a, H_a) + \mathrm{MMD}(U_v, H_v)), \tag{2}$$

where MMD() represents the MMD loss that measures the differences between two input features within the dataset. N signifies the total number of training data. Mathematically, the MMD() is defined as:

$$\mathrm{MMD}(p, q) = \|\phi(p) - \phi(q)\|_{\mathcal{H}}, \tag{3}$$

where p, q are input feature vectors in this study, which can be represented by the modality features of each data sample and its corresponding intra-modality emotion heterogeneity feature Token. ϕ represents a mapping from the original feature space to the reproducing kernel Hilbert space $\mathcal{H}$ (Borgwardt et al., 2006). This space mapping refers to mapping the low-dimensional space where the sample points cannot be separated by a straight line to the high-dimensional space where the sample points can be separated by a plane. A reproducing kernel Hilbert space is a function space that satisfies the condition of mapping from the current d-dimensional space to a Hilbert space. By minimizing the loss $\mathcal{L}^m_{sim}$, the distribution differences between the features of each modality and its corresponding intra-modality heterogeneity feature Token will be as similar as possible, so that H_a, H_v, and H_t contain more modality-specific information.

Inter-modality orthogonal loss. Moreover, to disentangle the intra-modality heterogeneity feature Tokens more properly, we employ the inter-modality orthogonal loss $\mathcal{L}^m_{orth}$ to make the emotion-related heterogeneity features of different modalities uncorrelated with each other, *i.e.*, independent of each other. As a result, we assert that H_a, H_v, and H_t are expected to be orthogonal to each other. To this end, inspired by Liu et al. (2017), the inter-modality orthogonal loss $\mathcal{L}^m_{orth}$ is represented as:

$$\mathcal{L}^m_{orth} = \frac{1}{N}\sum_{i=1}^{N}(\|H_v^{\top} H_a\|_2^2 + \|H_v^{\top} H_t\|_2^2 + \|H_a^{\top} H_t\|_2^2), \tag{4}$$

where $\|\|_2^2$ represents the orthogonal loss and is used to calculate the orthogonal projection loss between each pair of intra-modality emotion heterogeneity feature Token. For example, the $\|H_a^{\top} H_t\|_2^2$ refers to the orthogonal constraint between the intra-modality heterogeneity feature Tokens of the text modality and the audio modality. As $\|H_a^{\top} H_t\|_2^2$ gets smaller, the correlation between H_a and H_t gets smaller. This means that much of the emotion information unique to each modality is retained. During the learning process, minimizing $\mathcal{L}^m_{orth}$ serves to encourage the dot-products of heterogeneous feature tokens from different modalities to approach zero. This ensures that they become orthogonal to each other and do not incorporate emotion information from other modalities.

Multimodal disentanglement loss. To separate the inter-modality emotion consistency feature Token from the intra-modality emotion heterogeneity feature Token, we devise the multimodal disentanglement loss $\mathcal{L}^c_{dis}$ to make the disentangled inter-modality emotional consistency feature Token H_c and intra-modality emotion heterogeneity feature Token H_a, H_v, H_t independent of each other. To achieve this, following the orthogonal loss (Liu et al., 2017), the $\mathcal{L}^c_{dis}$ is denoted as follows:

$$\mathcal{L}^c_{dis} = \frac{1}{N}\sum_{i=1}^{N}(\|H_c^{\top} H_a\|_2^2 + \|H_c^{\top} H_v\|_2^2 + \|H_c^{\top} H_t\|_2^2). \tag{5}$$

Through training, the Token separation encoder can effectively disentangle inter-modality consistent feature Token from each intra-modality heterogeneity feature Token by forcing the dot-products of each pair of features to converge to zero.

3.3. Token mutual Transformer

With the disentangled emotion-related inter-modality consistency feature Token and each intra-modality heterogeneity feature Token, by formulating feature Tokens into query, key, and value representations, the normal Transformer can model the relations between query and key to achieve effective fusion between query and value. It is important to obtain a holistic and robust query representation for the Transformer to achieve promising emotion recognition performance. However, we found that existing Transformer methods mainly rely on the query representation from a specific modality, which could tend to focus on emotion features associated with the query modality representation and ignore subtle emotion interactions, resulting in incomplete fusion and affecting the recognition performance. To address this limitation and obtain more comprehensive and robust emotion representation, we further devise the Token mutual Transformer to completely explore the contributions of different disentangled features in terms of emotion interactions for valid information fusion by using bi-directional query learning in two cross-modal encoders, resulting in more robust emotion recognition performance. The detailed architecture of the Token mutual Transformer is shown in Fig. 3(b).

Formally, given three disentangled intra-modality heterogeneity feature Tokens H_a, H_v, H_t as input, we first splice them to form the heterogeneity feature tensor denoted as $H_s = \{con(H_a, H_v, H_t)\}$. Then, we employ two parallel cross-modal encoders with bi-directional query learning to fully fuse and interact the emotion information of H_s and H_c. The two parallel cross-modal encoders can be represented as $\mathcal{T}_{c\rightarrow s}()$ and $\mathcal{T}_{s\rightarrow c}()$, respectively. They use the H_s and H_c as the query representations, respectively, to fuse more comprehensive and complementary emotion representations by adaptive mutual learning. Each cross-modal encoder follows the typical Cross-Transformer encoder structure (Tsai et al., 2019b) and includes multi-head self-attention, normalization layers, and MLP layers.

More specifically, using the intra-modality heterogeneity feature tensor H_s as query and inter-modality consistency feature Token H_c as value and key, the cross-modal encoder $\mathcal{T}_{c\rightarrow s}()$ is used for heterogeneity to consistent emotion fusion, which tends to explore the contribution of the heterogeneity features for emotion interactions. $\mathcal{T}_{c\rightarrow s}()$ uses the query tensor as a reference and transforms the value tensor into the desired output based on the relations between the q and k. When transforming, multi-head attention mechanisms are performed to achieve relation modeling and data fusion. The learning process of the $\mathcal{T}_{c\rightarrow s}()$ can be implemented as:

$$H_{f1} = \mathcal{T}_{c\rightarrow s}(q = H_s, k/v = H_c), \tag{6}$$

where H_{f1} is the heterogeneity-to-consistent fusion features that learn more meaningful information related to emotions by reconstructing the emotion information of H_s and receiving the information from H_c.

Moreover, using the disentangled H_c as query and H_s as key and value, we follow the other cross-modal encoder $\mathcal{T}_{s\rightarrow c}()$ for consistent to heterogeneity emotion fusion, which tends to explore the contribution of the consistent features for emotion interactions. We implement the $\mathcal{T}_{s\rightarrow c}()$ as:

$$H_{f2} = \mathcal{T}_{s\rightarrow c}(q = H_c, k/v = H_s), \tag{7}$$

where H_{f2} is the consistent-to-heterogeneity fusion features.

In conclusion, through the bi-directional query learning in the two cross-model encoders, we can obtain two fused emotion representations, namely, H_{f1}, H_{f2}, which can fully explore the subtle emotion interactions from different views. Finally, we use a simple summation operation to further integrate the H_{f1}, H_{f2}, thus obtaining the final comprehensive multimodal emotion representation Y_o for emotion classification. In practice, we also used a more complicated attention mechanism to integrate the two learned fused representations, and the results in the ablation study (see Section 4.5.7) show that simple summation fusion obtains the optimal performance.

3.4. Overall learning objectives

For training, the TMT model includes four learning objectives, namely the intra-modality similarity loss $\mathcal{L}^m_{sim}$, inter-modality orthogonal loss $\mathcal{L}^m_{orth}$, multimodal disentanglement loss $\mathcal{L}^c_{dis}$, and emotion prediction loss $\mathcal{L}_{\text{pre}}$. Since the emotion labels in this study are multiple, *i.e.*, including both discrete emotion category labels and continuous emotion ratings. Therefore, we introduce the cross-entropy loss (Tsai et al., 2019a) as the emotion learning loss $\mathcal{L}_{\text{pre}}$ for classification and the mean square error (MSE) (Hazarika et al., 2020b) as the emotion learning loss $\mathcal{L}_{\text{pre}}$ for regression. Mathematically, the $\mathcal{L}_{\text{pre}}$ is given by:

$$\mathcal{L}_{pre} = \begin{cases} -\frac{1}{N}\sum_{i=0}^{N} y_i \cdot \log \hat{y}_i & \text{for classification} \\ \\ \frac{1}{N}\sum_{i=0}^{N} \|y_i - \hat{y}_i\|_2^2 & \text{for regression} \end{cases} \tag{8}$$

where y_i represents the prediction results, $\hat{y}_i$ represents the actual labels, and N signifies the number of training samples. Mathematically, the overall learning objectives of TMT can be written as:

$$\mathcal{L} = \alpha\mathcal{L}^m_{sim} + \beta(\mathcal{L}^m_{orth} + \mathcal{L}^c_{dis}) + \mathcal{L}_{pre} \tag{9}$$

where α, β is the weight distribution of each subspace constraint loss to balance multi-task learning.

4. Experiments

4.1. Datasets

We evaluated the TMT model using three widely used multimodal emotion datasets: CMU-MOSI (Zadeh et al., 2016a), CMU-MOSEI (Zadeh et al., 2018b), and CH-SIMS (Yu et al., 2020).

The CMU-MOSI dataset (Zadeh et al., 2016b) is a widely used multimodal emotion recognition dataset that includes video, text, and audio modalities. It consists of 89 videos from 89 speakers, with a total of 2,199 video clips. Each segment has been manually annotated with an emotion score ranging from −3 to 3, categorized into seven levels. Specifically, −3 indicates extremely negative emotions, while 3 signifies extremely positive emotions.

The CMU-MOSEI dataset (Zadeh et al., 2018b) is among the most extensive multimodal datasets, encompassing video, text, and audio modalities. It includes videos from over 1,000 narrators and has 23,453 video clips. The dataset provides two labels of sentiment and emotion, with emotions categorized into six categories: anger, happiness, sadness, surprise, fear, and disgust. The sentiment strength is annotated on a continuous scale ranging from −3 to 3, covering emotions from extremely negative to extremely positive. The video and audio features are sampled at 15 Hz and 20 Hz, respectively.

The CH-SIMS dataset (Yu et al., 2020) is a popular Chinese fine-grained multimodal emotion dataset that contains 2,281 refined video clips with three modalities: video, text, and audio. CH-SIMS annotates each modality separately to better capture their interactions, unlike the previous two datasets. The labels range from −1 to 1 and are divided into negative, neutral, and positive categories with 11 different strengths. Negative emotions include −1.0 and −0.8; weak negative emotions include −0.6, −0.4, and −0.2; neutral is labeled as 0; weak positive emotions include 0.2, 0.4, and 0.6, and positive emotions include 0.8 and 1.0.

4.2. Implementation details

The experiments described in this paper were conducted using PyTorch 1.11.0 and were trained on an NVIDIA GeForce RTX 2080Ti GPU. Both the Token separation encoder in the multimodal emotion Token disentanglement module and the cross-modal Transformer encoder in the Token mutual Transformer module have two layers. Following the typical CrossTransformer encoder structure (Tsai et al., 2019a), each layer of these encoders includes a 8-head multi-head self-attention, 2-level normalization, and a 256-dim MLP. A batch size of 16 was used for all three datasets: CMU-MOSI, CMU-MOSEI, and CH-SIMS. The Adam optimizer was employed for learning, with a learning rate of 0.001 for the CMU-MOSI dataset and 0.002 for the CMU-MOSEI and CH-SIMS datasets. The inter-modality emotion consistency feature Token has a dimension of 6 × 256, while each modality's heterogeneity features Token has a dimension of 2 × 256.

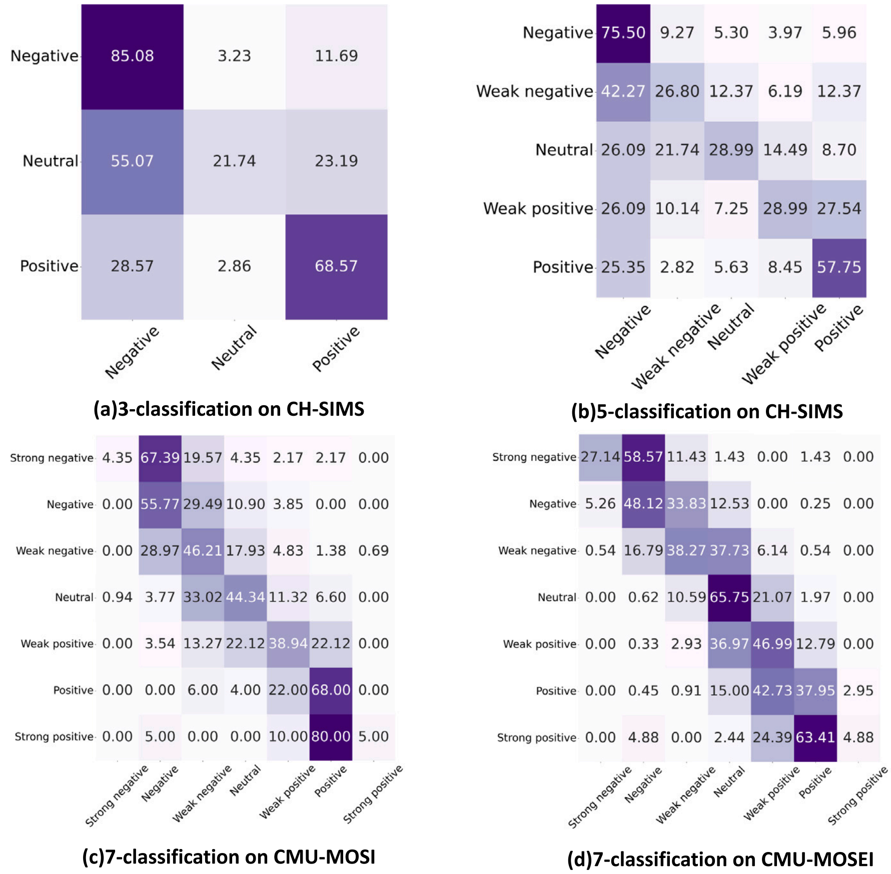


Fig. 4. The Confusion matrix of TMT on CMU-MOSI/CMU-MOSEI/CH-SIMS datasets, respectively.

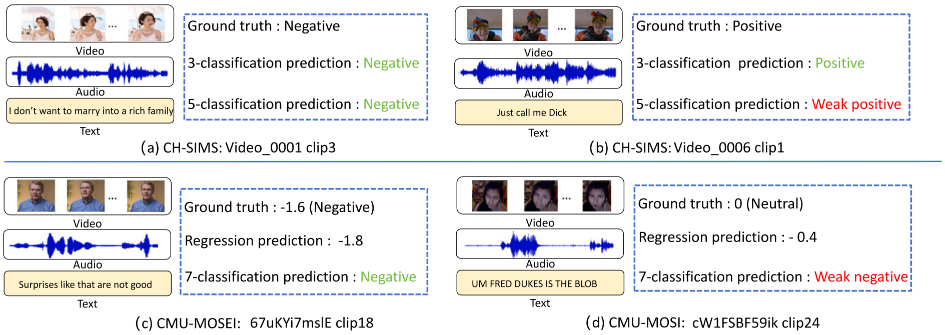

Video
Audio
I don't want to marry into a rich family
Text
Ground truth : Negative
3-classification prediction : Negative
5-classification prediction : Negative
(a) CH-SIMS: Video_0001 clip3
Video
Audio
Just call me Dick
Text
Ground truth : Positive
3-classification prediction : Positive
5-classification prediction : Weak positive
(b) CH-SIMS: Video_0006 clip1
Video
Audio
Surprises like that are not good
Text
Ground truth : -1.6 (Negative)
Regression prediction : -1.8
7-classification prediction : Negative
(c) CMU-MOSEI: 67uKYi7mslE clip18
Video
Audio
UM FRED DUKES IS THE BLOB
Text
Ground truth : 0 (Neutral)
Regression prediction : - 0.4
7-classification prediction : Weak negative
(d) CMU-MOSI: cW1FSBF59ik clip24

Fig. 5. Visualization of the samples predicted by the TMT model, where the highlighted ones in green indicate correctly predicted results and the highlighted ones in red indicate incorrectly predicted results.

4.3. Overall performance

The confusion matrices of TMT on CMU-MOSI/CMU-MOSEI/CH-SIMS datasets are shown in Fig. 4, respectively. Furthermore, we also visualized some prediction results on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets for analysis in Fig. 5. As can be seen in Fig. 5(a) and Fig. 5(c), our model predicts accurate results on both classification and regression tasks which proves the robustness of our model. As depicted in Fig. 5(b), for the sample Video_0006 clip1, although our model might make an erroneous prediction of 'Weak positive' in the 5-classification task, it accurately predicts the correct classification in the 3-classification task. This results in an accuracy of 68.57% for correctly predicting the 'Positive' class in the three-classification task (see Fig. 4(a)), but decreases to 57.75% in the 5-classification task depicted in Fig. 4(b). Additionally, discrepancies between regression values and the ground truth in the CMU-MOSI dataset may contribute to the misclassification in the classification task. From Fig. 5(d), it can be observed that on the sample cW1FSBF59ik clip24, our model's regression prediction is −0.4, while the ground truth is 0. This also leads to an error in the 7-classification task. Consequently, the misclassification of Neutral as Weak negative occurs at a rate of 33.02% in Fig. 4(c).

In summary, while our model has achieved promising results in both classification and regression tasks, there is room for improvement, especially in high-dimensional classification tasks, to enhance robustness.

Table 2
Comparison results on the CMU-MOSI dataset.

	CMU-MOSI				
Model	Acc-2 ↑	F1 ↑	Acc-7 ↑	MAE ↓	Corr ↑
MFM (Tsai et al., 2019c)	81.7	81.6	35.4	0.877	0.706
TFN (Zadeh et al., 2017b)	80.8	80.7	34.9	0.901	0.698
ICCN (Sun et al., 2020b)	83.0	83.0	39.0	0.860	0.710
MulT (Tsai et al., 2019b)	83.0	82.8	42.0	0.871	0.698
MISA (Hazarika et al., 2020b)	83.4	83.6	42.3	0.783	0.761
Self-MM (Yu et al., 2021)	86.0	86.0	45.8	**0.713**	0.798
CubeMLP (Sun et al., 2022)	85.6	85.5	45.5	0.770	0.760
TETFN (Wang et al., 2023)	86.1	86.1	–	0.717	0.800
TMT	**86.4**	**86.5**	**47.3**	0.718	**0.801**

Table 3
Comparison results on the CMU-MOSEI dataset.

	CMU-MOSEI				
Model	Acc-2 ↑	F1 ↑	Acc-7 ↑	MAE ↓	Corr ↑
MFM (Tsai et al., 2019c)	84.4	84.3	51.3	0.568	0.717
TFN (Zadeh et al., 2017b)	82.5	82.1	50.2	0.593	0.700
ICCN (Sun et al., 2020b)	84.2	84.2	51.6	0.565	0.713
MulT (Tsai et al., 2019b)	82.5	82.3	51.8	0.580	0.703
MISA (Hazarika et al., 2020b)	85.5	85.3	52.2	0.555	0.756
Self-MM (Yu et al., 2021)	85.2	85.3	53.5	0.530	0.765
CubeMLP (Sun et al., 2022)	85.1	84.5	–	**0.529**	0.760
TETFN (Wang et al., 2023)	85.2	85.3	–	0.551	0.748
TMT	**86.5**	**86.5**	**53.7**	0.542	**0.775**

Table 4
Comparison results on the CH-SIMS dataset.

	CH-SIMS					
Model	Acc-2 ↑	F1 ↑	Acc-3 ↑	F1 ↑	Acc-5 ↑	F1 ↑
MFN (Zadeh et al., 2018a)	78.26	78.62	65.79	–	41.19	–
TFN (Zadeh et al., 2017b)	**82.06**	**82.72**	66.16	–	39.74	–
LMF (Liu et al., 2018)	79.74	80.56	66.48	–	39.74	–
MulT (Tsai et al., 2019b)	78.84	80.00	67.13	–	38.24	–
MISA (Hazarika et al., 2020b)	80.09	79.49	64.99	62.25	41.79	41.23
Self-MM (Yu et al., 2021)	80.04	80.44	65.47	–	41.53	–
ALMT (Zhang et al., 2023)	81.19	81.57	**68.93**	–	45.73	–
TMT	80.53	81.11	68.71	**74.30**	**48.14**	**58.62**

4.4. Evaluation protocals

To comprehensively evaluate the model, we utilized the Pearson correlation coefficient (Corr) and mean absolute error (MAE) as evaluation indicators. Additionally, Acc-2 and F1 were employed as binary classification evaluation criteria. Acc-2 was evaluated using negative/non-negative and negative/positive classifications. For the 7-classification task, we employed the Acc-7 metric for evaluation. It is worth noting that MAE is the only indicator where a smaller value is considered better, while the other four indicators benefit from larger values.

In this paper, the proposed TMT model was employed and compared across three extensively utilized multimodal emotion datasets: CMU-MOSI (Zadeh et al., 2016b), CMU-MOSEI (Zadeh et al., 2018b), and CH-SIMS (Yu et al., 2020). Our experimental results indicate that our method surpasses the performance of existing state-of-the-art methods, including TFN (Zadeh et al., 2017b), MFM (Tsai et al., 2019c), MULT (Tsai et al., 2019b), MISA (Hazarika et al., 2020b),ICCN (Sun et al., 2020b), Self-MM (Yu et al., 2021), CubeMLP (Sun et al., 2022) and TETFN (Wang et al., 2023) on these three datasets. Fig. 4 illustrates the confusion matrix of TMT for 3-class and 5-class emotion recognition on the CH-SIMS dataset and 7-class emotion recognition on the CMU-MOSI and CMU-MOSEI datasets, respectively.

4.4.1. Comparison on the CMU-MOSI dataset

Table 2 presents the comparison results on the CMU-MOSI dataset. The TMT model demonstrates superior performance compared to the current state-of-the-art models across nearly all indicators. The Acc-2 accuracy achieves 86.43%, the F1 score reaches 86.5, and the Acc-7 accuracy attains 47.3%. Moreover, the MAE value is reduced to 0.718, indicating improved accuracy, and the correlation coefficient (Corr) is enhanced to 0.801. Our TMT method outperforms Self-MM by 1.5% and CubeMLP by 1.8% on the Acc-7 metric. Compared to TETFN, our TMT achieves state-of-the-art performance on all metrics. This is attributed to our Token-learning-based disentangling approach, which has learned a more comprehensive multimodal representation, facilitating the subsequent fusion process. These advancements signify significant improvements in the classification accuracies for both two and seven categories compared to the previous state-of-the-art models.

4.4.2. Comparison on the CMU-MOSEI dataset

Table 3 displays the comparison results of our proposed TMT method and existing models on the CMU-MOSEI dataset. Our TMT model outperformed the current state-of-the-art models in almost all indicators. The Acc-2 accuracy of our model is 86.5%, which is 2.1% higher than MFM's (Tsai et al., 2019c) accuracy. Furthermore, our model achieves an Acc-7 accuracy of 53.7%, surpassing MISA's (Hazarika et al., 2020b) accuracy by 1.6%. These results demonstrate the superiority of our approach on the CMU-MOSEI dataset. Our TMT method outperforms Self-MM by 0.01 and CubeMLP by 0.015 on the Corr metric. Compared to the sub-optimized result TETFN, TMT achieves state of the art performance on all metrics. The reason for our leading position is that our TMT simplifies the complexity of the model through Token-learning-based disentangling approach while obtaining a more comprehensive representation. Additionally, it complements and fully integrates the disentangled features.

4.4.3. Comparison on the CH-SIMS dataset

Table 4 presents the comparison results of our proposed method and existing models on the CH-SIMS dataset. Our approach outperformed the current mainstream model in all indicators, with TMT achieving the best performance. Specifically, our model achieved an Acc-5 accuracy of 48.14%, significantly higher than the current state-of-the-art model MISA (Hazarika et al., 2020b) by 6.35%. Our model achieves 3.24% and 6.61% improvement over Self-MM on Acc-3 and Acc-5 metrics, respectively. Compared with the latest method ALMT, our TMT is still 2.41% ahead on Acc-5. This is because our TMT method fully disentangles and fuses the interacting inter-modality emotion consistency features and intra-modality emotion heterogeneity features. This demonstrates that our method, which utilizes multimodal emotion Token disentanglement learning to obtain a comprehensive and robust emotion representation, improves the accuracy of multimodal emotion recognition.

4.5. Ablation studies

To comprehensively assess the effectiveness of our proposed model, we have provided detailed ablation results for the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets in the subsequent sections. In the following sections, we provide a comprehensive description of the results obtained from the ablation experiments.

4.5.1. Effects of disentangled feature Tokens

To demonstrate the effectiveness of the disentangled inter-modality emotion consistency feature Token and intra-modality emotion heterogeneity feature Token, we conducted ablation experiments where we removed either the consistency features or heterogeneity features from the model to observe its performance in Table 5. During the multimodal emotion mutual Transformer, we only send the inter-modality

Table 5
Effects of disentangled feature Tokens.

Model	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	MAE ↓	Corr ↑	MAE ↓	Corr ↑	Acc-2 ↑	Acc-5 ↑
TMT	**0.718**	**0.801**	**0.542**	**0.775**	**80.53**	**48.14**
w/o-inter-modality emotion consistency feature Token	0.790	0.740	0.781	0.750	76.59	45.95
w/o-intra-modality emotion heterogeneity feature Token	0.750	0.760	0.693	0.760	77.02	46.14

Table 6
Effects of key modules in TMT.

Model	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	MAE ↓	Corr ↑	MAE ↓	Corr ↑	Acc-2 ↑	Acc-5 ↑
TMT	**0.718**	**0.801**	**0.542**	**0.775**	**80.53**	**48.14**
w/o- multimodal emotion Token disentanglement module	0.733	0.778	0.543	0.763	78.56	46.52
w/o- Token mutual Transformer module	0.747	0.780	0.544	0.771	77.62	46.24

emotion consistency feature Token (or intra-modality emotion heterogeneity feature Token) to the Token mutual Transformer module. The results in Table 5 show that removing either the consistency or heterogeneity features leads to a decrease in emotion recognition accuracy. When the emotion consistency features were discarded, the MAE on the CMU-MOSI/CMU-MOSEI dataset increased to 0.790/0.781, respectively. When the heterogeneity features were discarded, the MAE on the CMU-MOSI/CMU-MOSEI dataset increased to 0.750/0.690, respectively. The Acc-5 indicators on the CH-SIMS dataset were reduced by 2.19% and 2.0%, respectively. This indicates that both features disentangled by the multimodal emotion Token disentanglement module are meaningful. Besides, comparing the above two ablation experiments, it can be observed that the influence of retaining only consistency features is significantly smaller than that of retaining only heterogeneity features, so the emotion consistency features extracted by the model are more important for emotion recognition than the emotion heterogeneity features.

4.5.2. *Effects of key modules*

Impact of the multimodal emotion Token disentanglement. To evaluate the contribution of the multimodal emotion Token disentanglement module, we removed the emotion disentanglement module in TMT and directly replaced them with MLPs in MISA (Hazarika et al., 2020b) to obtain emotion representations. The experimental results are shown in the "w/o- multimodal emotion Token disentanglement module" in Table 6. The experimental results on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets reveal that removing the multimodal emotion token disentanglement module leads to a notable reduction in the model's recognition performance. Specifically, the relative MAE increases by 0.015 on CMU-MOSI, 0.001 on CMU-MOSEI, respectively. The Acc-2 indicator on the CH-SIMS dataset was reduced by 1.97%. This demonstrates that the multimodal emotion Token disentanglement module significantly improves the results of multimodal sentiment recognition due to more efficient learning of emotional information in the modality.

Impact of the Token mutual Transformer. To demonstrate the influence of the Token mutual Transformer module, we removed the fusion module in the TMT and replaced it with a simple row-wise splicing operation. Specifically, with the obtained disentangled features H_s and H_c, each having a dimension of $\mathrm{B} \times 6 \times 256$ (where B represents the batch size), we directly concatenate the H_s and H_c, resulting in dimensions $\mathrm{B} \times 12 \times 256$. Subsequently, we perform averaging along the first dimension to obtain the multimodal feature with the dimension of $\mathrm{B} \times 256$ for emotion prediction. The experimental results are shown in the "w/o-Token mutual Transformer module" in Table 6. The results show that the direct use of row-wise splicing operation in the later fusion makes the performance of the model decline in different magnitudes on these three datasets. Meanwhile, the MAE on the CMU-MOSI/CMU-MOSEI datasets increased by 0.029/0.002, respectively. The Acc-2 indicator on the CH-SIMS dataset was reduced by 2.90%. The experimental results prove that the Token mutual Transformer module is beneficial to improve the performance of the model.

Table 7
Influence of different regularization terms in TMT.

Model	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	MAE ↓	Corr ↑	MAE ↓	Corr ↑	Acc-2 ↑	Acc-5 ↑
TMT	**0.718**	**0.801**	**0.542**	**0.775**	**80.53**	**48.14**
w/o- $\mathcal{L}^{c}_{\mathrm{dis}}$	0.750	0.770	0.740	0.760	79.14	47.29
w/o- $\mathcal{L}^{m}_{\mathrm{orth}}$	0.780	0.760	0.783	0.764	77.62	45.87
w/o- $\mathcal{L}^{m}_{\mathrm{sim}}$	0.750	0.776	0.750	0.774	78.48	46.10

4.5.3. *Effects of different regularization terms*

We set up three loss functions to constrain the model to disentangle out emotion heterogeneity features as well as consistent features. To demonstrate the importance of different loss functions, we explored how the model performs without one of them. The effects of different regularization terms are shown in Table 7. From the Table 7, we can see that these three losses have different impacts on MAE performance on three different datasets. The effect variation can be attributed primarily to differences in dataset scales and collection backgrounds. For example, the CMU-MOSI dataset consists of 2199 samples, the CH-SIMS dataset contains 2281 samples, and the CMU-MOSEI dataset is larger with 23,453 samples. Furthermore, the CMU-MOSI and CMU-MOSEI datasets are collected from English video backgrounds, whereas CH-SIMS datasets are obtained from Chinese video backgrounds. These affect the learning effect of feature decoupling and fusion during training.

Without $\mathcal{L}^{c}_{dis}$: The model performance decreased without using the multimodal disentanglement loss, and the MAE on these CMU-MOSI/CMU-MOSEI datasets also increased to 0.770/0.760, respectively. The Acc-5 indicator on the CH-SIMS dataset was reduced by 0.85%. The possible reason is that the lack of $\mathcal{L}^{c}_{dis}$ results in insufficient decoupling between the consistent and heterogeneous features, leading to a significant decline in model performance.

Without $\mathcal{L}^{m}_{\mathrm{orth}}$: The absence of inter-modality orthogonal loss resulted in a decrease in model performance. For instance, the Mean Absolute Error (MAE) on the CMU-MOSI/CMU-MOSEI datasets increased to 0.780/0.783, respectively.The Acc-5 indicator on the CH-SIMS dataset was reduced by 2.27%. This occurred because the learned heterogeneity features among the modalities were not effectively separated, leading to a decline in model performance.

Without $\mathcal{L}^{m}_{sim}$: When the intra-modality similarity loss was removed, the MAE on the CMU-MOSI/CMU-MOSEI dataset also increased to 0.750/0.750, respectively. The Acc-5 indicator on the CH-SIMS dataset was reduced by 2.04%. The lack of MMD loss makes the data distribution of the intra-modality emotion heterogeneity feature Token too different from the data distribution of the input feature vector of the modality.

In addition, due to the significant differences in these datasets, we find that the three losses have also slightly different impacts on Corr performance. After removing the multimodal disentanglement loss

Table 8
Effects of different modalities.

Model	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	MAE ↓	Corr ↑	MAE ↓	Corr ↑	Acc-2 ↑	Acc-5 ↑
TMT	**0.718**	**0.801**	**0.542**	**0.775**	**80.53**	**48.14**
w/o-Audio	0.750	0.775	0.540	0.761	79.25	47.18
w/o-Text	0.780	0.750	0.750	0.761	77.58	46.01
w/o-video	0.761	0.750	0.548	0.760	78.46	46.78

$\mathcal{L}^{c}_{dis}$, inter-modality orthogonal loss $\mathcal{L}^{m}_{\text{orth}}$, and intra-modality similarity loss $\mathcal{L}^{m}_{sim}$, the Corr metric decreased by 0.031/0.041/0.025 on the CMU-MOSI dataset, by 0.015/0.011/0.001 on the CMU-MOSEI dataset, respectively.

4.5.4. *Effects of different modalities*

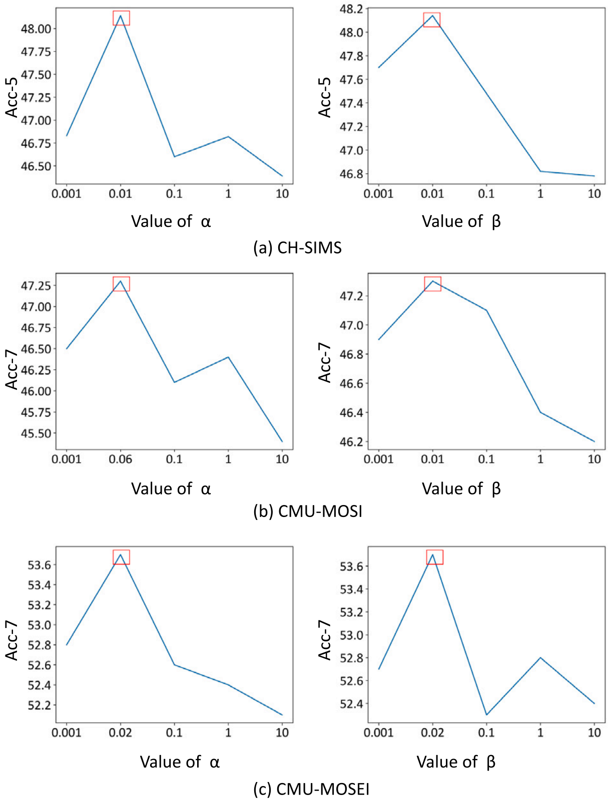
To assess the impact of different modalities on emotion recognition, we systematically excluded the combined feature vectors of the text, video, and audio modalities and analyzed their effects. The experimental results are displayed in Table 8. As shown in the Table 8, the removal of any modality leads to a lower correlation (Corr) and a higher Mean Absolute Error (MAE) for the model. Specifically, when the text modality is omitted, the Corr decreases by 0.051/0.014 on the CMU-MOSI/CMU-MOSEI datasets, and the MAE increases by 0.062/0.208, respectively. The Acc-5 indicator on the CH-SIMS dataset was reduced by 2.13%. These findings highlight the complementary nature of the interaction between the three modalities, leading to improved emotion recognition accuracy. Notably, the most significant drop in performance occurs when the text modality is removed. For example, in the CMU-MOSEI dataset, the MAE increased from 0.542 to 0.750. This suggests that textual information contains crucial emotion cues that dominate in multimodal emotion recognition. Hence, the removal of textual information has a substantial impact on the model's performance.

4.5.5. *Effects of different loss weight parameters*

Fig. 6 illustrates the key parameter selection in Eq. (9). We conducted extensive experiments with varying loss weight parameters, *i.e.*, α and β, on the CMU-MOSI, CMU-MOSEI and CH-SIMS dataset, respectively. As depicted in the figure, the highest accuracy (Acc-5) reached 48.14% when we set α to 0.01 and β to 0.01 on the CH-SIMS dataset. Acc-7 achieved 47.3% with α set to 0.06 and β to 0.01 on the CMU-MOSI dataset. Additionally, Acc-7 reached 53.7% when α was set to 0.02 and β to 0.02 on the CMU-MOSEI dataset. The red squares in the figure represent the best performance results with the optimal parameter choices. We observed that the performance indicator dropped when the parameter settings were either too large or too small.

4.5.6. *Effects of different disentanglement methods*

As shown in Table 9, we conducted a contrast experiment on CMU-MOSI using different decoupling methods to demonstrate the effectiveness of our Token Separation Encoder-based approach. MISA (Hazarika et al., 2020b) projects each modality into two distinct subspaces through the design of constraints and encoders. Yang et al. (2022a) have designed a Feature Separation Multimodal Emotion Recognition (FDMER) method, which addresses modality heterogeneity by projecting each mode into a modality-invariant subspace and a modality-specific subspace. By employing adversarial learning strategies, they refine the public and private representations of feature separation. MFSA (Yang et al., 2022b) introduces an approach for learning effective multimodal representations in asynchronous sequences, with a focus on feature disentanglement. The results demonstrate that our Token separation encoder outperforms other disentanglement methods. In terms of the Acc-7 indicator, it has a relative improvement of 11.8% over MISA, 14.3% over MFSA and 12.4% over FDMER on the CMU-MOSI dataset. In terms of the Corr indicator, it has a relative improvement of 2.5% over MISA, 7.0% over MFSA and 0.3% over FDMER on the CMU-MOSEI dataset. This can be attributed to the effectiveness of the Token learning for feature disentangling.



(a) CH-SIMS

(b) CMU-MOSI

(c) CMU-MOSEI

Fig. 6. Comparison results on the three datasets with different loss weight parameters α and β, the red squares in the figure represent the best performance results with the optimal parameter choices.

4.5.7. *Effects of different fusion methods*

To discuss the effects of different fusion methods, we compared different fusion technologies, including GRU (Chung et al., 2014), LSTM (Greff et al., 2017), and TFN (Zadeh et al., 2017b) for feature fusion on CMU-MOSI, CMU-MOSEI and CH-SIMS datasets, respectively. The details are shown in Table 10. Obviously, our TMT achieved the best performance when using Token mutual Transformer for feature fusion, demonstrating that Token mutual Transformer can obtain a more complementary feature for emotion recognition due to the use of bi-directional query learning to fully interact and integrate the emotion consistency and heterogeneity information by exploring their mutual contribution in emotion interactions.

4.5.8. *Effects of different feature integration mechanisms in Token mutual Transformer*

With the two bi-directional interaction features H_{f1}, H_{f2} obtained by the Token mutual Transformer module, we compared different feature integration mechanisms, including summation, concatenation, and self-attention for them on the three datasets. The self-attention fusion process follows a typical Transformer encoder mechanism, whose calculation equation can be seen in Table 11. Here the values of q, k and v result from the direct concatenation of H_{f1} and H_{f2}. The comparison results are shown in Table 11. The simple method of Summation we adopted can achieve better results on the three datasets. In addition, We observed that different fusion methods have little effect

Table 9
Comparison results on the CMU-MOSI and CMU-MOSEI dataset with different disentanglement methods.

Model	CMU-MOSI			CMU-MOSEI		
	Acc-7 ↑	MAE ↓	Corr ↑	Acc-7 ↑	MAE ↓	Corr ↑
MISA (Hazarika et al., 2020b)	42.3	0.783	0.761	52.2	0.555	0.756
MFSA (Yang et al., 2022b)	41.4	0.856	0.722	53.2	0.574	0.724
FDMER (Yang et al., 2022a)	44.1	0.724	0.788	**54.1**	**0.536**	0.773
TMT	**47.3**	**0.718**	**0.801**	53 .7	0.542	**0.775**

Table 10
Effects of different fusion methods.

Model	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	Acc-7 ↑	Corr ↑	Acc-7 ↑	Corr ↑	Acc-2 ↑	Acc-5 ↑
Token mutual Transformer(Ours)	**47.3**	**0.801**	**53.7**	**0.775**	**80.53**	**48.14**
LSTM (Greff et al., 2017)	44.2	0.785	51.4	0.763	78.77	45.73
GRU (Chung et al., 2014)	44.3	0.789	52.3	0.765	77.24	46.17
TFN (Zadeh et al., 2017b)	44.6	0.788	52.4	0.768	78.56	47.48

Table 11
Comparison results of different feature integration mechanisms in Token mutual Transformer on three datasets.

Model	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	Acc-7 ↑	Corr ↑	Acc-7 ↑	Corr ↑	Acc-2 ↑	Acc-5 ↑
Summation(Ours)	**47.3**	**0.801**	**53.7**	**0.775**	**80.53**	48.14
Concatenation	45.8	0.792	52.4	0.768	80.35	**49.01**
Self-attention (softmax $\left(\frac{qk^T}{\sqrt{d_k}}\right)v$)	46.7	0.793	52.5	0.764	80.50	48.36

Table 12
Effects of the dimension of each input feature vector.

Model	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	Acc-7 ↑	Corr ↑	Acc-7 ↑	Corr ↑	Acc-2 ↑	Acc-5 ↑
256(Ours)	**47.3**	**0.801**	**53.7**	**0.775**	**80.53**	**48.14**
128	44.9	0.786	52.2	0.761	79.87	47.48
512	46.2	0.798	51.5	0.763	78.34	47.26

on the final results, demonstrating that the H_{f1}, H_{f2} have learned the comprehensive and complementary emotion information.

4.5.9. Effects of the dimension of each input feature vector

As shown in Table 12, in order to verify the influence of the sequence dimensions of each input feature vector on the experimental results, different sequence dimensions including 128, 256, 512 were selected for the experiment, respectively. The experimental results show that the overall performance of the model on the three datasets is best when the value of the sequence dimension of each input feature vector is set to 256. Therefore, considering the trade-off between efficiency and accuracy, we empirically set the value of d to 256 (see Table 12).

4.5.10. Effects of different numbers of layers in the Token separation encoder

As shown in Table 13, we present the experimental results of different numbers of layers in the Token separation encoder on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, respectively. We find that the best performance was obtained when choosing 2 layers in the Token separation encoder. It is worth noting that when the model's layer number is set to 4, although the model achieves the best performance in Acc7 on the CMU-MOSI dataset, its performance in Corr is comparatively poor. Therefore, based on these observations, we empirically choose a model with two layers.

4.5.11. Effects of different numbers of layers in the Token mutual Transformer

As shown in Table 14, we presents the experimental results of different numbers of layers in the Token mutual Transformer on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, respectively. We conducted ablation experiments with layer numbers set to 1, 2, 4 and 8. The results indicated that the Token mutual Transformer with two-level encoders achieved the optimal performance.

4.5.12. Analysis on model complexity

As shown in Table 15, we compared the parameters and FLOPs (Molchanov et al., 2017) of TMT with other state-of-the-art Transformer-based methods. The different hyper-parameter configurations for each dataset may lead to a slight difference in parameter and complexity calculation. We calculated the model parameters and FLOPs under the hyper-parameter settings on the CMU-MOSI. Following Liu et al. (2023c), we calculated the modality disentangling and the modality fusion modules, except for the pre-trained Bert text extractor. In the comparative analysis, our proposed Token-disentangling Mutual Transformer (TMT) demonstrates notable advancements over the competing methods, MuLT and MISA. TMT achieves a substantial improvement in both parameter efficiency and computational complexity, with a reduced parameter count of 1.31M. This improvement is attributed to its straightforward approach of the Token disentanglement module that does not require additional parameter learning. Specifically, we learn the decoupled consistency features and heterogeneity features through the initialized Tokens. The dimension of inter-modality consistency feature Token is 6 × 256 in this paper, and the dimension of the other three intra-modality heterogeneity feature Tokens is 2 × 256. This way, fewer parameters are used to store features for learning. This approach utilizes fewer parameters to store features for learning. Consequently, compared to other methods, our TMT can maintain higher accuracy and efficiency.

Table 13
Effects of different numbers of layers in the Token separation encoder.

Model	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	Acc-7 ↑	Corr ↑	Acc-7 ↑	Corr ↑	Acc-2 ↑	Acc-5 ↑
2(Ours)	47.3	**0.801**	**53.7**	**0.775**	**80.53**	**48.14**
1	45.6	0.785	50.9	0.765	79.43	45.95
4	**48.1**	0.792	52.4	0.763	80.09	47.92
8	43.4	0.801	51.3	0.766	80.31	46.17

Table 14
Effects of different numbers of layers in the Token mutual Transformer.

Model	CMU-MOSI		CMU-MOSEI		CH-SIMS	
	Acc-7 ↑	Corr ↑	Acc-7 ↑	Corr ↑	Acc-2 ↑	Acc-5 ↑
2(Ours)	**47.3**	**0.801**	**53.7**	**0.775**	**80.53**	**48.14**
1	45.5	0.790	51.4	0.763	79.87	45.95
4	44.0	0.797	52.3	0.763	78.99	47.48
8	45.4	0.795	52.5	0.769	79.65	47.26

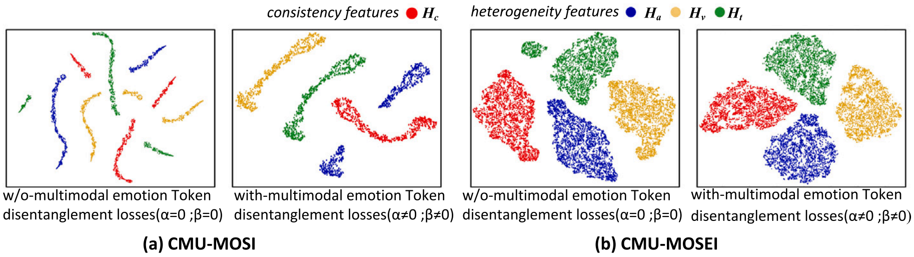


Fig. 7. Effects of the proposed multimodal emotion Token disentanglement losses for feature disentanglement. (a) Feature visualization on the CMU-MOSI dataset, (b) Feature visualization on the CMU-MOSI dataset. H_c, H_a, H_v, and H_t represent the emotion consistent features, audio heterogeneity features, video heterogeneity features, and text heterogeneity features, respectively. Among them, $\alpha \neq 0, \beta \neq 0$ represents the disentangled features with using the multimodal emotion Token disentanglement losses in Eq. (9), while $\alpha = 0, \beta = 0$ represents the disentangled features without using the multimodal emotion Token disentanglement losses, *i.e.*, removing the multimodal disentanglement loss, inter-modality orthogonal loss and intra-modality similarity loss from the Eq. (9).

Table 15
Analysis on model complexity.

Method	Parameters	FLOPs	Acc-7 on MOSI
MuLT (Tsai et al., 2019b)	2.57M	1.9G	42.0
MISA (Hazarika et al., 2020b)	3.10M	1.7G	42.3
TMT	**1.31M**	**1.2G**	**47.3**

Note: the parameters of other Transformer-based methods was calculated by authors from open source code with default hyper-parameters on CMU-MOSI.

4.6. Visualization and analysis

4.6.1. Visualization of disentangled features with different disentanglement losses

In order to verify the effectiveness of the proposed multimodal emotion Token disentanglement losses, Fig. 7 shows the visualization results of the disentangled emotion heterogeneity features and emotion consistent features by using t-SNE (van der Maaten and Hinton, 2008), on the CMU-MOSI and CMU-MOSI datasets, respectively. In this figure, $\alpha \neq 0, \beta \neq 0$ represents the disentangled features with using the multimodal emotion Token disentanglement losses; $\alpha = 0, \beta = 0$ represents the disentangled features without using the multimodal emotion Token disentanglement losses (*i.e.*, removing the multimodal disentanglement loss, inter-modality orthogonal loss and intra-modality similarity loss). Here, we select all sample disentangled features of the whole test set for visualization.

As depicted in Fig. 7, the introduction of multimodal emotion Token disentanglement losses effectively disentangle the consistency features (H_c) from heterogeneity features (H_a, H_v and H_t). It is worth noting that the audio heterogeneity features are not effectively aggregated, possibly due to their heightened sensitivity to emotion-irrelevant information (namely noises) compared to video and text features. Moreover, the selected Transformer architecture in our model may be more conducive to feature learning in vision and text modalities. In addition, we observed that scale differences between the CMU-MOSI and CMU-MOSEI datasets lead to variations in separation degrees. The larger CMU-MOSEI dataset enables the model to train on more samples, achieving a higher degree of separation compared to the CMU-MOSI dataset.

4.6.2. Visualization of the loss functions

In Fig. 8, we visualized the variations of inter-modality orthogonal loss $\mathcal{L}^{m}_{\text{orth}}$, multimodal disentanglement loss $\mathcal{L}^{c}_{dis}$, intra-modality similarity loss $\mathcal{L}^{m}_{sim}$, emotion prediction loss $\mathcal{L}_{\text{pre}}$ and the overall learning loss $\mathcal{L}$ of TMT during training. The blue lines represent the learning process of each loss function on the CMU-MOSI dataset, while the red lines represent the learning process on the CMU-MOSEI dataset. From the figures, it can be observed that all loss functions demonstrate clear convergence by the 20th epoch. This indicates efficient and rapid training of our proposed model, as well as its strong generalization ability.

4.6.3. Visualization of attention in Token mutual Transformer

To assess the model's ability to integrate heterogeneous and consistent features, we visualized the learned average attention matrix for the inter-modality consistent feature H_c and intra-modality heterogeneity feature H_s in the Token mutual Transformer module on the CMU-MOSI dataset, as shown in Fig. 9. Specifically, we randomly selected a sample to visualize the multi-head average attention matrixes of the last layer in the Token mutual Transformer, which are represent as $\mathcal{T}_{c\rightarrow s}()$ and $\mathcal{T}_{s\rightarrow c}()$. Each grid in Fig. 9 represents the mutual attention weight between heterogeneity features and consistency features, with darker colors indicating higher attention weights. Higher attention weights represent that one decoupled feature learns

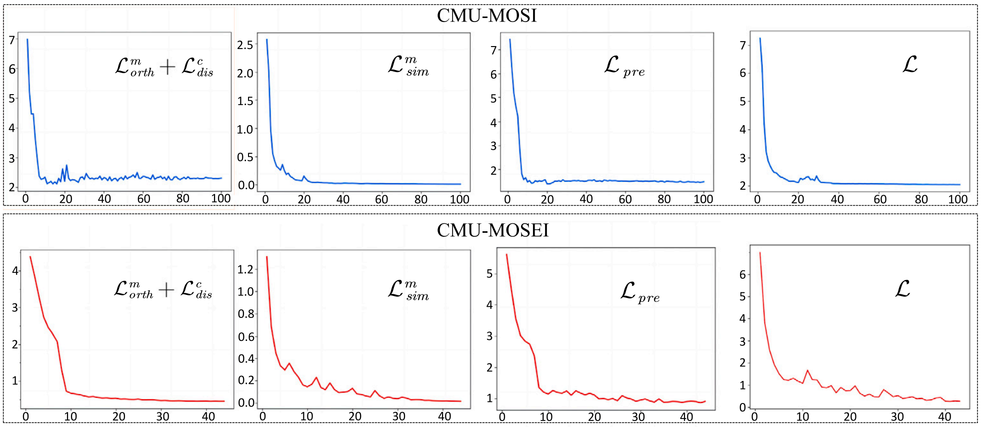
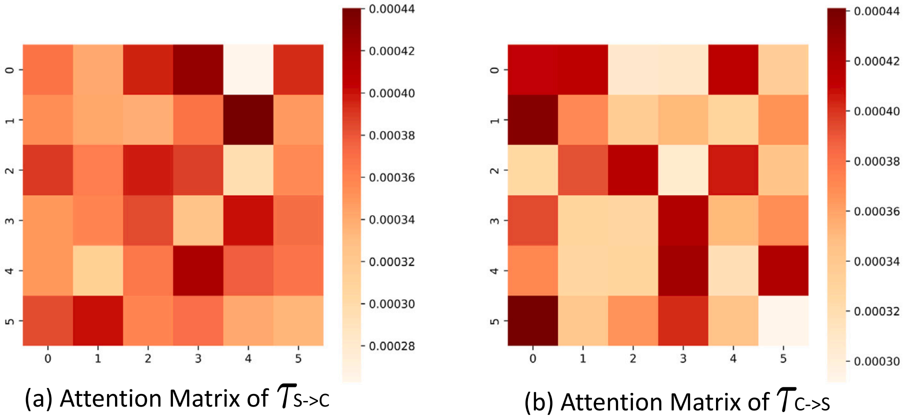


Fig. 8. The learning process for each loss objective in TMT during training.



(a) Attention Matrix of $\mathcal{T}_{S\text{->}C}$ (b) Attention Matrix of $\mathcal{T}_{C\text{->}S}$

Fig. 9. Visualization of attention in Token mutual Transformer. Obviously, the learned attention weights in two cross-modal encoders of the Token mutual Transformer are complementary, indicating that more comprehensive and complementary multimodal emotion information is obtained. Note: darker colors indicate higher attention weights for learning.

more valuable information from another decoupled feature, indicating a stronger learned fusion feature. As evident from the figure, the attention distributions differ when the heterogeneity feature and the consistency feature are used as query terms (Q), respectively. This disparity enables them to achieve complementary learning effects, resulting in more comprehensive features and improved results.

4.6.4. *Visualization of the input unimodal features and final fused features*

As shown in Fig. 10, we use t-sne to visualize the input unimodal features (*i.e.*, U_a, U_v and U_t) and the final fused features Y_0, respectively, on the 5-class emotion recognition task. The visualization results indicate that, in contrast to unimodal features with overlapping distributions for emotion classification, the final fused features demonstrate more distinct clustering, particularly for positive and negative emotions. This demonstrates that TMT has the potential to improve classification accuracy by integrating diverse and complementary information from each modality.

5. Conclusion

In this paper, we propose an effective multi-modality emotion recognition method called **T**oken-disentangling **M**utual **T**ransformer (TMT). Our method effectively disentangles and interacts with emotion-related inter-modality emotion consistency features and intra-modality emotion heterogeneity features, facilitating robust multimodal emotion recognition. TMT consists of two main parts: multimodal emotion Token disentanglement and Token mutual Transformer. To achieve multimodal emotion Token disentanglement, we introduce a novel Token separation encoder with its corresponding emotion disentanglement losses into the Transformer framework. This approach effectively separates emotion-related inter-modality emotion consistency features and intra-modality heterogeneity features from multimodal features. With the disentangled features, the Token mutual Transformer module employs two bidirectional cross-modality Transformers to perform bidirectional query interaction and fusion, resulting in a more comprehensive multimodal emotion representation. Extensive experiments were conducted on three challenging multimodal emotion datasets (CMU-MOSI, CMU-MOSEI, CH-SIMS) to evaluate the performance of our method. The results demonstrate that our method outperforms existing multimodal emotion recognition methods, achieving state-of-the-art performance. Despite the effectiveness of our method, we find that our method does not capture the problem of the 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.

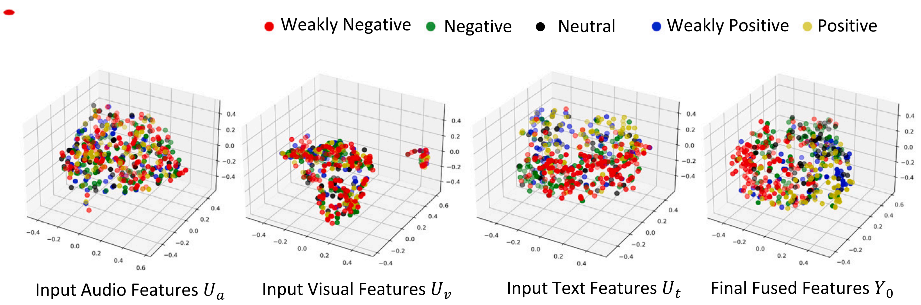


Fig. 10. Visualization of the input unimodal features (U_a, U_v and U_t) and the final fused features (Y_0) on CH-SIMS dataset, respectively.

CRediT authorship contribution statement

Guanghao Yin: Conceptualization, Methodology, Project administration, Validation, Writing – review & editing. **Yuanyuan Liu:** Methodology, Writing – review & editing. **Tengfei Liu:** Data curation, Methodology, Software, Visualization, Writing – original draft. **Haoyu Zhang:** Data curation, Writing – review & editing. **Fang Fang:** Writing – review & editing. **Chang Tang:** Writing – review & editing. **Liangxiao Jiang:** Writing – review & editing.

Declaration of competing interest

We wish to confirm that there are no known conflicts of interest associated with this manuscript and there has been no significant financial support for this work that could have influenced its outcome.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the National Natural Science Foundation of China grant (62076227), Natural Science Foundation of Hubei Province grant (2023AFB572) and Hubei Key Laboratory of Intelligent Geo-Information Processing (KLIGIP-2022-B10).

References

An, J., Wan Zainon, W.M.N., 2023. Integrating color cues to improve multimodal sentiment analysis in social media. Eng. Appl. Artif. Intell. 126, 106874. http://dx.doi.org/10.1016/j.engappai.2023.106874.

Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H., Schölkopf, B., Smola, A.J., 2006. Integrating structured biological data by kernel maximum mean discrepancy. In: Proceedings 14th International Conference on Intelligent Systems for Molecular Biology 2006, Fortaleza, Brazil, August 6-10, 2006. pp. 49–57. http://dx.doi.org/10.1093/bioinformatics/btl242.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers. In: Proceedings of the 16th European Conference on Computer Vision, vol. 12346, pp. 213–229.

Chen, Y., Joo, J., 2021. Understanding and mitigating annotation bias in facial expression recognition. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, pp. 14960–14971. http://dx.doi.org/10.1109/ICCV48922.2021.01471.

Chen, W., Xing, X., Xu, X., Pang, J., Du, L., 2022. SpeechFormer: A hierarchical efficient framework incorporating the characteristics of speech. In: Proceedings of the 23rd Annual Conference of the International Speech Communication Association. pp. 346–350.

Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555, arXiv:1412.3555.

Delbrouck, J.-B., Tits, N., Brousmiche, M., Dupont, S., 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. In: Second Grand-Challenge and Workshop on Multimodal Language. Challenge-HML, pp. 1–7.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations. ICLR 2021, Virtual Event, Austria, May 3-7, 2021.

Greff, K., Srivastava, R.K., Koutník, J., Steunebrink, B.R., Schmidhuber, J., 2017. LSTM: a search space odyssey. IEEE Trans. Neural Netw. Learn. Syst. 28 (10), 2222–2232. http://dx.doi.org/10.1109/TNNLS.2016.2582924.

Han, W., Chen, H., Gelbukh, A.F., Zadeh, A., Morency, L., Poria, S., 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In: ICMI '21: International Conference on Multimodal Interaction, MontrÉAl, QC, Canada, October 18-22, 2021. pp. 6–15.

Han, W., Chen, H., Poria, S., 2021b. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. arXiv preprint arXiv:2109.00412.

Hazarika, D., Zimmermann, R., Poria, S., 2020a. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In: Proceedings of the 28th ACM International Conference on Multimedia. ACM, pp. 1122–1131.

Hazarika, D., Zimmermann, R., Poria, S., 2020b. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In: MM '20: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1122–1131.

He, J., Su, B., Sheng, Z., Zhang, C., Yang, H., 2023. Adversarial invariant-specific representations fusion network for multimodal sentiment analysis. In: International Conference on Image, Signal Processing, and Pattern Recognition. ISPP 2023, vol. 12707, SPIE, pp. 930–942.

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.

Liang, T., Lin, G., Feng, L., Zhang, Y., Lv, F., 2021. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8148–8156.

Liang, P.P., Liu, Z., Zadeh, A., Morency, L., 2018. Multimodal language analysis with recurrent multistage fusion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. pp. 150–161.

Lin, K., Wang, L., Liu, Z., 2021. End-to-end human pose and mesh reconstruction with transformers. In: 2021 IEEE/CVF CONFERENCE on COMPUTER VISION and PATTERN RECOGNITION, CVPR 2021. pp. 1954–1963.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: 2021 IEEE/CVF International Conference on Computer Vision. ICCV 2021, pp. 9992–10002.

Liu, P., Qiu, X., Huang, X., 2017. Adversarial multi-task learning for text classification. In: Barzilay, R., Kan, M. (Eds.), Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers, Association for Computational Linguistics, pp. 1–10. http://dx.doi.org/10.18653/v1/P17-1001.

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.

Liu, Y., Wang, W., Feng, C., Zhang, H., Chen, Z., Zhan, Y., 2023a. Expression snippet transformer for robust video-based facial expression recognition. Pattern Recognit. 138, 109368.

Liu, Y., Wang, W., Zhan, Y., Feng, S., Liu, K., Chen, Z., 2023b. Pose-disentangled contrastive learning for self-supervised facial representation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023, IEEE, pp. 9717–9728. http://dx.doi.org/10.1109/CVPR52729.2023.00937.

Liu, Y., Zhang, H., Zhan, Y., Chen, Z., Yin, G., Wei, L., Chen, Z., 2023c. Noise-resistant multimodal transformer for emotion recognition. http://dx.doi.org/10.48550/arXiv.2305.02814, CoRR abs/2305.02814, arXiv:2305.02814.

Liu, W., Zhong, X., Zhou, Z., Jiang, K., Wang, Z., Lin, C., 2023d. Dual-recommendation disentanglement network for view fuzz in action recognition. IEEE Trans. Image Process. 32, 2719–2733. http://dx.doi.org/10.1109/TIP.2023.3273459.

Lv, F., Chen, X., Huang, Y., Duan, L., Lin, G., 2021. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In: IEEE Conference on Computer Vision and Pattern Recognition. CVPR, Computer Vision Foundation / IEEE, pp. 2554–2562.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 2579–2605.

Mao, H., Yuan, Z., Xu, H., Yu, W., Liu, Y., Gao, K., 2022. M-SENA: an integrated platform for multimodal sentiment analysis. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, pp. 204–213.

Molchanov, P., Tyree, S., Karras, T., Aila, T., Kautz, J., 2017. Pruning convolutional neural networks for resource efficient inference. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, URL: https://openreview.net/forum?id=SJGCiw5gl.

Park, G., Im, W., 2016. Image-text multi-modal representation learning by adversarial backpropagation. arXiv preprint arXiv:1612.08354.

Singh, N., Kapoor, R., 2023. Multi-modal expression detection (MED): A cutting-edge review of current trends, challenges and solutions. Eng. Appl. Artif. Intell. 125, 106661. http://dx.doi.org/10.1016/j.engappai.2023.106661.

Sun, Z., Sarma, P., Sethares, W., Liang, Y., 2020a. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 8992–8999.

Sun, Z., Sarma, P.K., Sethares, W.A., Liang, Y., 2020b. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence. pp. 8992–8999.

Sun, H., Wang, H., Liu, J., Chen, Y., Lin, L., 2022. CubeMLP: A MLP-based model for multimodal sentiment analysis and depression estimation. http://dx.doi.org/10.48550/ARXIV.2207.14087, CoRR abs/2207.14087, arXiv:2207.14087.

Tang, C., Li, Z., Wang, J., Liu, X., Zhang, W., Zhu, E., 2022. Unified one-step multi-view spectral clustering. IEEE Transactions on Knowledge and Data Engineering 35 (6), 6449–6460.

Tang, C., Zheng, X., Zhang, W., Liu, X., Zhu, X., Zhu, E., 2023. Unsupervised feature selection via multiple graph fusion and feature weight learning. Science China Information Sciences 66 (5), 152101.

Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al., 2021. Mlp-mixer: An all-mlp architecture for vision. In: Advances in neural information processing systems, vol. 34, pp. 24261–24272.

Tsai, Y.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R., 2019a. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Conference of the Association for Computational Linguistics. ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, pp. 6558–6569.

Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R., 2019b. Multimodal transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Conference of the Association for Computational Linguistics. pp. 6558–6569.

Tsai, Y.H., Liang, P.P., Zadeh, A., Morency, L., Salakhutdinov, R., 2019c. Learning factorized multimodal representations. In: 7th International Conference on Learning Representations. ICLR 2019, New Orleans, la, USA, May 6-9, 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 5998–6008.

Wang, D., Guo, X., Tian, Y., Liu, J., He, L., Luo, X., 2023. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. Pattern Recognit. 136, 109259. http://dx.doi.org/10.1016/J.PATCOG.2022.109259.

Wang, Z., Wan, Z., Wan, X., 2020. TransModality: An End2End fusion method with transformer for multimodal sentiment analysis. In: WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24. pp. 2514–2520.

Yan, W., Gu, M., Ren, J., Yue, G., Liu, Z., Xu, J., Lin, W., 2023a. Collaborative structure and feature learning for multi-view clustering. Inf. Fusion 98, 101832.

Yan, W., Zhang, Y., Lv, C., Tang, C., Yue, G., Liao, L., Lin, W., 2023b. GCFAgg: Global and cross-view feature aggregation for multi-view clustering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19863–19872.

Yang, D., Huang, S., Kuang, H., Du, Y., Zhang, L., 2022a. Disentangled representation learning for multimodal emotion recognition. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1642–1651.

Yang, D., Kuang, H., Huang, S., Zhang, L., 2022b. Learning modality-specific and -agnostic representations for asynchronous multimodal language sequences. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1708—-1717.

Yu, W., Xu, H., Meng, F., Zhu, Y., Ma, Y., Wu, J., Zou, 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.

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, (no. 12), pp. 10790–10797.

Yuan, Z., Li, W., Xu, H., Yu, W., 2021a. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In: ACM Multimedia Conference, Virtual Event.

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

Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L., 2017a. Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 1103–1114.

Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.-P., 2017b. Tensor fusion network for multimodal sentiment analysis. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pp. 1103–1114, arXiv preprint arXiv:1707.07250.

Zadeh, A., Liang, P.P., Mazumder, N., Poria, S., Cambria, E., Morency, L., 2018a. Memory fusion network for multi-view sequential learning. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. EAAI-18, New Orleans, Louisiana, USA, February 2-7, 2018, pp. 5634–5641.

Zadeh, A., Liang, P.P., Poria, S., Cambria, E., Morency, L., 2018b. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pp. 2236–2246.

Zadeh, A., Zellers, R., Pincus, E., Morency, L., 2016a. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. CoRR abs/1606.06259.

Zadeh, A., Zellers, R., Pincus, E., Morency, L., 2016b. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. IEEE Intell. Syst. 82–88.

Zeng, Y., Li, Z., Chen, Z., Ma, H., 2024. A feature-based restoration dynamic interaction network for multimodal sentiment analysis. Eng. Appl. Artif. Intell. 127, 107335. http://dx.doi.org/10.1016/j.engappai.2023.107335.

Zhang, F., Liu, K., Liu, Y., Wang, C., Zhou, W., Zhang, H., Wang, L., 2024. Multitarget domain adaptation building instance extraction of remote sensing imagery with domain-common approximation learning. IEEE Transactions on Geoscience and Remote Sensing 1. http://dx.doi.org/10.1109/TGRS.2024.3376719.

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: Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. EMNLP 2023, Singapore, December 6-10, 2023, Association for Computational Linguistics, pp. 756–767, URL: https://aclanthology.org/2023.emnlp-main.49.

Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., Hu, R., Chai, H., Keutzer, K., 2020. An end-to-end visual-audio attention network for emotion recognition in user-generated videos. 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, NY, USA, February 7-12, 2020, AAAI Press, pp. 303–311. http://dx.doi.org/10.1609/aaai.v34i01.5364.

Zhong, X., Lu, T., Huang, W., Ye, M., Jia, X., Lin, C., 2022. Grayscale enhancement colorization network for visible-infrared person re-identification. IEEE Trans. Circuits Syst. Video Technol. 32 (3), 1418–1430. http://dx.doi.org/10.1109/TCSVT.2021.3072171.