



TF-MERC: Integrating Time-Frequency Information for Multimodal Emotion Recognition in Conversation

Jiawei Cheng
chengjiawei248019@gmail.com
Chongqing University of Technology
Chongqing, China

Xiaofei Zhu*
zxf@cqu.edu.cn
Chongqing University of Technology
Chongqing, China

Zhou Yang
200310007@fzu.edu.cn
Fuzhou University
Fuzhou, China

Abstract

Multimodal emotion recognition in conversations aims to accurately detect emotions by integrating audio, text, and video modalities, playing an important role in various systems. Existing approaches utilize convolutional and recurrent networks to learn short-term emotional information from individual modalities, or employ graph and attention mechanisms to integrate long-term emotional information from multiple modalities. These methods effectively combine emotional information within the conversational content in the time domain. However, psychological research shows that emotional information are not only conveyed in the time domain but also in the frequency domain (e.g., pitch and speech rate). To capture emotions from a more comprehensive perspective, we propose TF-MERC, a framework that integrates both time and frequency domains. TF-MERC uses a multi-domain alignment module to learn modality information within the time or frequency domains. It then employs FATransformer to deeply integrate the multimodal associations between the time and frequency domains, providing a more comprehensive approach for emotion prediction. Experimental results show that TF-MERC outperforms state-of-the-art methods, achieving superior performance across multiple datasets.

CCS Concepts

• **Information systems** → **Sentiment analysis**; • **Computing methodologies** → **Discourse, dialogue and pragmatics**.

Keywords

Fourier transform, Multimodal fusion, Emotion recognition

ACM Reference Format:

Jiawei Cheng, Xiaofei Zhu, and Zhou Yang. 2025. TF-MERC: Integrating Time-Frequency Information for Multimodal Emotion Recognition in Conversation. In *Proceedings of the 48th International ACM ICMR Conference on Research and Development in Information Retrieval (ICMR '25)*, June 30–July 3, 2025, Chicago, IL, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3731715.3733447>

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
ICMR '25, Chicago, IL, USA.

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1592-1/25/07
<https://doi.org/10.1145/3731715.3733447>

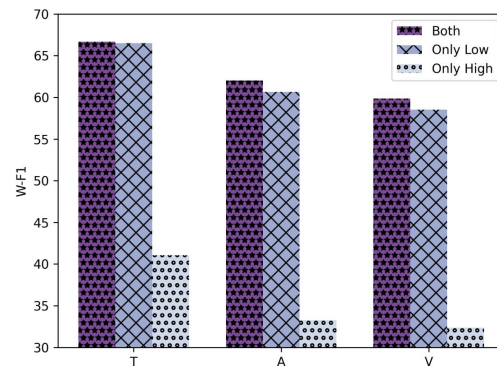


Figure 1: Emotion accuracy from different frequency perspectives, where T, A, and V represent text, audio, and visual modalities, respectively.

1 Introduction

As a key area in affective computing, multimodal emotion recognition in conversations aims to accurately detect emotions by integrating audio, text, and visual modalities. It has been widely applied in dialogue systems [31], social media analysis [5], and health support [28], yielding significant benefits.

Early studies [19, 33] use recurrent and convolutional networks to capture dialogue emotions in the text modality. These studies primarily focus on short-term dependencies within the text modality, overlooking the long-term dependencies across multiple modalities.

Recent studies introduce audio and visual modalities, using graph-based [32] and attention-based [30] methods to integrate long-term dependencies in multimodal information. Graph-based methods construct multimodal information as heterogeneous graphs and design cross-modal feature complementarity modules [6, 7], two-stage multi-source information fusion [13], and curriculum learning [22] to integrate long-term information between modalities. Attention-based methods modify the attention structure to capture cross-modal and intra-modal information [16, 35], facilitating the alignment and fusion of long-term modality information. These methods, focusing on the time domain, integrate both short-term and long-term emotional information in conversations, achieving promising results.

However, according to dialogue psychology research [10, 17], emotions are not only carried in the time domain but also exist in the frequency domain [29, 34]. For example, when expressing “happy,” conversations often include high-frequency speech and faster speech rates, with noticeable speech fluctuations. In contrast, expressing “sadness” typically involves lower pitch and slower

speech, with smoother fluctuations. To further verify the impact of frequency information, we conduct pilot experiments on the MELD [24] dataset. Specifically, we use fast fourier transform functions to extract high-frequency and low-frequency information from text, audio, and visual modalities. We then restrict the dialogue content to retain full frequency, high, and low-frequency information to assess the impact of frequency information. As shown in Figure 1, results indicate that omitting frequency information, especially low-frequency data, significantly reduces the emotion recognition accuracy. This highlights the critical role of frequency information in multimodal emotion recognition. Since time and frequency domain information carry emotional content from different perspectives, considering both dimensions comprehensively is beneficial for emotion recognition. However, how to capture the deep associations between different domain information and effectively integrate the emotional content remains a challenge.

In this paper, we propose TF-MERC, a novel framework for multimodal emotion recognition in conversations, which comprehensively considers both time and frequency perspectives. TF-MERC employs a multi-domain alignment module to capture the internal associations within time or frequency perspectives, facilitating internal information learning. It then introduces the FATransformer, which effectively integrates information between the two perspectives using the time-domain attention mechanism and frequency-domain attention mechanism. The time-domain attention captures emotion-related information that changes over time, while the frequency-domain attention focuses on the impact of frequency fluctuations. By combining these two attention mechanisms, TF-MERC further integrates the deep associations of emotion across both perspectives, leading to more accurate emotion prediction. Experimental results on widely used datasets demonstrate that TF-MERC outperforms existing state-of-the-art methods, achieving superior performance.

In summary, our contributions are as follows:

- We introduce the frequency perspective to more comprehensively capture emotions in conversations, avoiding the incompleteness of previous methods that rely solely on the time perspective.
- We propose TF-MERC, which employs multi-domain alignment modules and FATransformer to integrate deep associations within and across perspectives, promoting accurate emotion recognition.
- Experimental results show that TF-MERC achieves state-of-the-art performance on multiple public datasets.

2 Related Work

2.1 Multimodal Emotion Recognition

In recent years, MERC has become a key focus within the affective computing community, attracting considerable attention. The incorporation of multimodal data offers a multidimensional viewpoint, facilitating a more detailed comprehension of emotions. Therefore, researchers have increasingly adopted multimodal fusion techniques, merging text, audio, and visual emotion cues to improve the performance of MERC. This multimodal fusion methods are broadly categorized into two types: graph-based and attention-based.

Graph-based method: The core idea of graph-based fusion methods is to represent discourse as nodes within a graph, thereby capturing the emotional connections between discourses through graph learning techniques. For instance, the CMCF-SRNet [35] proposes a semantic refinement module that fully leverages emotional semantic relationship information from a global perspective. GA2MIF [13] employs a multi-head directed graph attention network to model intra-modal and inter-modal interactions, alleviating the accumulation of redundant information through contextual windows. Similarly, GraphCFC [14] models multi-modal dialogue as a directed graph with variable context and extracts different types of edges from the graph for graph attention learning. MultiDAG [22], on the other hand, integrates text, audio, and visual features into an undirected graph framework for learning conversational context information and employs curriculum learning to mitigate data imbalance issues.

Attention-based method: The core of this approach is to enhance the modality representations by exploiting emotional semantic information both intra-modal and inter-modal interaction information. For example, DialogueTRM [20] proposes a direct and effective intra-modal and inter-modal emotional dynamic strategy to cater to the contextual preferences of different modalities and achieve multi-granularity fusion. CTNet [16] utilize the transformer-based structure to capture intra- and inter-modal interactions among different modalities, and model temporal information in the utterance by considering both word-level lexical features and segment-level audio features. EmoCaps [15] utilizes Transformer to extract emotional vectors from each modality and fuse them into an emotional capsule. SCMM [30] proposes the self-adaptive context and modal-interaction framework, which attempts to model different ranges of context dependency as well as captures the specific contribution of each modality. MGLRA [21] creates a memory block for each modality to store single-modal emotional semantic information and employs intra-modal attention mechanism and cross-modal attention mechanism respectively to realize information extraction within modalities and feature fusion across modalities.

However, these methods primarily focus on the time information within conversation, neglecting the latent frequency information underlying the conversation. TF-MERC simultaneously considers both time and frequency information to fully exploit the emotional semantic information across different domain perspective.

2.2 Fourier Domain Learning

The Fourier transform has long been a cornerstone in digital and graphic signal processing [2, 23, 25]. Its applications span computer vision [3, 8] and natural language processing [11, 12], where it has been integrated to enhance model performance. Recent advancements have extended its use to long-term time-series forecasting and recommendation systems. For example: TFDNet [18] introduces a time-frequency enhanced encoder with trend and seasonal components for time-series data. FMLP-Rec [36] employs learnable filters to adaptively reduce noise in the frequency domain. BSARec [26] leverages frequency signals to mitigate the over-smoothing issue in Transformer-based models.

Since conversations can be viewed as a specific type of time-series data, transforming them into frequency signals offers a novel way to analyze emotional transitions. Inspired by these advancements, we attempt to introduce frequency signals in MERC to improve the performance of emotion recognition.

3 Methodology

3.1 Fourier Transform

Discrete Fourier Transform (DFT) is a popular computing method for data analysis, signal processing and machine learning. Here, 1D DFT is used for our TF-MERC. Given a finite sequence $\{x_i\}_{i=1}^n$, the 1D DFT convert the original sequence into the frequency signals in the frequency domain by:

$$X_k = \sum_{i=1}^n x_i W_n^{ik}, 1 \leq k \leq n$$

where n represents the sequence length, W_n^{ik} is the rotation factor, X_k is a complex number represents the signal with frequency $\omega_k = 2\pi k/n$. Therefore, we can decompose a series of values into different frequency components. It is noteworthy that the DFT is a one-to-one unique mapping operation between the time domain and the frequency domain. The frequency representation X_k can be transformed back into the time domain representation through the Inverse Discrete Fourier Transform (IDFT), which is formulated as:

$$x_i = \frac{1}{n} \sum_{k=1}^n X_k W_n^{-ik}.$$

For a real input x_i , it has been proven that its DFT is conjugate symmetric, i.e., $X_k = X_{n-k}^*$, where $*$ denotes the conjugate operation. This indicates that half of the DFT contains the complete frequency characteristics. If we perform Inverse Discrete Fourier Transform (IDFT) on $\{X_k\}_{k=1}^{\lceil n/2 \rceil}$, a real signal can be recovered. The Fast Fourier Transform (FFT) algorithm is a fast algorithm for computing the DFT, reducing the complexity to compute DFT from $O(N^2)$ to $O(N \log N)$. The Inverse FFT (IFFT), which has a similar form to the DFT. In this paper, we denote FFT and IFFT by \mathcal{F} and \mathcal{F}^{-1} , respectively

3.2 Problem Definition

In MERC, a conversation is defined as a sequence of utterances $U = \{u_1, u_2, \dots, u_n\}$ uttered by m speakers, where n is the number of utterances. The i -th utterance representation \mathbf{u}_i is represented by three different modalities denoted as $\mathbf{u}_i = \{\mathbf{u}_i^t, \mathbf{u}_i^a, \mathbf{u}_i^v\}$, where $\mathbf{u}_i^t \in \mathbb{R}^{d_t}$, $\mathbf{u}_i^a \in \mathbb{R}^{d_a}$, $\mathbf{u}_i^v \in \mathbb{R}^{d_v}$ are the corresponding representations of text, audio, visual modality, respectively. d_t , d_a and d_v are the dimensions of the three modalities. The goal of the task is to predict the emotion label y_i for a given utterance u_i based on the multimodal information of utterances in the conversation.

The proposed TF-MERC is shown in Figure 2, which consists of three main modules: the multi-domain alignment module, the information aggregation module and the emotion classification module.

3.3 Modality Encoding

For fair comparison, similar to [32], we use RoBERTa to obtain text features. For extracting audio features and visual features, we utilize OpenSmile, an audio feature extraction toolkit and a pre-trained DenseNet model. To handle the inconsistent dimensions in multimodal features, we use 1D convolutional operation to map each modal feature \mathbf{U}_m feature into fixed-size representation \mathbf{H}_m by:

$$\mathbf{H}_m = \text{Conv1D}(\mathbf{U}_m, k_m). \quad (1)$$

where $\mathbf{H}_m \in \mathbb{R}^{n \times d}$, $\mathbf{h}_i^m \in \mathbf{H}_m$, $m \in \{t, a, v\}$ and $\mathbf{U}_m \in \mathbb{R}^{n \times d_m}$ represents the raw feature of modality m , which is extracted by the corresponding feature extractor. k_m means kernel size.

3.4 Multi-Domain Alignment

To capture the internal associations and eliminate the gaps between modalities, we introduce a multi-domain alignment module. This module primarily consists of frequency domain alignment and time domain alignment.

The purpose of frequency domain alignment is to maintain semantic consistency in the frequency domain by bringing the frequency representations that are the same across modalities closer together and pushing those that are different further apart. We first utilize the Fast Fourier Transform (FFT) to convert the temporal representations into frequency representations. The frequency representation for each modality can be obtained as follows:

$$\tilde{\mathbf{H}}_m = \mathcal{F}(\mathbf{H}_m), \quad (2)$$

where $\tilde{\mathbf{H}}_m \in \mathbb{R}^{f \times d}$, $f = \lceil n/2 \rceil$ and \mathcal{F} indicates FFT. Frequency domain alignment is defined as follows:

$$\mathcal{L}_{Fa}^{m1, m2} = \sum_{i=1}^f \log \frac{\exp(\text{sim}(\tilde{\mathbf{h}}_{m1}^i, \tilde{\mathbf{h}}_{m2}^i)/\tau)}{\sum_{j=1}^f \exp(\text{sim}(\tilde{\mathbf{h}}_{m1}^i, \tilde{\mathbf{h}}_{m2}^j)/\tau)}, \quad (3)$$

$$\mathcal{L}_{Fa} = \mathcal{L}_{Fa}^{t,a} + \mathcal{L}_{Fa}^{t,v} + \mathcal{L}_{Fa}^{a,v}, \quad (4)$$

where $m1, m2 \in \{t, a, v\}$, $m1 \neq m2$. τ is temperature parameter.

The purpose of time domain alignment is to maintain semantic consistency in the time domain by bringing the representations of the same utterances from different modalities closer and pushing the temporal representations of different utterances further apart. We define time domain alignment as follows:

$$\mathcal{L}_{Ta}^{m1, m2} = \sum_{i=1}^n \log \frac{\exp(\text{sim}(\mathbf{h}_{m1}^i, \mathbf{h}_{m2}^i)/\tau)}{\sum_{j=1}^n \exp(\text{sim}(\mathbf{h}_{m1}^i, \mathbf{h}_{m2}^j)/\tau)}, \quad (5)$$

$$\mathcal{L}_{Ta} = \mathcal{L}_{Ta}^{t,a} + \mathcal{L}_{Ta}^{t,v} + \mathcal{L}_{Ta}^{a,v}, \quad (6)$$

Finally, hyperparameter γ_1 is used to adjust the relative importance of the two alignment strategies in multi-domain alignment module:

$$\mathcal{L}_{Align} = (1 - \gamma_1) \mathcal{L}_{Ta} + \gamma_1 \mathcal{L}_{Fa}. \quad (7)$$

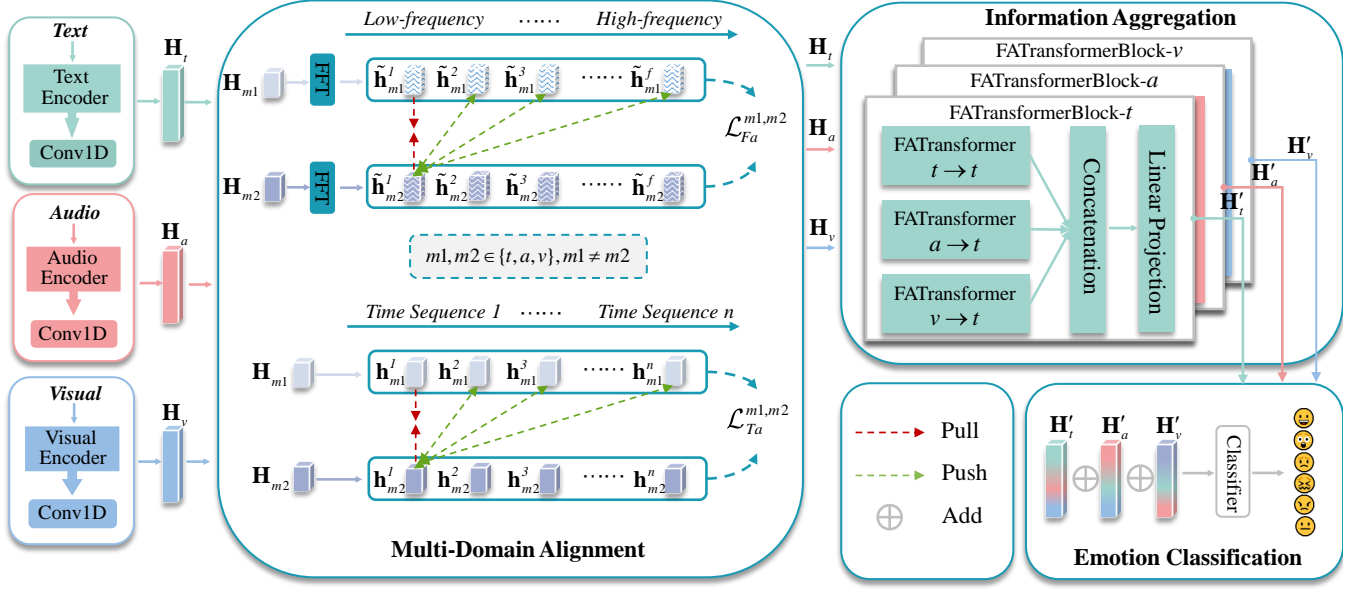


Figure 2: Overview of TF-MERC. TF-MERC processes input multimodal data by encoding raw features and using 1D convolution to unify dimension size. In the Multi-Domain Alignment phase, it aligns semantics across modalities in both time and frequency domains. During Information Aggregation, the FATransformer extracts cross-modal interactions from these domains, while modality-specific FATransformerBlocks aggregate this information to enhance each modality. Finally, the enhanced representations are combined into a multimodal representation, which is fed into a classifier for prediction.

3.5 Information Aggregation

To capture cross-modal interaction information from time and frequency domain, we propose the FATransformer (Fourier Augmented Transformer) shown in Figure 3, which integrates both time-domain and frequency-domain attention mechanisms. The time-domain attention captures captures emotion-related information that changes over time, while the frequency-domain attention focuses on the impact of frequency fluctuations.

Time-Domain Attention: Given two modality representations \mathbf{H}_{m_1} and \mathbf{H}_{m_2} where $m_1, m_2 \in \{t, a, v\}$ (different from multi-domain alignment module, here m_1 can equal to m_2), we apply a set of linear layers to project \mathbf{H}_{m_1} into $Q_{m_1}^{time} \in \mathbb{R}^{n \times d}$ and \mathbf{H}_{m_2} into $K_{m_1}^{time} \in \mathbb{R}^{n \times d}$, $V_{m_1}^{time} \in \mathbb{R}^{n \times d}$. Then, $\mathbf{H}_{m_2 \rightarrow m_1}^{time}$ is obtained by:

$$\mathbf{H}_{m_2 \rightarrow m_1}^{time} = \text{Softmax}(Q_{m_1}^{time} (K_{m_2}^{time})^T) V_{m_2}^{time}, \quad (8)$$

where $\mathbf{H}_{m_2 \rightarrow m_1}^{time} \in \mathbb{R}^{n \times d}$. In this way, we can exploit the semantic information in the time domain. Moreover, we incorporate the dropout, skip connection and layer normalization operations to alleviate the gradient vanishing and unstable training problems as:

$$\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{time} = \text{LayerNorm}(\mathbf{H}_{m_1} + \text{Dropout}(\mathbf{H}_{m_2 \rightarrow m_1}^{time})), \quad (9)$$

Freq-Domain Attention: For exploiting frequency fluctuations of emotion in the frequency domain, we apply another set of linear layers to project \mathbf{H}_{m_1} and \mathbf{H}_{m_2} into $Q_{m_1}^{freq} \in \mathbb{R}^{n \times d}$, $K_{m_1}^{freq} \in \mathbb{R}^{n \times d}$ and $V_{m_1}^{freq} \in \mathbb{R}^{n \times d}$. After that, we apply FFT to them and obtain $\tilde{Q}_{m_1}^{freq} \in \mathbb{R}^{f \times d}$, $\tilde{K}_{m_1}^{freq} \in \mathbb{R}^{f \times d}$ and $\tilde{V}_{m_1}^{freq} \in \mathbb{R}^{f \times d}$. $\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq}$ is

held by executing the self-correlation operation as follows:

$$\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq} = (\tilde{Q}_{m_1}^{freq} \odot (\tilde{K}_{m_2}^{freq})^*) \odot \tilde{V}_{m_2}^{freq}, \quad (10)$$

where $\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq} \in \mathbb{R}^{f \times d}$. \odot represents element-wise product and $*$ means the conjugate operation. Because the attention weights are derived solely from the frequency representations of the same utterance, we refer to this operation as self-correlation. We then use the inverse FFT operation to convert $\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq}$ back into the time domain. The process is defined as follows:

$$\mathbf{H}_{m_2 \rightarrow m_1}^{freq} = \mathcal{F}^{-1}(\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq}), \quad (11)$$

where $\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq} \in \mathbb{R}^{n \times d}$. The dropout, skip connection, and layer normalization operations are applied as follows:

$$\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq} = \text{LayerNorm}(\mathbf{H}_{m_1} + \text{Dropout}(\mathbf{H}_{m_2 \rightarrow m_1}^{freq})), \quad (12)$$

where $\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq} \in \mathbb{R}^{n \times d}$. The hyperparameter γ_2 is used to combine $\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{time}$ and $\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq}$ as:

$$\tilde{\mathbf{H}}_{m_2 \rightarrow m_1} = (1 - \gamma_2)\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{time} + \gamma_2\tilde{\mathbf{H}}_{m_2 \rightarrow m_1}^{freq}, \quad (13)$$

MLP and ReLU activation function are applied to further capture the non-linearity characteristics. The computation is defined as:

$$\hat{\mathbf{H}}_{m_2 \rightarrow m_1} = (\text{ReLU}(\tilde{\mathbf{H}}_{m_2 \rightarrow m_1} W^1 + b^1)) W^2 + b^2, \quad (14)$$

where $W^1, W^2 \in \mathbb{R}^{d \times d}$ and $b^1, b^2 \in \mathbb{R}^d$. Then, we also perform dropout, skip connection and layer normalization operations as

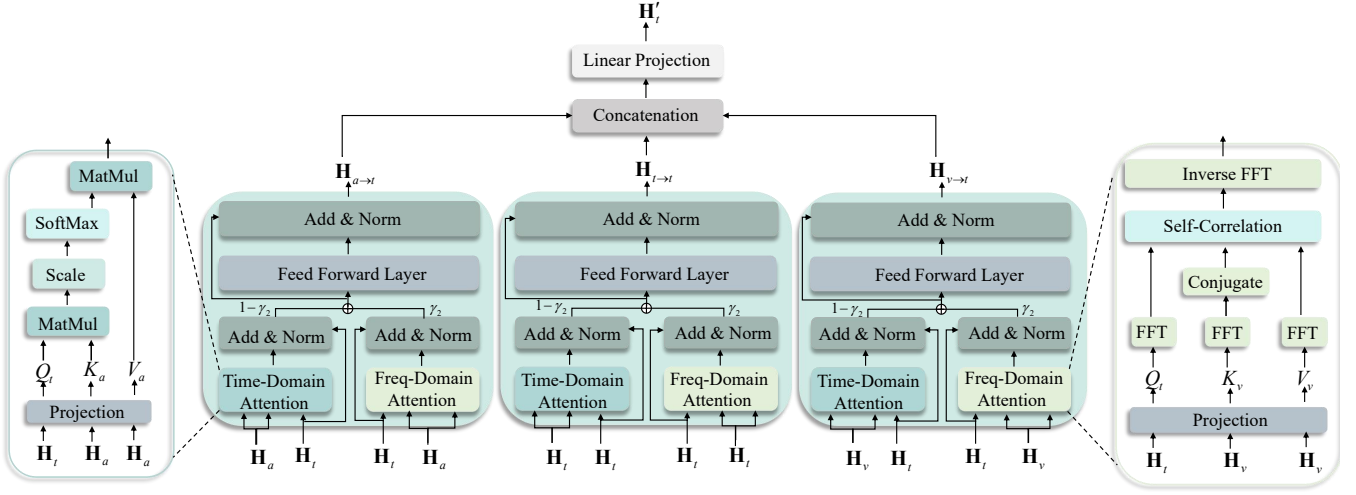


Figure 3: Illustration of FATransformerBlock-t. It consists of three FATransformers, i.e., FATransformer_{a→t}, FATransformer_{t→t} and FATransformer_{v→t}. Other modality-specific FATransformerBlocks are the same structure with FATransformerBlock-t.

in Eq.9 to generate the output $\mathbf{H}_{m_2 \rightarrow m_1}$. In the end, we define FATransformer through Equations (8-14) as follows:

$$\mathbf{H}_{m_2 \rightarrow m_1} = \text{FATransformer}_{m_2 \rightarrow m_1}(\mathbf{H}_{m_1}, \mathbf{H}_{m_2}, \mathbf{H}_{m_2}), \quad (15)$$

Based on the FATransformer, we design modality-specific FATransformerBlocks shown in Figure 3. Each block selects one modality as the central modality and aggregates interaction information from other modalities to enhance its representation. We create three blocks for text, visual, and audio modalities. Since their structures are identical, we illustrate the text modality block as an example. When text modality is the central modality, the block performs unidirectional aggregation (text-text, audio-text, and visual-text) via the FATransformer. The operations are defined as follows:

$$\mathbf{H}_{t \rightarrow t} = \text{FATransformer}_{t \rightarrow t}(\mathbf{H}_t, \mathbf{H}_t, \mathbf{H}_t), \quad (16)$$

$$\mathbf{H}_{a \rightarrow t} = \text{FATransformer}_{a \rightarrow t}(\mathbf{H}_t, \mathbf{H}_a, \mathbf{H}_a), \quad (17)$$

$$\mathbf{H}_{v \rightarrow t} = \text{FATransformer}_{v \rightarrow t}(\mathbf{H}_t, \mathbf{H}_v, \mathbf{H}_v), \quad (18)$$

where $\mathbf{H}_{t \rightarrow t}, \mathbf{H}_{t \rightarrow t}, \mathbf{H}_{t \rightarrow t} \in \mathbb{R}^{n \times d}$. Concatenation operation and fully connected layer are employed to obtain the final text modality representation as:

$$\mathbf{H}'_t = [\mathbf{H}_{t \rightarrow t} || \mathbf{H}_{a \rightarrow t} || \mathbf{H}_{v \rightarrow t}]W^3 + b^3, \quad (19)$$

where $||$ indicates concatenation operation. $\mathbf{H}'_t \in \mathbb{R}^{n \times d}$ and $W^3 \in \mathbb{R}^{3d \times d}, b^3 \in \mathbb{R}^d$. In this way, text modality is augmented by integrating the time domain and frequency domain semantic information from other modalities. Meanwhile, we also define FATransformerBlock through Equations (16-19) as follows:

$$\mathbf{H}'_t = \text{FATransformerBlock} - t(\mathbf{H}_t, \mathbf{H}_a, \mathbf{H}_v). \quad (20)$$

where $\mathbf{H}'_t \in \mathbb{R}^{n \times d}$. Additionally, we can consider audio modality and visual modality as the centers for information aggregation, thereby obtaining $\mathbf{H}'_a \in \mathbb{R}^{n \times d}$ and $\mathbf{H}'_v \in \mathbb{R}^{n \times d}$ as in Equation 20.

3.6 Emotion Classification

In the final TF-MERC, $\mathbf{H}'_t, \mathbf{H}'_a$ and \mathbf{H}'_v are summed up to derive the final joint multimodal representation, which is fed into the classifier to obtain the final emotion prediction. The process is defined as follows:

$$\hat{Y} = \text{Classifier}(\mathbf{H}'_t + \mathbf{H}'_a + \mathbf{H}'_v), \quad (21)$$

where $\hat{Y} \in \mathbb{R}^{n \times C}$, C represents the number of emotion categories. The classifier is composed of MLP layer, ReLU activation function and Softmax layer. We choose the categorical cross-entropy loss function as the task loss. The loss is shown below:

$$\mathcal{L}_{task} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^C y_{i,j} \log(\hat{Y}_{i,j}), \quad (22)$$

where $y_{i,j}$ represents the true label of utterance u_i . Finally, overall loss is defined as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{task} + \mathcal{L}_{Align}. \quad (23)$$

4 Experiments

4.1 Datasets and Evaluations

We conduct experiments on two well-known benchmark datasets, i.e., IEMOCAP [1] and MELD [24].

IEMOCAP: This dataset comprises two-way conversations involving ten speakers, with a total of 151 conversations and 7,433 utterances. It is segmented into five sessions, with the first four utilized for training and the final session reserved for testing. Each utterance within the dataset is annotated with one of six emotions, including *Happy, Sad, Neutral, Angry, Excited, Frustrated*.

MELD: Different from IEMOCAP, MELD is a multi-speaker conversation dataset with three or more speakers in a conversation. It is collected from the TV series *Friends*, which includes 1,433 conversations and 13,708 utterances. Each utterance is categorized under

Model	IEMOCAP							ACC	W-F1
	Happy	Sad	Neutral	Angry	Excited	Frustrated			
DialogueRNN	32.20	80.26	57.89	62.82	73.87	59.76	-	62.89	
DialogueGCN	51.57	80.48	57.69	53.95	72.81	57.33	-	62.89	
CTNet	51.30	79.90	65.80	67.20	<u>78.70</u>	58.80	-	67.50	
MMGCN	45.14	77.16	64.36	68.82	74.71	61.40	-	66.26	
MM-DFN	42.22	78.98	66.42	<u>69.77</u>	75.56	66.33	-	68.18	
GA2MIF	46.15	84.50	68.38	70.29	75.99	66.49	69.75	70.00	
AdaIGN	53.04	<u>81.47</u>	<u>71.26</u>	65.87	76.34	<u>67.79</u>	<u>70.49</u>	<u>70.74</u>	
HAUCL	<u>53.57</u>	82.04	68.61	66.44	75.60	68.23	70.30	70.27	
TF-MERC	59.94 [†]	81.43	71.30 [†]	65.33	78.95 [†]	66.61	71.45 [†]	71.50 [†]	

Table 1: Overall performance comparison on IEMOCAP dataset. The best and second best are in bold and underlined, respectively. '·' indicates original paper do not report. '†' means the improvement of TF-MERC is significant at $p < 0.05$ based on t-test. The metric for each emotion category is F1-score.

Model	MELD					ACC	W-F1
	Neutral	Surprise	Sadness	Joy	Angry		
DialogueRNN	76.97	47.69	20.41	50.92	45.52	-	57.66
DialogueGCN	75.97	46.05	19.60	51.20	40.83	-	56.36
CTNet	77.40	52.70	32.50	56.00	44.60	-	60.50
MMGCN	76.33	48.15	22.93	53.02	46.09	-	58.31
MM-DFN	77.76	50.69	22.93	54.78	47.82	-	59.46
GA2MIF	76.92	49.08	27.18	51.87	48.52	61.65	58.94
AdaIGN	<u>79.75</u>	60.53	<u>43.70</u>	64.54	<u>56.15</u>	67.62	<u>66.79</u>
HAUCL	-	-	-	-	-	<u>68.05</u>	66.72
TF-MERC	79.90 [†]	<u>58.85</u>	44.07 [†]	<u>64.46</u>	58.36 [†]	68.23 [†]	66.98 [†]

Table 2: Overall performance comparison on MELD dataset. The best and second best are in bold and underlined, respectively. '·' indicates original paper do not report. '†' means the improvement of TF-MERC is significant at $p < 0.05$ based on t-test. The metric for each emotion category is F1-score.

one of the seven emotions, i.e., *Neutral, Surprise, Fear, Sadness, Joy, Disgust, Angry*.

Evaluation Metrics: We evaluate the model’s performance based on two metrics, i.e., the weighted average accuracy (ACC) and the weighted average F1-score (W-F1). Additionally, we provide the F1-score for each emotion category to offer a comprehensive assessment of performance except for Fear and Disgust classes on MELD due to insufficient training samples for statistically significant results.

4.2 Implementation Details and Baselines

The proposed TF-MERC is implemented by Pytorch on RTX 4090Ti 24G GPU and trained for 30 epochs. The fixed-size dimension is 200 for IEMOCAP (100 for MELD). Batch size is 8 for IEMOCAP (4 for MELD). We use Adam [9] as the optimizer and set the learning rate to $5e-4$ for IEMOCAP ($4e-5$ for MELD). We set γ_1 and γ_2 to 0.4 and 0.2 for IEMOCAP, 0.3 and 0.2 for MELD. The results reported in our experiments are averages of 5 random runs on the test set.

In order to validate the performance of the proposed method TF-MERC in the MERC task, we introduce state-of-the-art methods for comparison: DialogueRNN [19], DialogueGCN [4], CTNet [16], MMGCN [7], MM-DFN [6], GA2MIF [13], AdaIGN [27], HAUCL [32].

4.3 Overall Results

We conduct comparative analyses with state-of-the-art (SOTA) models on the IEMOCAP and MELD datasets. Experimental results present in Table 1 and Table 2. Experimental results indicate that TF-MERC demonstrates competitive performance on both datasets. Specifically, on the IEMOCAP dataset, TF-MERC exhibits promising performance, significantly improving W-F1 by 0.76% compared to AdaIGN. Although some other methods achieve the highest F1-score for a particular emotion classification, for example, HAUCL achieving the best F1-score on *Sad* and *Frustrated* and CMCF-SRNet achieving the best F1-score on *Angry*. The obvious improvement on W-F1 shows that TF-MERC can identify fine-grained emotion compared with other methods. Similarly, on the MELD

Methods	IEMOCAP		MELD	
	ACC	W-F1	ACC	W-F1
w/o FDAlignment	70.28	70.37	68.13	66.85
w/o TDAlignment	70.31	70.50	68.11	66.91
w/o MDA	70.39	70.49	68.02	66.84
w/o FDAttention	69.39	69.52	68.13	66.94
w/o TDAttention	68.32	68.29	68.01	66.80
w/o FATransformerBlock	61.77	61.73	68.09	66.62
Our	71.45	71.50	68.23	66.98

Table 3: Ablation Study. W/o FDAlignment, w/o TDAlignment and w/o MDA mean remove frequency domain alignment, time domain alignment and multi-domain alignment module. W/o FDAttention and w/o TDAttention indicate remove the freq-domain attention and time-domain attention. We also replace FATransformerBlock with fully connected layer to valid its effectiveness.

dataset, TF-MERC also outperforms baselines and achieves competitive performance on W-F1. Although AdaIGN achieves the best F1-score on *Surprise* and *Joy*, our proposed TF-MERC achieves second F1-score on this emotions and best F1-score on other emotions. It also shows that TF-MERC’s robustness of recognizing emotion.

In summary, compared to MCF-SRNet, AdaIGN and HAUCL that utilize graph-structured learning to capture temporal semantic information in conversations, TF-MERC is capable of learning richer and more fine-grained semantic information from both the time and frequency domains of conversations, thereby demonstrating robustness in emotion recognition.

4.4 Ablation Study

In this subsection, we have conducted ablation studies to verify the effectiveness and necessity of each module in our proposed model, with the results shown in Table 3. From these experiments, we have made the following key findings:

- Removing the multi-domain alignment module (MDA) leads to a significant drop in our model’s performance on various metrics. This observation highlights the importance of dual alignment in the frequency and time domains for maintaining semantic consistency among modalities.
- In the FATransformerBlock, we note a decline in performance when either the Freq-Domain Attention (FDAttention) or Time-Domain Attention (TDAttention) is removed. The performance further deteriorates to the worst outcome when the entire FATransformer is removed. This indicates that different domain information offers the model distinct emotional cues. Thus, by considering this domain-specific information concurrently, the model can receive a richer set of emotional cues, enabling it to learn more robust modality representations.

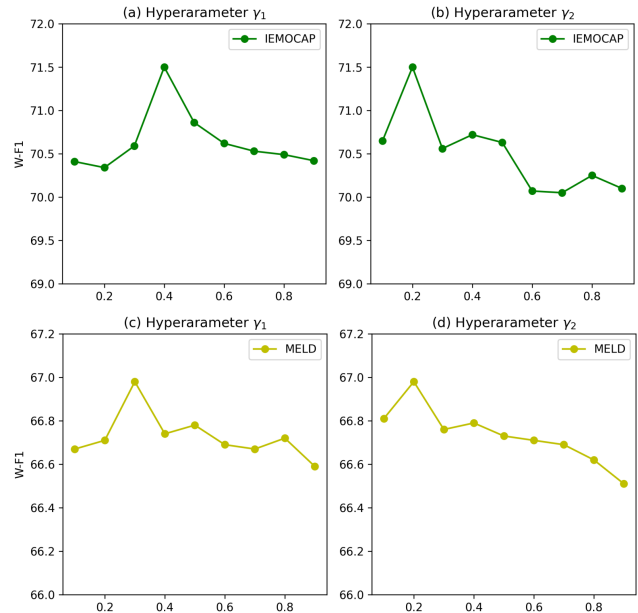


Figure 4: Hyperparameter analysis of TF-MERC on MELD and IEMOCAP dataset. All experiments test the results while fixing all other parameters with the best performance.

4.5 Hyperparameter Analysis

In this section, we have examined the impact of hyperparameters γ_1 and γ_2 on the model’s performance, where γ_1 and γ_2 are used to balance the importance of the time domain and frequency domain in the multi-domain alignment module and the FATransformer, respectively. To ensure controlled experiments, we adjusted one hyperparameter at a time. The results are displayed in Figure 4. It can be observed that in the IEMOCAP dataset, performance gradually improved with the increase in the values of γ_1 and γ_2 , reaching their peaks at 0.4 and 0.2, respectively, followed by a decline and eventually stabilizing. A similar pattern occurred in the MELD dataset, where the values of γ_1 and γ_2 reached their peaks at 0.3 and 0.2, with performance decreasing when the values were either too high or too low. This indicates that each domain plays a significant role.

4.6 Impact of Frequency Information

To verify whether considering frequency domain information can provide additional useful information, we divide the test sets of the two datasets according to the length of the conversations. Figure 5 shows the experiments of TF-MERC and the baseline model on sub-testsets of different lengths. The experimental results indicate that our model outperforms the baseline model in any sub-testsets. At the same time, it can be observed that compared to other sub-test sets (such as G2 and G3), our model significantly surpasses the baseline model in G1. This situation arises because the conversations in G1 are generally shorter, producing sparser emotional cues, and relying solely on emotional cues in the time domain cannot effectively recognize emotions. However, our proposed TF-MERC incorporates both time domain and frequency domain information,

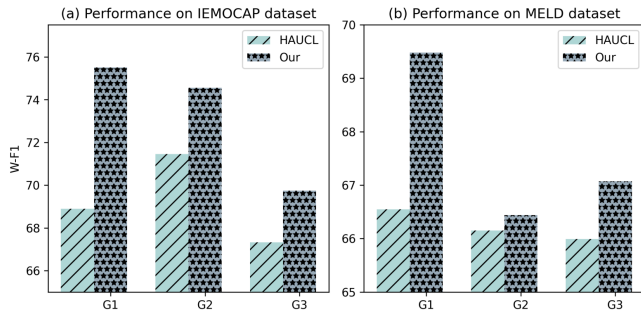


Figure 5: Experimental results on different sub-test sets. We divide the test sets of both datasets based on conversation length. In IEMOCAP, G1 represents the set of conversations with lengths less than 40, G2 represents the set of conversations with lengths in the range [40, 70), and G3 represents the set of conversations with lengths greater than or equal to 70. In MELD, G1 represents the set of conversations with lengths less than 10, G2 represents the set of conversations with lengths in the range [10, 20), and G3 represents the set of conversations with lengths greater than or equal to 20.

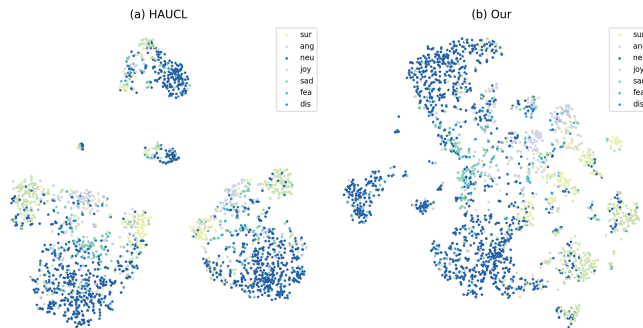


Figure 6: Visualization of HAUCL and our proposed TF-MERC on MELD dataset.

compensating for the lack of emotional cues in short conversations. This also proves the effectiveness of frequency information.

4.7 Visualization

In this section, we present in Figure 6 the multimodal features learned by our proposed model and HAUCL on the MELD dataset, to demonstrate the discriminability of the multimodal representations learned by our model. To visualize these representations, we employ the t-SNE method to transform the high-dimensional multimodal representations into a two-dimensional form. Additionally, we assign different colors to each emotional category for better observation. From Figure 6, it can be observed that in the multimodal representations learned by HAUCL, samples of the same emotion are more dispersed, and the boundaries between different emotional categories are rather vague. In contrast, the multimodal representations of the same emotional samples learned by our proposed

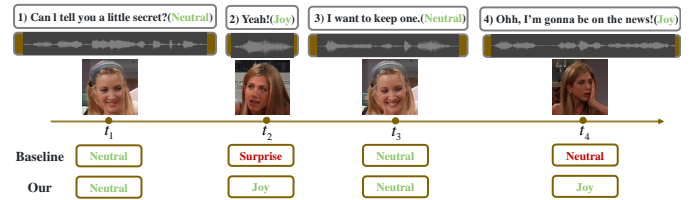


Figure 7: Case study. This conversation unfolds with temporal sequence, consists of four utterances, and the green-colored text represents the true emotion of each utterance.

model are more compact, and there is a clear separation between categories.

4.8 Case Study

In this section, we conduct case study to further validate the effectiveness of our model in short conversations, as demonstrated by the conversation example shown in Fig. 7, which consists of only four utterances. It is evident that the sequential contextual semantic information generated in such a short conversation is very sparse. Therefore, graph-based methods like HAUCL, which rely on learning emotional dependencies between utterances, may be interfered with by the sparsity of utterances and thus fail to identify the true emotions of the utterances. In contrast, TF-MERC not only learns sequential information but also observes subtle emotional changes from a frequency perspective, capturing more delicate emotional cues and ultimately accurately identifying the emotions conveyed by the utterances.

5 Conclusion and Future Work

In this paper, we proposed TF-MERC, a novel framework for multi-modal emotion recognition in conversations that integrates both time and frequency domain perspectives. By leveraging a multi-domain alignment module and the FATransformer, TF-MERC captured deep associations between time and frequency information, enabling more accurate emotion prediction. Experimental results on multiple benchmark datasets demonstrate that TF-MERC outperforms existing state-of-the-art methods, achieving superior performance. Future work will focus on further refining domain integration and applying it to more complex emotional recognition scenarios.

Although TF-MERC indicates that the use of frequency signals can enhance the accuracy of emotional recognition, there is significant potential to understand how frequency signals express the inertia and changes of emotions.

Acknowledgments

This work was supported by the Natural Science Foundation of Chongqing, China (CSTB2022NSCQ-MSX1672); the National Natural Science Foundation of China (62472059); the Chongqing Talent Plan Project, China (CSTC2024YCJH-BGZX0022); the Major Project of Science and Technology Research Program of Chongqing Education Commission of China (KJZD-M202201102); the Open Research Fund of Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education (CBDIS202403).

References

- [1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42 (2008), 335–359.
- [2] Mark Cheung, John Shi, Oren Wright, Lavendar Y Jiang, Xujin Liu, and José MF Moura. 2020. Graph signal processing and deep learning: Convolution, pooling, and topology. *IEEE Signal Processing Magazine* 37, 6 (2020), 139–149.
- [3] Tao Dai, Jianping Wang, Hang Guo, Jinmin Li, Jinbao Wang, and Zexuan Zhu. 2024. FreqFormer: Frequency-aware Transformer for Lightweight Image Super-resolution. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*. ijcai.org, 731–739.
- [4] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540* (2019).
- [5] Luis Romero Gomez, Tess Watt, Kehinde O. Babaagba, Christos Chrysoulas, Aydin E. Homyay, Raghuraman Rangarajan, and Xiaodong Liu. 2023. Emotion Recognition on Social Media Using Natural Language Processing (NLP) Techniques. In *Proceedings of the 2023 6th International Conference on Information Science and Systems*. 113–118.
- [6] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7037–7041.
- [7] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 5666–5675.
- [8] Kui Jiang, Junjun Jiang, Xianming Liu, Xin Xu, and Xianzheng Ma. [n. d.]. FM-RNet: Image Deraining via Frequency Mutual Revision. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*. AAAI. AAAI Press, 12892–12900.
- [9] Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [10] Devpriya Kumar and Narayanan Srinivasan. 2011. Emotion perception is mediated by spatial frequency content. *Emotion* 11, 5 (2011), 1144.
- [11] An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. 2024. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 18426–18434.
- [12] James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824* (2021).
- [13] Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023. GA2MIF: graph and attention based two-stage multi-source information fusion for conversational emotion detection. *IEEE Transactions on affective computing* 15, 1 (2023), 130–143.
- [14] Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia* 26 (2023), 77–89.
- [15] Zijiang Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion Capsule based Model for Conversational Emotion Recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 1610–1618.
- [16] Zheng Lian, Bin Liu, and Jianhua Tao. 2021. CTNet: Conversational transformer network for emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 985–1000.
- [17] Ke Liu, Jingzhao Hu, and Jun Feng. 2023. Speech Emotion Recognition Based on Low-Level Auto-Extracted Time-Frequency Features. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 1–5.
- [18] Yuxiao Luo, Ziyu Lyu, and Xingyu Huang. 2023. TFDNet: Time-Frequency Enhanced Decomposed Network for Long-term Time Series Forecasting. *arXiv preprint arXiv:2308.13386* (2023).
- [19] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence*. AAAI. AAAI Press, 6818–6825.
- [20] Yuzhao Mao, Qi Sun, Guang Liu, Xiaojie Wang, Weiguo Gao, Xuan Li, and Jianping Shen. 2020. DialogueTrm: Exploring the intra-and inter-modal emotional behaviors in the conversation. *arXiv preprint arXiv:2010.07637* (2020).
- [21] Tao Meng, Fuchen Zhang, Yuntao Shou, Hongen Shao, Wei Ai, and Keqin Li. 2024. Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).
- [22] Cam-Van Thi Nguyen, Cao-Bach Nguyen, Quang-Thuy Ha, and Duc-Trong Le. 2024. Curriculum Learning Meets Directed Acyclic Graph for Multimodal Emotion Recognition. *arXiv preprint arXiv:2402.17269* (2024).
- [23] I Pitas. 2000. Digital Image Processing Algorithms and Applications. *John Wiley & Sons Inc google schola* 2 (2000), 133–138.
- [24] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 527–536.
- [25] K Deergha Rao and Madiseti NS Swamy. 2018. *Digital signal processing: Theory and practice*. Springer.
- [26] Yehjin Shin, Jeongwhan Choi, Hyowon Wi, and Noseong Park. 2024. An attentive inductive bias for sequential recommendation beyond the self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 38. 8984–8992.
- [27] Geng Tu, Tian Xie, Bin Liang, Hongpeng Wang, and Ruifeng Xu. 2024. Adaptive Graph Learning for Multimodal Conversational Emotion Detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence*. AAAI. AAAI Press, 19089–19097.
- [28] Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Emotion Recognition in Conversation via Dynamic Personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. 5711–5722.
- [29] Shiyun Xu, Yinghan Cao, Zehua Zhang, and Mingjiang Wang. 2025. Two-stage UNet with channel and temporal-frequency attention for multi-channel speech enhancement. *Speech Commun.* 166 (2025), 103154.
- [30] Haozhe Yang, Xianqiang Gao, Jianlong Wu, Tian Gan, Ning Ding, Feijun Jiang, and Liqiang Nie. 2023. Self-adaptive context and modal-interaction modeling for multimodal emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2023*. 6267–6281.
- [31] Zhou Yang, Zhaochun Ren, Wang Yufeng, Haizhou Sun, Chao Chen, Xiaofei Zhu, and Xiangwen Liao. 2024. An Iterative Associative Memory Model for Empathetic Response Generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.).
- [32] Zijian Yi, Ziming Zhao, Zhishu Shen, and Tiejia Zhang. 2024. Multimodal Fusion via Hypergraph Autoencoder and Contrastive Learning for Emotion Recognition in Conversation. In *Proceedings of the 32nd ACM International Conference on Multimedia*. ACM, 4341–4348.
- [33] Haidong Zhang and Yekun Chai. 2021. Coin: Conversational interactive networks for emotion recognition in conversation. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*. 12–18.
- [34] Peng Zhang, Meijuan Li, Hui Zhao, Yida Chen, Fuqiang Wang, Ye Li, and Wei Zhao. 2024. Lightweight Fusion Model with Time-Frequency Features for Speech Emotion Recognition. In *27th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2024, Tianjin, China, May 8-10, 2024*. IEEE, 3017–3022.
- [35] Xiaoheng Zhang and Yang Li. 2023. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13099–13110.
- [36] Kun Zhou, Hui Yu, Wayne Xin Zhao, and Ji-Rong Wen. 2022. Filter-enhanced MLP is all you need for sequential recommendation. In *Proceedings of the ACM web conference 2022*. 2388–2399.