TDFNet: Transformer-Based Deep-Scale Fusion Network for Multimodal Emotion Recognition

Zhengdao Zhao, Yuhua Wang, Guang Shen, Yuezhu Xu, and Jiayuan Zhang

***Abstract*—As deep learning technology research continues to progress, artificial intelligence technology is gradually empowering various fields. To achieve a more natural human-computer interaction experience, how to accurately recognize emotional state of speech interactions has become a new research hotspot. Sequence modeling methods based on deep learning techniques have promoted the development of emotion recognition, but the mainstream methods still suffer from insufficient multimodal information interaction, difficulty in learning emotion-related features, and low recognition accuracy. In this article, we propose a transformer-based deep-scale fusion network (TDFNet) for multimodal emotion recognition, solving the aforementioned problems. The multimodal embedding (ME) module in TDFNet uses pretrained models to alleviate the data scarcity problem by providing a priori knowledge of multimodal information to the model with the help of a large amount of unlabeled data. Furthermore, a mutual transformer (MT) module is introduced to learn multimodal emotional commonality and speaker-related emotional features to improve contextual emotional semantic understanding. In addition, we design a novel emotion feature learning method named the deep-scale transformer (DST), which further improves emotion recognition by aligning multimodal features and learning multiscale emotion features through GRUs with shared weights. To comparatively evaluate the performance of TDFNet, experiments are conducted with the IEMOCAP corpus under three reasonable data splitting strategies. The experimental results show that TDFNet achieves 82.08% WA and 82.57% UA in RA data splitting, which leads to 1.78% WA and 1.17% UA improvements over the previous state-of-the-art method, respectively. Benefiting from the attentively aligned mutual correlations and fine-grained emotion-related features, TDFNet successfully achieves significant improvements in multimodal emotion recognition.**

***Index Terms*—Deep-scale fusion transformer, multimodal embedding, multimodal emotion recognition, mutual correlation, mutual transformer.**

I. Introduction

MULTIMODAL emotion recognition, which refers to processing multimedia resources and detecting the emotional state, is a very active research topic in affective computing and has been widely studied in recent years [1], [2]. With the rapid development of artificial intelligence (AI), it has become increasingly popular to study ways of improving the automatic human-machine application experience in the field of human-computer interaction (HCI) [3]. If people's current emotional states can be accurately grasped and relevant responses can be obtained during the interaction between AI products and people, users' experience with AI products may be necessarily improved to some extent. Thus, emotion recognition plays a vital role in HCI to ensure effective and accessible interactions with machines [4]. In consideration of its great significance in commodity recommendation, public opinion monitoring, man.machine dialog [5], emotion recognition is definitely a valuable research field.

Over the past few decades, more efforts have been made to investigate valuable and effective methods [6], [7], [8], [9], [10], [11] for improving the performance of multimodal emotion recognition. In the early stage, various probabilistic statistical models (e.g., hidden Markov models and Gaussian mixture models) [12], [13] were mainly applied by researchers to classify the emotional state of utterances. In these methods, a global statistics framework of utterances is designed by Gaussian mixture models using derived features of speech signals' natural pitch and energy contour. In recent years, due to the remarkable feature extraction and data fitting performance of deep neural networks (DNNs), various DNN architectures have been applied in multimodal emotion recognition [14], [15], [16], [17], [18]. Conventionally, these methods include two stages: 1) A preprocessing system computes low-level descriptors from multimedia resources and extracts robust and universal multimodal embeddings. 2) A multimodal fusion mechanism aggregates multimodal emotion-related features into utterance-level representations. Compared with traditional methods, DNNs are more feasible for refining emotional features from abundant data and more effective in distinguishing emotions. However, some limitations have not yet been solved, including the following: *1) Data Scarcity*. There is a common problem of data scarcity in multimedia corpora [19], [20], [21], [22]. The powerful fitting ability of a DNN is easily disturbed by noise information, and insufficient emotion-labeled data, especially with an imbalanced data distribution, adds to the effect of noise-contaminated periods. *2) Indistinguishable Emotional Features*. Speech and text are composed of multiple small-scale fractions. Achieving integrated semantic understanding requires building relationships between multi-scale units and learning the affective states that

Manuscript received 6 June 2022; revised 10 April 2023 and 25 June 2023; accepted 9 September 2023. Date of publication 18 September 2023; date of current version 20 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62072135 and in part by the High Performance Research Center of Harbin Engineering University. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Panayiotis Georgiou. *(Corresponding author: Yuhua Wang.)*

Zhengdao Zhao, Yuhua Wang, Guang Shen, and Yuezhu Xu are with the High Performance Computing Research Center, Harbin Engineering University, Harbin 150001, China (e-mail: 1099361413@hrbeu.edu.cn; wangyuhua@hrbeu.edu.cn; shenguang@hrbeu.edu.cn; xuyuezhu@hrbeu.edu.cn).

Jiayuan Zhang is with the High Performance Computing Laboratory, Harbin Engineering University, Harbin 150001, China (e-mail: sp15128933354@126.com).

Digital Object Identifier 10.1109/TASLP.2023.3316458

exist between different words, phrase, and tone collocations. Deep-scale features represent emotional features that automatically capture multimodal information at the local level by the model, while fine-grained emotional features are used to distinguish a broader range of utterance sentiment states. A practicable way to utilize deep-scale and fine-grained emotional features, which reflects the association between multi-scale acoustic units and fine-grained emotional states (e.g., happy, sad, angry, and neutral), is the key to compatibly modeling the emotional states of utterances [23], [24]. *3) Fragile Multimodal Interaction*. The embedding lengths of both speech and text are usually misaligned in an application, and a multimodal features fusion shortages occur at the segment level, restricting the extraction of mutual correlations in deep-scale aspects [25], [26], [27], [28]. At the same time, the lack of constructing correspondences between speaker features and multimodal features limits the model to learn the emotional vocal characteristics of speakers in different contexts. Learning speaker-related mutual correlations, which are used to construct associations between speaker vocal features and emotional expressions, to improve the performance of emotion recognition.

In solving the data scarcity problem, researchers have practiced methods such as data augmentation [17], semisupervised learning [29], and pretrained embedding [19]. Data augmentation methods relieve the abovementioned problems to some extent. Nevertheless, data augmentation methods (e.g., noise injection, random period masking, and segment replacement) hardly build representative features for different emotion categories [30]. Semisupervised learning methods seem to be negligibly influenced by the problems mentioned above. Models are pretrained with labeled data to acquire general emotional features and leverage a large amount of unlabeled data to screen in-domain features by learning the prior distribution. Considering the high computational cost in semisupervised learning, it is still a challenge to extend its strengths of iterating and transferring applications [31]. It is a more feasible fact that utilizing a pretrained embedding for multimodal emotion recognition has distinguished feature extraction capabilities and applications. In recent years, transformer-based [32] unsupervised learning pretrained embeddings have proven effective in natural language processing [33]. A transformer can effectively model the relationships between information and fit vast data very well by benefiting from the global self-attention computing characteristic, and it is capable of generating robust representations that are universal and imply semantic relationships. How to use pretrained embeddings to enhance emotion recognition is a valuable research direction.

To extract more prominent emotional features, current DNN methods mainly adopt four typical architectures, convolution neural networks with recurrent neural networks (CRNNs) [34], the capsule network (CapsNet) [35], graph neural network (GNN) [36], and transformer [28]. A CRNN architecture uses a convolution neural network (CNN) to match emotional patterns in spatial information, and a recurrent neural network (RNN) is used to aggregate these patterns into utterance-level representations. Different from the RNN methods that aggregate on the whole-time sequence, the emotional features extracted by the CRNN represent the specific feature sequence of emotion in utterances [37]. The CapsNet architecture is designed to construct the relationship between features extracted by a CNN. Researchers have leveraged this strength in emotion recognition to better model the related spatial information in speech. Practically, capsules can store the relationships of each feature, and RNNs can aggregate them to form discriminate representations [38]. However, experimental results in [38] indicate that CapsNet has relatively poor performance in recognizing happy emotions due to limited training data and its special emotion characteristics. Specifically, the Happy emotion relies more on context contrastive information than other categories. A GNN architecture relies on the information transmission between nodes to capture dependencies in a graph. In emotion recognition, a CNN can only be applied to conventional Euclidean data, which can be regarded as the instantiation of graphs. Researchers have leveraged GNNs to reform utterances into frames and construct a multiconnection graph to extend the emotion recognition application area [39]. Restricted by the graph construction of speech, the performance of GNN methods still has room for improvement [40]. The transformer, proposed in [32], are used to address the text translation problem with robust self-attention computation and adequate data fitting ability. In emotion recognition, the encoder from the transformer is utilized to model a multimodal embedding into a hidden emotional representation. However, the transformer computes self-attention in a global range and might be unproductive in terms of modeling local relationships. We consider that the emotional states of utterances are associated with local and fine-grained areas. A method is required to select the emotion-related segment and calculate these deep-scale emotional features.

Comparative methods mainly adopt two strategies to address mutual correlations, including frame-level fusion [27] and decision-level fusion [41]. Frame-level fusion works on speech and text frames, and each frame represents a short utterance period. When fusing frame-level mutual correlations, a particular architecture is used to calculate the frame-to-frame relevance of speech and text [25]. Emotional representation is formed in the early stage, and the performance mainly depends on the frame processing and fusion mechanism. Decision-level fusion works on the features of each modality, which denotes utterance-level representations from multiple modalities. It modifies the weight of multimodal features and emphasizes emotion-related features to organize an emotional state effectively. This strategy utilizes mutual correlations in multimodal emotion recognition, but it ignores correlations in deep-scale information [26]. How to solve the problem of different granularity of multimodal information and learning more effective mutual correlation is an important way to improve the level of multimodal information interaction.

In this article, a transformer-based deep-scale fusion network (TDFNet) for multimodal emotion recognition is proposed. It uses pretrained models to generate robust and universal embeddings for speech and text. TDFNet captures and calculates deep-scale features from attentively aligned multimodal information, supplemented with speaker-related mutual correlations extracted from the mutual transformer to learn and construct the speaker's emotional features in multimodal information to

improve recognition accuracy in complex contexts. Then, TDFNet properly aggregates these multimodal features and effectively classifies the emotional categories of utterances. This method has the following three main strengths:

- The multimodal embedding module generates robust and universal embeddings for speech and text.
- The deep-scale transformer module extracts fine-grained information in attentively aligned multimodal embeddings and aggregates emotion-related features effectively.
- The mutual transformer models and captures speaker-related mutual correlations from corresponding feature sets.

This article is organized as follows. In Section II, related methods are discussed. In Section III, the TDFNet architecture is explained in detail, and the available multimodal fusion strategy is elaborated. The experiments are reported in Section IV. We finally summarize the method in Section V.

II. Related Work

In this section, we introduce research works related to multimodal emotion recognition. They consist of three parts: unsupervised feature learning, multi-scale feature extraction, and multimodal fusion strategy.

1) Unsupervised Feature Learning: Various researches [5], [19], [20], [21], [22] have proven that a universal and robust pretrained embedding can improve the performance of classification tasks. Neumann et al. [19] investigated learning representations for large unlabeled speech corpora and proposed a CNN-based emotion classifier with an unsupervised autoencoder to integrate representations. The experimental results indicated that pretraining an autoencoder with unlabeled corpora could leverage the unsupervised prior information about emotions for classification in supervised learning. Eskimez et al. [20] noted that unsupervised feature learning techniques could alleviate data scarcity problems. They implemented four kinds of unsupervised feature learning methods and demonstrated that they could improve speaker-independent automatic speech emotion recognition. Li et al. [21] investigated the use of the contrastive predictive coding method to acquire salient representations from unlabeled datasets. Lin et al. [5] applied a novel semisupervised learning framework, DeepEmoCluster, for attribute-based SER tasks. Their approach maximized the emotional separation of K-means clusters, encouraging the model to learn latent representations and achieve competitive results. Zhang et al. [42] proposed an unsupervised pretrained method to handle the limitation of the labeled data size. This method used a transformer-based encoder to learn a general and robust high-level representation. The experimental results showed that these pretrained embeddings could significantly improve the performance in four emotion categories. Paraskevopoulos et al. [43] studied the application of linear and nonlinear dimensionality reduction algorithms to extract low-level feature representations in speech emotion recognition. Lian et al. [44] investigated combinations of unsupervised representation learning strategies, future observation prediction, and transfer learning methods (such as fine-tuning and super columns). The experimental results showed that their method was superior to the current advanced unsupervised learning strategy.

2) Multi-Scale Feature Extraction A dependent relationship in the long or short term exists between segmented information in utterances, so multiscale feature extraction methods have recently been utilized for emotion recognition. Many experiments have shown that multiscale features can benefit the model in acquiring emotion-related and fine-grained representations. Yoon et al. [14] proposed a framework using acoustic information and linguistic data and an attention mechanism named multihop. Multihop attention calculates the relevant segment of the text data corresponding to the audio signal. Liu et al. [15] declared that capsule networks could overcome the shortcomings of CNNs and capture shallow global features from spectrograms. They proposed a local-global aware deep representation learning system and improved the performance of vanilla CapsNet. Peng et al. [16] proposed a simple but effective neural network structure using acoustic and linguistic information in speech. The network uses multiscale convolutional layers to obtain hidden audio and text representations. Xu et al. [17] applied multiscale area attention to focus on emotional features with different granularities in a deep convolution neural network so that the classifier could benefit from attention sets with different scales. Chen et al. [34] proposed a multiscale fusion framework based on audio and text information named STSER. A multiscale fusion strategy, including feature fusion and ensemble learning, was adopted to improve the overall performance. Chen et al. [41] proposed a new method of simultaneous recognition of temporal and semantic consistency in the network. The experimental results showed that these temporal and semantic enhanced architectures could significantly improve the performance.

3) Multimodal Fusion Strategy: The key to integrating different modality information into one stable and distinguishing representation is the strategy of multimodal fusion. Generally, fine and closed interactions are more effective in representing mutual correlations. Huang et al. [8] used a transformer model to integrate audio-visual modes at the model level. Specifically, after encoding audio and video modes, the common semantic feature space generated multimodal emotional intermediate representation. Pepino et al. [9] studied different methods of classifying emotions from speech using acoustic and text-based features. They found that fusion acoustics and text-based systems were beneficial to both datasets, although only slight differences were observed in the fusion methods evaluated. Pryasad et al. [25] proposed a method based on deep learning that uses and integrates text and acoustic data for emotion classification. Sebastian et al. [26] studied the application of a long short-term memory (LSTM) recurrent neural network and a CNN in text-based emotion recognition. The LSTM recurrent neural network had a pretrained word embedding, and the CNN had a discourse level descriptor for speech emotion recognition. Various fusion strategies were used in these models to obtain the total score of each emotion category. Guang et al. [27] explicitly simulated the dynamic interaction between audio and text at the word level through the interaction unit between two LSTM networks representing audio and text. They proposed a new multimodal speech emotion recognition fusion framework

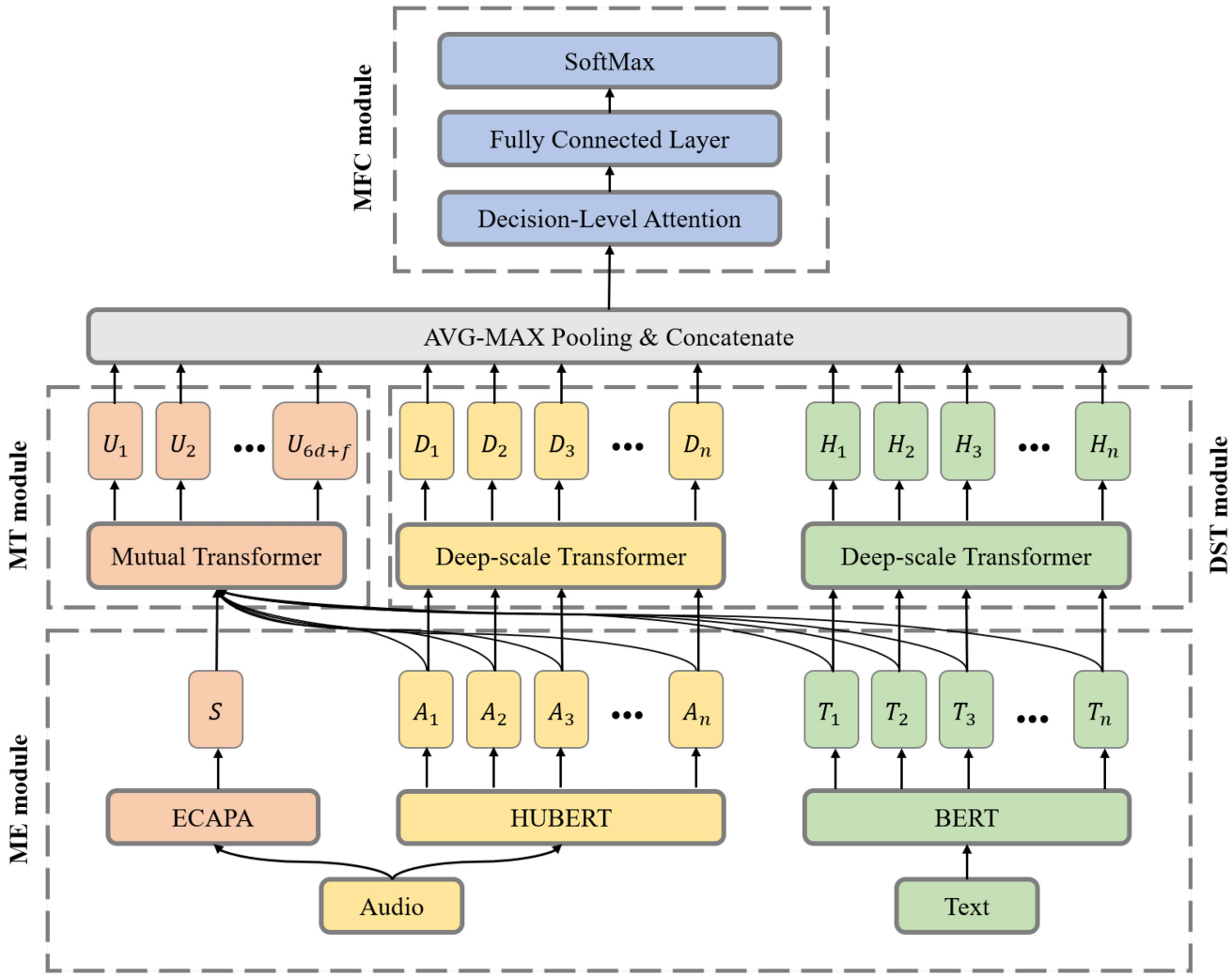

Fig. 1. Overview of the TDFNet architecture. The ME module uses three unsupervised pretrained models to generate universal embeddings A and T, and speaker features S for audio and text. The DST module attentively aligns and captures the deep-scale and fine-grained emotional features D and H from a multimodal embedding. The MT module computes the mutual correlations U from the multimodal embedding and utilizes the speaker-related information to enhance the emotion-related features. The MFC module aggregates these mutual correlations U and deep-scale features D and H into an utterance-level representation, deploying attention to salient emotional features and classifying the emotional states.

WISE-based on word-level interaction to adapt to the above components. Lian et al. [28] proposed a multimodal learning framework for conversational emotion recognition, which was referred to as a conversational transformer network. At the same time, they used word-level lexical features and fragment-level acoustic features as input to capture the time information in a discourse.

III. Methodology

In this section, TDFNet is proposed and explained in detail. The architecture of the TDFNet framework is described in Fig. 1. The TDFNet architecture consists of four components, including a multimodal embedding (ME) module, a deep-scale transformer (DST) module, a mutual transformer (MT) module and a multimodal fusion classification (MFC) module. All four elements are specifically introduced in the following subsections.

A. Overview of TDFNet Architecture

The TDFNet architecture mainly focuses on modeling the deep-scale features and speaker-related mutual correlations between audio and text, collecting and aggregating the fine-grained emotion features into an utterance-level representation to predict the emotion of certain multimedia resources. Receiving a series of audio-text pairs as inputs, the ME module utilizes unsupervised pretrained models to generate robust and universal embeddings for speech and text and supplements them with speaker features to enhance the model by acquiring speaker-related information. Then, the DST module attentively aligns the multimodal embedding and captures the deep scale information to extract the emotion-related features. The MT module uses speaker-related information and the mutual transformer to calculate the correlations between speech and text. The MFC module simultaneously fuses speaker-related mutual correlations and deep-scale features into an utterance-level representation,

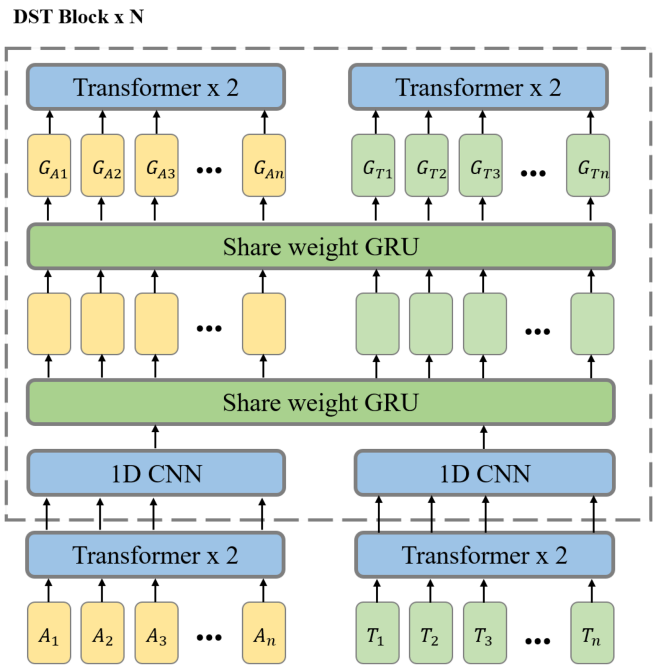

Fig. 2. Details of the DST module. The 1D CNN layer aims to downscale the sequence length of A and T. Two layers of weight-sharing GRU simultaneously aggregate A and T, applying the same weight to make the emotion-related periods salient. Aligned speech features G_A and text features G_T are fed into the transformer layer to further extract deep-scale features. N denotes the number of DST blocks applied.

attentively enhancing emotion-related features and classifying emotion states.

B. Multimodal Embedding

To benefit from the universal and robust multimodal embedding of the pretrained method, three unsupervised pretrained models are used in this architecture. In the ME module, the multimodal dataset $M = \{A_i, T_i\}_{i=1}^{n}$ is defined with n speech-text pairs, in which A and T represent the speech and text, respectively. In the audio path, HUBERT and ECAPA [45] from SpeechBrain [46] are specially utilized to extract audio embedding and speaker features. Practically, HUBERT is applied to extract audio embedding $A = \{a_1, a_2, \ldots, a_n\} \in \mathbb{R}^{n\times d}$, wherein $n = 512$ denotes the sequence length and $d = 1024$ denotes the embedding size. The ECAPA is applied to extract speaker features $S \in \mathbb{R}^f$ from the audio signal, wherein $f = 192$ denotes the embedding size. In the text path, BERT is applied to extract text embedding $T = \{t_1, t_2, \ldots, t_n\} \in \mathbb{R}^{n\times d}$, wherein the $n = 128$ denotes the number of words and $d = 768$ denotes the embedding size.

C. Deep-Scale Transformer

To capture the fine-grained emotional features from the multimodal embedding, the DST module is proposed to model the local coherence from the deep-scale information. As Fig. 2 illustrates, the DST module consists of three parts. First, two transformer layers are applied to extract features A' and T' in the global range. The multimodal features A' and T' are first downscaled with a 1D CNN layer. The sequence lengths of A' and T' are reshaped to $n = 64$ and embedding size is $d = 384$. In this experimental setting, each DST block downscales the sequence length and embedding size into $\frac{n}{2}$ and $\frac{d}{2}$, respectively. Then the weight-sharing GRU simultaneously aggregates A' and T', applying the same weight to attentively align information and make the emotion-related periods salient. The attentively aligned multimodal embedding $G_A \in \mathbb{R}^{\frac{n}{2}\times\frac{d}{2}}$ and $G_T \in \mathbb{R}^{\frac{n}{2}\times\frac{d}{2}}$ are the output of weight-sharing GRU.

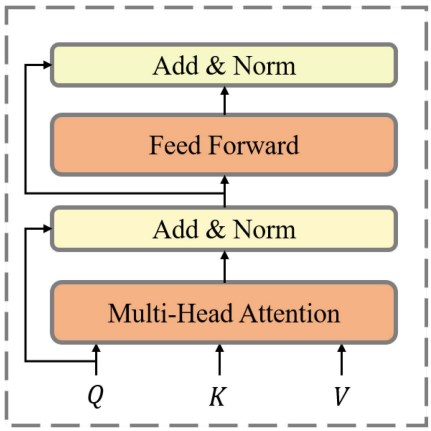

Fig. 3. Vanilla architecture of the transformer encoder. Q, K and V represent the query, key and value of inputs, respectively.

$$G_h = GRU_1(a_i), i = 1, 2, \ldots, \frac{n}{2},$$
$$G_A = GRU_2(G_{h_i}), i = 1, 2, \ldots, \frac{n}{2}. \quad (1)$$
$$G_h = GRU_1(t_i), i = 1, 2, \ldots, \frac{n}{2},$$
$$G_T = GRU_2(G_{h_i}), i = 1, 2, \ldots, \frac{n}{2}. \quad (2)$$

where GRU_i denotes the $i-$th GRU and n denotes the sequence length of A' and T'. G_h denotes the hidden representation between the consecutive GRUs. G_A and G_T denote the attentively aligned representation aggregated from A' and T', respectively.

Then, the transformer encoder calculates the attention from multimodal downscaled features. As illustrated in Fig. 3, the transformer calculates the self-attention by a multihead attention mechanism. Given feature X, it is processed into a query, a key, and a value, represented as Q, K, and V, respectively.

$$Q = W_Q X,$$
$$K = W_K X,$$
$$V = W_V X. \quad (3)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_K}}\right)V. \quad (4)$$

where W_Q, W_K and W_V denote the weight matrices for Q, K, and V respectively. $Attention(Q, K, V)$ denotes the self-attention calculated from Q, K and V respectively.

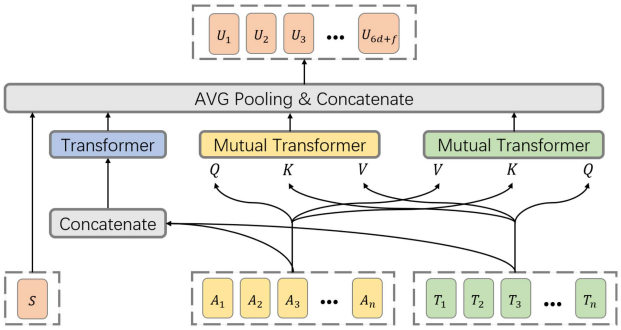

Fig. 4. Details of the MT module. The mutual transformer calculates the mutual attention from A and T and their concatenation. The mutual correlations are average pooled and concatenated into an utterance-level representation with speaker features. U is the output of the MT module.

In the DST module, the transformer layer calculates the self-attention $D \in \mathbb{R}^{\frac{n}{2} \times \frac{d}{2}}$ and $H \in \mathbb{R}^{\frac{n}{2} \times \frac{d}{2}}$ for speech and text, respectively. The formula is shown as follows.

$$D = L(G_A + \sigma(W^T(L(G_A + Attn(G_A))) + b)),$$
$$H = L(G_T + \sigma(W^T(L(G_T + Attn(G_T))) + b)), \quad (5)$$

where L denotes the LayerNorm [47] function and σ denotes the GELU [48] activation. $Attn$ is the abbreviation of the self-attention function. D and H denote the output of the DST module for A and T, respectively. W and b are learnable parameters.

It is effective to capture deep-scale information by stacking multiple DST blocks, and each block generates more local and fine-grained emotion features. We set $N = 2$ to acquire the deep-scale information in two levels. The outputs of the final DST block are $D \in \mathbb{R}^{\frac{n}{2^N} \times \frac{d}{2^N}}$ and $H \in \mathbb{R}^{\frac{n}{2^N} \times \frac{d}{2^N}}$.

D. Mutual Transformer

TDFNet computes multimodal mutual correlations and speaker-related emotion features with the MT module. The MT module is illustrated in Fig. 4. There are two mutual transformer layers to calculate the mutual correlations $C_{AT} \in \mathbb{R}^{n \times d}$ and $C_{TA} \in \mathbb{R}^{n \times d}$ from $A \in \mathbb{R}^{n \times d}$ and $T \in \mathbb{R}^{n \times d}$, respectively. As previously mentioned, the mutual transformer receives Q and K, V from different modalities. In particular, the C_{AT} mutual transformer receives the speech embedding A as Q and text embedding T as K and V. C_{TA} receives T as Q and A as K and V. Meanwhile, a particular transformer is applied to model the correlations $C \in \mathbb{R}^{2n \times d}$ of the concatenation of A and T, and the speaker features $S \in \mathbb{R}^f$ are added into utterance-level representations. Then, the formulations are as follows.

$$Attn(A, T) = softmax\left(\frac{Q_a K_t^T}{\sqrt{d_{K_t}}}\right) V_t,$$
$$Attn(T, A) = softmax\left(\frac{Q_t K_a^T}{\sqrt{d_{K_a}}}\right) V_a. \quad (6)$$
$$C_{AT} = L\left(A + \sigma\left(W^T(L(A + Attn(A, T))) + b\right)\right),$$
$$C_{TA} = L\left(T + \sigma\left(W^T(L(T + Attn(T, A))) + b\right)\right), \quad (7)$$
$$C = Transformer(Concat(A, T)). \quad (8)$$
$$U = Concat(AvgPool(S, C, C_{AT}, C_{TA})). \quad (9)$$

where Q_a, K_a, V_a denote the query, key, and value from speech embedding A, respectively, and the Q_t, K_t, and V_t from text embedding T, respectively. C, C_{AT}, and C_{TA} denote the output of mutual transformer. $AvgPool$ represents the average pooling function. $U \in \mathbb{R}^{(6d+f)}$ is the output of the MT module. W and b are learnable parameters.

E. Multimodal Fusion Classification

To make the emotion-related features salient in the utterance-level representation, the MFC module utilizes a decision-level attention mechanism to refine the pretrained embeddings $A \in \mathbb{R}^{n \times d}$ and $T \in \mathbb{R}^{n \times d}$, the deep-scale speech features $D = \{D_1, D_2, \ldots, D_i\} \in \mathbb{R}^{\frac{n}{2^i} \times \frac{d}{2^i}}, i = 1, 2, \ldots, N$, text features $H = \{H_1, H_2, \ldots, H_i\} \in \mathbb{R}^{\frac{n}{2^i} \times \frac{d}{2^i}}, i = 1, 2, \ldots, N$ and speaker-related mutual correlations $U \in \mathbb{R}^{(6d+f)}$. As shown in Fig. 1, there is an AVG-MAX pooling layer to statistically pool D, H, and U. Then, these features are concatenated as $O \in \mathbb{R}^{\left(\sum_{i=1}^{N} \frac{d}{2^{i-2}} + 12d + f\right)}$. The formulations are shown below.

$$O = Concat(Pool(A, T, A', T', U, D, H)). \quad (10)$$
$$\alpha = sigmoid(W^T O),$$
$$P = \sigma(W^T(\alpha^T O) + b). \quad (11)$$

where A and T denote the pretrained embeddings of speech and text, respectively. U denotes the mutual correlations, N denotes the number of DST blocks, D denotes the deep-scale speech features, and H denotes the deep-scale text features. $Pool$ denotes the AVG-MAX pooling layer, and O is the concatenation of these pooling features. α denotes the attention scores, and σ denotes the GELU activation function. $P \in \mathbb{R}^4$ is the emotion probability. W and b are learnable parameters.

IV. Experiments

In this section, TDFNet is evaluated with the IEMOCAP corpus, and the results are compared with those of previous state-of-the-art methods for three data splitting strategies. We elaborate on the experimental results in detail and perform ablation studies, reflecting the contributions of each module in the TDFNet method. Then, we describe a visualized presentation with the t-SNE [49] dimension reduction method and analyze the performance of this method.

A. Dataset Description

We adopt the widely used Interactive Emotional Dyadic Motion Capture (IEMOCAP) [50] corpus, which contains approximately 12 hours of multimedia data, covering video, speech, and motion capture of face and text transcriptions. IEMOCAP contains over 10 k utterances about improvisations and scripts emotion-annotated audio from 10 sample subjects (5 male and 5 female) in five sessions. This method uses the part containing speech and text transcriptions. Five emotion categories, $anger$, $happiness$, $excitement$, $sadness$, and $neutral$, are used to

compare the performance of the proposed method with those of the baseline. There is a data imbalance problem in different emotion categories, and the $excitement$ and $happiness$ parts are merged in the same way as in previous work. The final multimodal dataset contains a total of 5531 utterances (1103$anger$, 1636$happiness$, 1084$sadness$, and 1708$neutral$).

Previous studies [51], [52] have demonstrated that data splitting strategies are important and that different settings significantly influence the experimental results. To conduct fair and comparable experiments with the IEMOCAP dataset, we adopt three suitable data splitting strategies, similar to previous work: leave-one-session-out (LOSO) [19], leave-one-person-out (LOPO) [7], and random (RA) [53].

B. Experimental Setup

1) Implementation Details: We extract speech embedding $A \in \mathbb{R}^{512\times1024}$ from HUBERT, text embedding $T \in \mathbb{R}^{128\times768}$ from BERT and speaker features $S \in \mathbb{R}^{192}$ from ECAPATDNN. There are 12 heads and 4096 feedforward sizes for each transformer layer, and the transformer encoders in the DST module and MT module contain two transformer layers. In the DST module, there are 768 hidden sizes in the first weight-sharing GRU layer, 384 hidden sizes in the latter layer. It adopts a batch size of 32 and AdamW [54] optimization during training, with a learning rate of 1e-4. To reduce the impact of overfitting, we set a maximum of 20 epochs and use an early stopping strategy (stopping when the validation set loss does not drop for 3 consecutive epochs) to obtain the model with the best identification performance.

2) Data Splitting: As mentioned above, there is no predefining data splitting in the IEMOCAP corpus. Thus, three comparable methods are used for data splitting. 1) LOSO. One session is left out in IEMOCAP for evaluation, and the remaining four sessions are used for training. A 5-fold cross-validation strategy is conducted, and the mean accuracy is calculated for the results. 2) LOPO. One person is left out in IEMOCAP for evaluation, and nine persons are used for training. A 10-fold cross-validation strategy is conducted, and the mean accuracy is calculated as the results. 3) Random (RA). The dataset is randomly split into a training \ test set with a 9\1 proportion with 10 folds, and the mean accuracy is calculated as the results.

3) Evaluation Metrics: For the data imbalance problem, we adopt the unweighted accuracy (UA) and weighted accuracy (WA) [55] metrics for assessment. The UA metric denotes the average accuracy of all emotion categories. The WA metric considers the data distribution and weighs the accuracy of each emotion category by the proportion of the data.

$$WA = \frac{\sum_{i=1}^{C} N_i * Accuracy_i}{\sum_{i=1}^{C} N_i}.$$

$$UA = \frac{1}{C}\sum_{i=1}^{C} Accuracy_i. \tag{12}$$

where C denotes the number of emotion categories and i represents the i-th emotion category. N_i denotes the data quantity of the i-th emotion category.

C. Experimental Results and Comparison

The proposed method is evaluated with the IEMOCAP corpus and three data splitting strategies. The experimental results are shown in Table I. The experimental results are compared with those of related methods with LOSO, LOPO and RA splitting, as shown in Tables II, III, and IV, respectively. Fig. 5, shows the confusion matrix of the recognition accuracy with the proposed TDFNet method. Fig. 6, demonstrated t-SNE downscaling for ablation experiments. Figs. 7 and 8, show the LOSS, UA, and WA variations of the TDFNet on the training set. Moreover, Figs. 9–11 illustrates the distribution of the features from speech, text, and TDFNet with the t-SNE dimension reduction method.

In LOSO splitting, the proposed method achieves 76.26% WA and 77.04% UA with the IEMOCAP corpus. The experimental results indicate that the performance of emotion prediction in untrained conversations is challenging because the model predicts emotion categories without preacquired speaker-related features. According to the results in Table II, the proposed method exceeds the state-of-the-art methods in previous works. There are 1.34% WA and 0.4% UA improvements over the TSIN methods. TDFNet achieves the highest scores for both the WA and UA metrics.

Compared with that of unimodal methods, TDFNet, benefiting from deep-scale features and multimodal mutual correlations, significantly enhances recognition performance. There are improvements of approximately 10% over the RNN-WPA (18.24%), ACNN-AE (17.50%), TBE (15.24%), and IAAN (10.74%) methods. Compared with multimodal methods, TDFNet attentively aligns the multimodal information by weight-sharing GRU layers. Thus, this method is more effective in building mutual correlations than LA-LSTM (6.14%) and TSIN (0.40%). According to Fig. 5(a), the most confusing emotion category is the $neutral$ category. It is considered that the data distribution of each emotion category is imbalanced in each session and in the indiscriminate features for the $neutral$ emotion. The emotion classification performance is mainly restricted by the $neutral$ emotion in LOSO splitting.

In LOPO splitting, the proposed method achieves 79.18% WA and 79.97% UA with the IEMOCAP corpus. There is a slight improvement over that with LOSO splitting because the TDFNet model can benefit from the speaker-related features in each session and enable the utilization of context-related information from the DST module. According to the experimental results in III, TDFNet further expands the performance gap with related methods, LSTM-CNN (14.07%), DRN (12.57%), DNN-BN (4.47%) and TSIN (1.86%). Compared with the related methods, as previously mentioned, TDFNet can effectively utilize speaker-related and context-related information, with apparent improvements. According to Fig. 5(b), the misclassified proportion of the $neutral$ emotion category is reduced. Due to the speaker-related information engaging, the emotion states of each speaker are learned, and more discriminate features are captured by the model to represent the $neutral$ emotion category.

In RA splitting, the proposed method achieves 82.08% WA and 82.57% UA, providing the background information for each circumstance. In fact, speaker-related information is useful for

TABLE I
PERFORMANCE OF UA AND WA (%) FOR THE IEMOCAP CORPORA WITH LOSO, LOPO, AND RA SPLITTING

LOSO	WA	UA	**LOPO**	WA	UA	**RA**	WA	UA
Session1	75.60%	77.90%	Ses01M	75.00%	75.76%	Fold1	83.54%	83.93%
Session2	77.64%	78.34%	Ses01F	78.66%	79.79%	Fold2	83.00%	83.44%
Session3	74.65%	74.79%	Ses02M	80.63%	83.95%	Fold3	81.56%	81.70%
Session4	75.87%	75.75%	Ses02F	80.14%	80.26%	Fold4	81.37%	81.91%
Session5	77.52%	78.41%	Ses03M	73.95%	74.30%	Fold5	80.29%	80.25%
-	-	-	Ses03F	81.00%	81.16%	Fold6	85.35%	85.81%
-	-	-	Ses04M	81.82%	82.08%	Fold7	80.65%	80.81%
-	-	-	Ses04F	76.62%	77.65%	Fold8	82.28%	82.27%
-	-	-	Ses05M	82.88%	83.35%	Fold9	80.11%	81.30%
-	-	-	Ses05F	81.07%	81.36%	Fold10	82.64%	84.29%
mean	**76.26%**	**77.04%**	**mean**	**79.18%**	**79.97%**	**mean**	**82.08%**	**82.57%**

All results are analyzed in terms of percentage (%). The mean UA and WA results are highlighted in bold. "-" Denotes that the corresponding result is not provided.

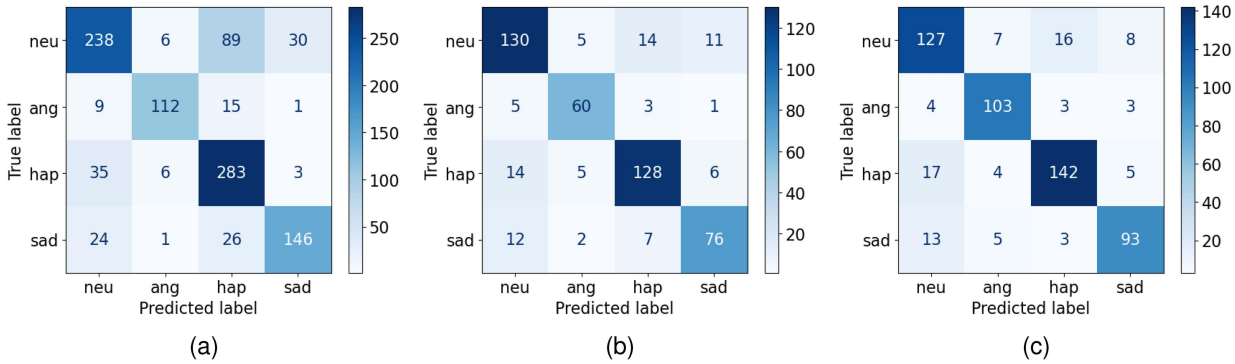

Fig. 5. Confusion matrices of recognition accuracy with the proposed TDFNet method: (a) Confusion matrix with LOSO splitting; (b) Confusion matrix with LOPO splitting; (c) Confusion matrix with RA splitting.

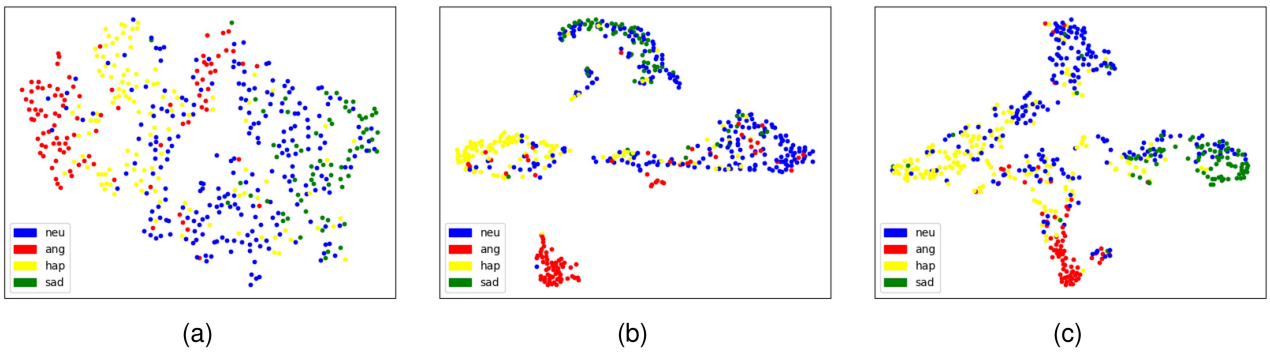

Fig. 6. Visualizations of ablation studies using t-SNE for LOPO splitting. Different emotion categories are shown in different colors and shapes: (a) w/o DST module; (b) w/o MT module; (c) w/o ME module.

emotion recognition. These experimental results demonstrate that this method can effectively leverage mutual correlations, speaker-related information, and deep-scale features to predict emotion categories. According to Table IV, TDFNet achieves state-of-the-art scores and is more effective than previous methods. The key to the significant improvements over related methods is that TDFNet can extract mutual correlations from speaker-related information and capture fine-grained emotional features from attentively aligned multimodal information.

Compared with the multiscale fusion methods, the DST module utilizes the weight-sharing GRU to align multimodal information attentively and extracts emotion-related features from multimedia. The lack of aligned multimodal information restricts the MSCNN (1.17%) and MHA-2 (4.93%). The WISE (6.17%) and TSIN (2.80%) methods do not have deep-scale information to model the fine-grained emotional features. The experimental results prove that the attentively aligned deep-scale information and speaker-related mutual correlations are

TABLE II
COMPARISONS OF UA AND WA WITH THE STATE-OF-THE-ART METHODS AND LOSO SPLITTING

Method	WA	UA
RNN-WPA [10]	63.50%	58.80%
ACNN-AE [19]	-	59.54%
TBE [42]	-	61.80%
EF-CS [9]	-	65.60%
IAAN [11]	64.70%	66.30%
LA-LSTM [23]	72.50%	70.90%
TSIN [41]	74.92%	76.64%
Our TDFNet	**76.26%**	**77.04%**

The experimental results are analyzed in terms of percentage (%). The best results are highlighted in bold. "-" Denotes that the corresponding result is not provided.

TABLE III
COMPARISONS OF UA AND WA WITH THE STATE-OF-THE-ART METHODS AND LOPO SPLITTING

Method	WA	UA
LSTM-CNN [4]	64.97%	65.90%
DRN [6]	-	67.40%
DNN-BN [7]	73.70%	75.50%
TSIN [41]	76.23%	78.11%
Our TDFNet	**79.18%**	**79.97%**

The experimental results are analyzed in terms of percentage (%). The best results are highlighted in bold. "-" Denotes that the corresponding result is not provided.

TABLE IV
COMPARISONS OF UA AND WA WITH THE STATE-OF-THE-ART METHODS AND RA SPLITTING

Method	WA	UA
TFCNN [15]	70.34%	70.78%
STSER [34]	71.06%	72.05%
MDRE [53]	71.80%	-
WISE [27]	75.90%	76.40%
MHA-2 [14]	76.50%	77.60%
TSIN [41]	78.74%	79.77%
MSCNN [16]	80.30%	81.40%
Our TDFNet	**82.08%**	**82.57%**

The experimental results are analyzed in terms of percentage (%). The best results are highlighted in bold. "-" Denotes that the corresponding result is not provided.

practical for multimodal emotion recognition. According to Fig. 5(c), $neutral - happiness$ and $neutral - sadness$ are the most misclassified emotion pairs in RA splitting. Compared with that with LOPO splitting, there are fewer misclassified data in the $neutral$ category, and $anger$ is more discriminated and significantly distinguished from other emotion categories.

D. Ablation Studies

To further demonstrate that the attentively aligned deep-scale features from the DST module and the speaker-related mutual correlations from the MT module are effective for multimodal emotion recognition, we conduct ablation studies to measure the contributions of each component in TDFNet. The experimental ablation results are shown in Table V.

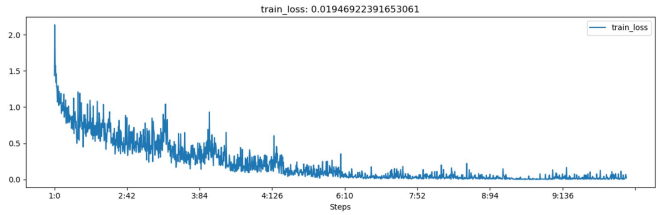

Fig. 7. LOSS variation of the TDFNet model on the training set.

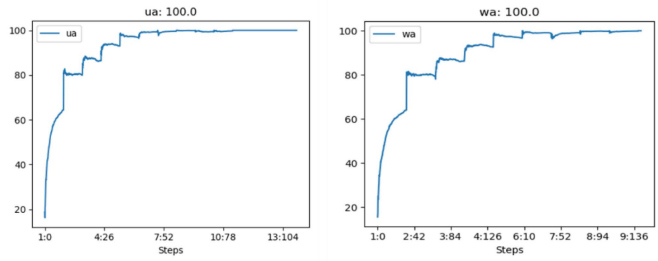

Fig. 8. UA and WA variations of the TDFNet model on the training set.

TABLE V
COMPARISONS OF WA AND UA FOR ABLATION STUDIES IN LOSO SPLITTING

Model	WA	UA
AUDIO PATH	71.49%	72.92%
TEXT PATH	67.68%	69.03%
w\o DST	74.82%	75.65%
w\o MT	73.88%	74.73%
w\o ME	74.58%	75.44%
our TDFNet	**76.26%**	**77.04%**

'w \ o' denotes the abbreviation of 'without'. The best results are highlighted in bold. All results are analyzed in terms of percentage (%).

As Table V shows, the audio path of TDFNet achieves 72.92% UA, and the text path achieves 69.03% UA with LOSO splitting. The unimodal method can achieve competitive performance from the robust and general pretrained embeddings in HUBERT and BERT. TDFNet without the DST module achieves 75.65% UA. The TDFNet model benefits from the multimodal feature alignment provided by the DST module to effectively learn and extract sentiment features from multimodal information and demonstrates that multimodal feature alignment can help improve performance in multimodal emotion recognition tasks. TDFNet without the MT module achieves 74.73% UA. The MT module improves the UA performance by 2.31% compared with that of TDFNet. The design concept of the MT module helps the TDFNet model learn the emotion characteristics of the speaker in each conversation and the common features of emotion in multimodal messages, and the experimental results also show that the MT module is very effective for multimodal emotion recognition. In addition, TDFNet without the ME module uses speech spectrogram features as the audio feature input and achieves 75.44% UA. The TDFNet model fails to effectively

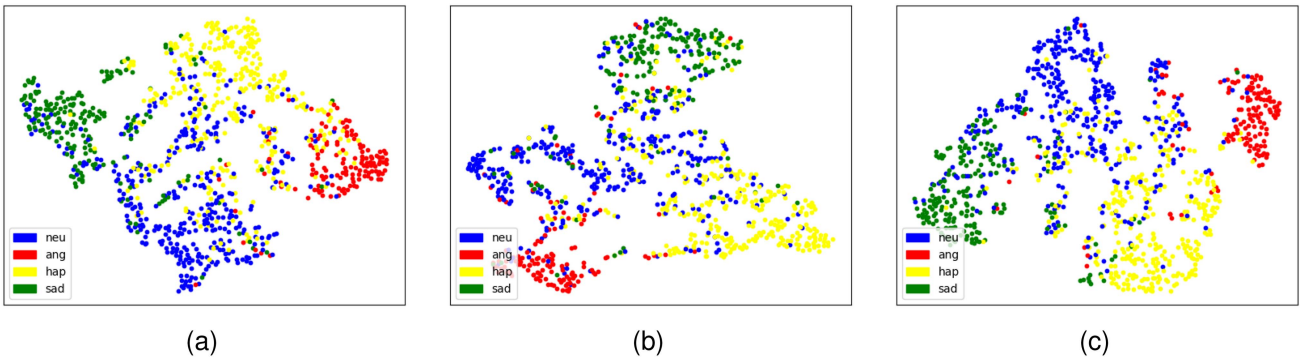

Fig. 9. Visualizations of the "Speech" model, the "Text" model and the TDFNet framework using t-SNE for LOSO splitting. Different emotion categories are shown in different colors and shapes: (a) Speech embedding; (b) Text embedding; (c) TDFNet embedding.

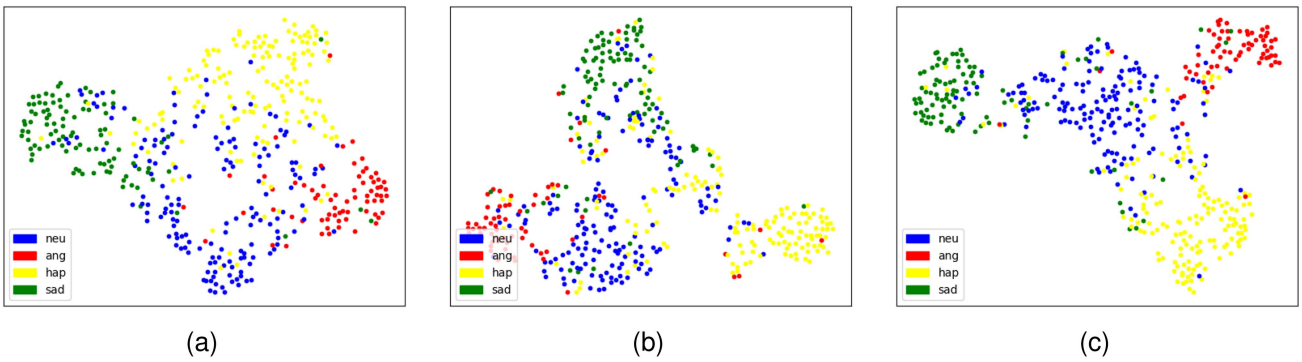

Fig. 10. Visualizations of the "Speech" model, the "Text" model and the TDFNet framework using t-SNE for LOPO splitting. Different emotion categories are shown in different colors and shapes: (a) Speech embedding; (b) Text embedding; (c) TDFNet embedding.

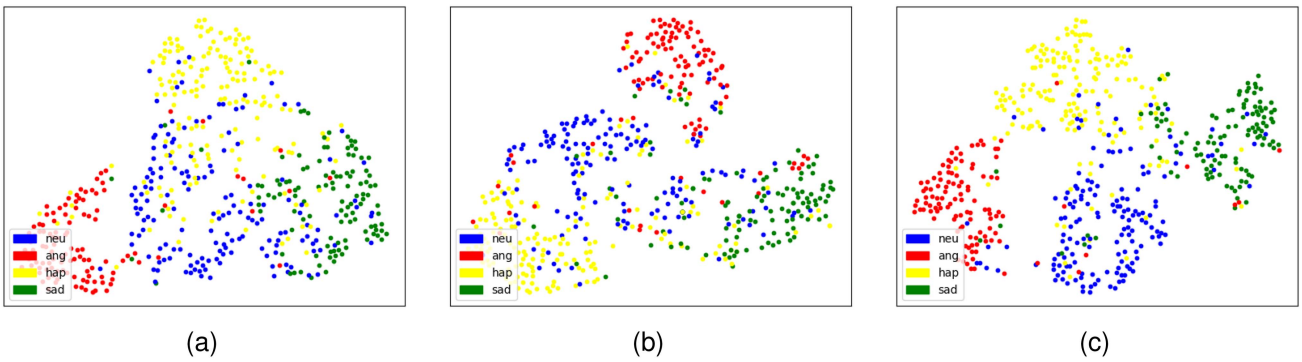

Fig. 11. Visualizations of the "Speech" model, the "Text" model and the TDFNet framework using t-SNE for RA splitting. Different emotion categories are shown in different colors and shapes: (a) Speech embedding; (b) Text embedding; (c) TDFNet embedding.

learn more emotionally relevant features in the speech spectrogram features due to the lack of a priori knowledge provided by the large amount of unlabeled data pretraining. The t-SNE embeddings for each of the above items are shown in Fig. 6.

E. Visualized Analysis

We adopt the t-SNE dimension reduction method to describe the distance of each emotion category feature aggregated by TDFNet. The t-SNE embeddings of speech, text, and TDFNet with LOSO, LOPO, and RA splitting are shown in Figs. 9, 10, and 11.

As illustrated in Fig. 9, the t-SNE embedding from TDFNet is more distinguished than that of the speech and text paths. In the speech and text paths, the $neutral$ emotion category is more challenging to distinguish from other emotions. In LOSO splitting, its performance is limited without preacquired speaker-related information and influenced by the imbalanced data distribution. In Fig. 10, with one speaker information trained and speaker-related mutual correlations modeled, the t-SNE embedding of TDFNet is more distinguished than that of LOSO splitting. In particular, $sadness$ and $anger$ are well separated. In Fig. 11, the t-SNE embedding of TDFNet with RA splitting aggregates the most distinguishing features. We

notice that the $neutral$ emotion data are misclassified into other emotion categories, but nearly none are misclassified into the $neutral$ category.

V. Conclusion

This article proposes a transformer-based deep-scale fusion network for multimodal emotion recognition. TDFNet benefits from robust and universal pretrained embeddings, attentively aligned deep-scale features, and speaker-related mutual correlations from multimedia resources. This method is specifically evaluated with the IEMOCAP corpus and three comparable and reasonable data splitting strategies. Furthermore, the experimental results show that TDFNet can effectively extract attentively aligned deep-scale features and aggregate the mutual correlations with speaker-related information into utterance-level representations. In fact, TDFNet achieves a significant improvement over the previous state-of-the-art methods. The visualized analysis is illustrated and reveals that TDFNet can generate a discriminate representation of emotions and improve the performance of multimodal emotion recognition.

References

[1] B. W. Schuller, "Speech emotion recognition two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[2] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[3] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47795–47814, 2021.

[4] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Proc. Interspeech*, 2018, pp. 247–251.

[5] W.-C. Lin, K. Sridhar, and C. Busso, "DeepEmoCluster: A semi-supervised framework for latent cluster representation of speech emotions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 7263–7267.

[6] R. Li, Z. Wu, J. Jia, S. Zhao, and H. Meng, "Dilated residual network with multi-head self-attention for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6675–6679.

[7] E. Kim and J. W. Shin, "DNN-based emotion recognition based on bottleneck acoustic features and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6720–6724.

[8] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3507–3511.

[9] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, "Fusion approaches for emotion recognition from speech using acoustic and text-based features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 6484–6488.

[10] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 2227–2231.

[11] S.-L. Yeh, Y.-S. Lin, and C.-C. Lee, "An interaction-aware attention network for speech emotion recognition in spoken dialogs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6685–6689.

[12] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2003, pp. I–401.

[13] S. Mao, D. Tao, G. Zhang, P. C. Ching, and T. Lee, "Revisiting hidden Markov models for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6715–6719.

[14] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 2822–2826.

[15] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 7174–7178.

[16] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale CNN and attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 3020–3024.

[17] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6319–6323.

[18] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 439–448.

[19] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 7390–7394.

[20] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5099–5103.

[21] M. Li et al., "Contrastive unsupervised learning for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6329–6333.

[22] J. Liu, Q. Zhong, L. Ding, H. Jin, B. Du, and D. Tao, "Unified instance and knowledge alignment pretraining for aspect-based sentiment analysis," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 2629–2642, 2023.

[23] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," in *Proc. Interspeech*, 2019, pp. 3569–3573.

[24] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Multimodal affective analysis using hierarchical attention strategy with word-level alignment," in *Proc. Conf. Assoc. Comput. Linguist. Meeting*, 2018, pp. 2225–2235.

[25] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Attention driven fusion for multi-modal emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 3227–3231.

[26] J. Sebastian and P. Pierucci, "Fusion techniques for utterance-level emotion recognition combining speech and transcripts," in *Proc. Interspeech*, 2019, pp. 51–55.

[27] G. Shen et al., "WISE: Word-level interaction-based multimodal fusion for speech emotion recognition," in *Proc. Interspeech*, 2020, pp. 369–373.

[28] Z. Lian, B. Liu, and J. Tao, "CTNet: Conversational transformer network for emotion recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 985–1000, 2021.

[29] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 1, pp. 31–43, Jan. 2018.

[30] A. Chatziagapi et al., "Data augmentation using GANs for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 171–175.

[31] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 2697–2709, 2020.

[32] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[33] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 4171–4186.

[34] M. Chen and X. Zhao, "A multi-scale fusion framework for bimodal speech emotion recognition," in *Proc. Interspeech*, 2020, pp. 374–378.

[35] X. Wu et al., "Speech emotion recognition using sequential capsule networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 3280–3291, 2021.

[36] J. Liu, Y. Song, L. Wang, J. Dang, and R. Yu, "Time-frequency representation learning with graph convolutional network for dialogue-level speech emotion recognition," in *Proc. Interspeech*, 2021, pp. 4523–4527.

[37] P. Tzirakis, J. Zhang, and B. W. Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5089–5093.

[38] X. Wu et al., "Speech emotion recognition using capsule networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 6695–6699.

[39] Q. Zhong, L. Ding, J. Liu, B. Du, H. Jin, and D. Tao, "Knowledge graph augmented network towards multiview representation learning for aspect-based sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 10098–10111, Oct. 2023.

[40] A. Shirian and T. Guha, "Compact graph architecture for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6284–6288.
[41] B. Chen, Q. Cao, M. Hou, Z. Zhang, G. Lu, and D. Zhang, "Multimodal emotion recognition with temporal and semantic consistency," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 3592–3603, 2021.
[42] R. Zhang, H. Wu, W. Li, D. Jiang, W. Zou, and X. Li, "Transformer based unsupervised pre-training for acoustic representation learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 6933–6937.
[43] G. Paraskevopoulos, E. Tzinis, N. Ellinas, T. Giannakopoulos, and A. Potamianos, "Unsupervised low-rank representations for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 939–943.
[44] Z. Lian, J. Tao, B. Liu, and J. Huang, "Unsupervised representation learning with future observation prediction for speech emotion recognition," in *Proc. Interspeech*, 2019, pp. 3840–3844.
[45] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *Proc. Interspeech*, 2020, pp. 3830–3834.
[46] M. Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," Mar. 2022. [Online]. Available: https://hal.science/hal-03601303
[47] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[48] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
[49] B. Tripathy, S. Anveshrithaa, and S. Ghela, "t-Distributed stochastic neighbor embedding (t-SNE)," in *Unsupervised Learning Approaches for Dimensionality Reduction and Data Visualization*. Boca Raton, FL, USA: CRC Press, 2021, pp. 127–135.
[50] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[51] V. Heusser, N. Freymuth, S. Constantin, and A. Waibel, "Bimodal speech emotion recognition using pre-trained language models," 2019, *arXiv:1912.02610*.
[52] G. Xu, W. Li, and J. Liu, "A social emotion classification approach using multi-model fusion," *Future Gener. Comput. Syst.*, vol. 102, pp. 347–356, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X1930888X
[53] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Lang. Technol. Workshop*, 2018, pp. 112–118.
[54] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019, *arXiv:1711.05101*.
[55] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Commun.*, vol. 120, pp. 11–19, 2020.

Zhengdao Zhao received the B.S. degree from the Harbin Engineering University, Harbin, China, in 2020. He is currently working toward the master's degree with High Performance Computing Laboratory, Harbin Engineering University. His research interests include deep learning, affective computing, speech emotion recognition, and multimodal emotion recognition.

Yuhua Wang received the Ph.D. degree in computer science and technology from Harbin Engineering University, Harbin, China, in 2011. He is currently an Associate Professor with the High Performance Computing Research Center, Harbin Engineering University. He has authored or coauthored more than 30 technical papers at prestigious journals and conferences. His research interests include speech emotion recognition, parallel computing, and image processing.

Guang Shen received the B.S. degree from the Harbin Engineering University, Harbin, China, in 2018, where he is currently working toward the master's degree. His research interests include deep learning, speech emotion recognition, and natural language processing.

Yuezhu Xu received the M.S. and Ph.D. degrees from the College of Computer Science and Technology, Harbin Engineering University, China, in 2005 and 2010, respectively. She is currently a Lecturer with the College of Computer Science and Technology, Harbin Engineering University. Her research interests include high performance computing and parallel computing, image recognition, speech emotion recognition, and industrial Big Data.

Jiayuan Zhang received the B.S. degree from Hohai University, Nanjing, China, in 2022. He is currently working toward the master's degree with Harbin Engineering University, Harbin, China. His research interests include deep learning, speech emotion recognition, and relevant applications.