Multi-channel Weight-sharing Autoencoder Based on Cascade Multi-head Attention for Multimodal Emotion Recognition

Jiahao Zheng, Sen Zhang, Zilu Wang, *Member, IEEE,* Xiaoping Wang, *Senior Member, IEEE,* and Zhigang Zeng, *Fellow, IEEE*

***Abstract*—Multimodal Emotion Recognition is challenging because of the heterogeneity gap among different modalities. Due to the powerful ability of feature abstraction, Deep Neural Networks (DNNs) have exhibited significant success in bridging the heterogeneity gap in cross-modal retrieval and generation tasks. In this work, a DNNs-based Multi-channel Weight-sharing Autoencoder with Cascade Multi-head Attention (MCWSA-CMHA) is proposed to generically address the affective heterogeneity gap in MER. Specifically, multimodal heterogeneity features are extracted by multiple independent encoders, and then a scalable heterogeneous feature fusion module (CMHA) is realized by connecting multiple multi-head attention modules in series. The core of the proposed algorithm is to reduce the heterogeneity between the output features of different encoders through the unsupervised training of MCWSA, and then to model the affective interactions between different modal features through the supervised training of CMHA. Experimental results demonstrate that the proposed MCWSA-CMHA achieves outperformance on two publicly available datasets compared with the state-of-the-art techniques. In addition, visualization experiments and approximation experiments are used to verify the effectiveness of each module in the proposed algorithm, and the experimental results show that the proposed MCWSA-CMHA can mine more emotion-related information among multimodal features compared with other fusion methods.**

***Index Terms*—Multimodal emotion recognition (MER), autoencoder, multi-head attention mechanism.**

I. Introduction

With the development of information technology, human-computer interaction has become an important research direction in the computer field. In the last few decades, researchers have devoted themselves to studying how to integrate emotion into human-computer interactions, which is called "affective computing" [1]. The purpose of affective computing is to allow machines to have the ability to model and interpret human emotions [2], so that they can be applied to voice call centers to provide better services [3], to develop voice-based medical diagnosis systems which can aid in the diagnosis of depression [4], and so on. One of the most important steps in affective computing is to realize emotion recognition and human beings usually express their emotions through voice, posture, facial expressions, and so on [5]. Besides the data of these different modalities are potentially interrelated and complementary to jointly construct human emotion expression forms. Therefore, the key to realizing emotion recognition lies in how to fuse multimodal information [6]. However, multimodal data is highly heterogeneous in nature [7], which brings great challenges to the research of emotion recognition based on multimodal data.

Fortunately, the rapid development of deep learning in recent decades has brought breakthroughs for multimodal fusion. Based on the Deep Neural Network (DNN), the most common fusion method is to extract features from multiple modalities through DNN, and then directly concatenate the obtained feature vectors [8] [9], but this method fails to learn the intra-modality and inter-modality dynamics. To address this issue, Tensor Fusion Network (TFN) [10], in which the unimodal, bimodal, and trimodal interactions are modeled using a 3-fold Cartesian product, provides a possible solution. Furthermore, low rank multimodal fusion technique (LMF) [11] based on TFN is proposed to realize cross-modal information interaction while reducing the computational cost. However, both TFN and LMF have complex network structures, which leads to high computational cost in the process of modal fusion. On the contrary, [12] proposed Bimodal Deep Autoencoder (BDA) based on Restricted Boltzmann Machine (RBM) [13] and autoencoder [14], which was completely different from the fusion method of TFN and LFM. BDA can fully explore and extract the intra-modality and inter-modality interaction information at a relatively low computational cost. However, due to the use of RBM, it is difficult for traditional optimization algorithms to be used for training RBM and its feature extraction ability is also insufficient, making the BDA unable to effectively learn discriminative affective features which are crucial for emotion recognition. In addition to the above conventional models, Tsai *et al*. [15] proposed a modal fusion model based on attention mechanisms which can learn long term dependencies across modalities from unaligned multimodal stream effectively, but this model only considered the interactions between paired modalities. Inspired by these models, in this work, we aim to build a BDA-based and attention-based model that can effectively mine the interactions between multiple modalities.

It should be noted that to achieve effective information interaction and fusion between modalities, data preprocessing methods and feature extraction methods are also crucial. It

This work was supported in part by the National Natural Science Foundation of China under Grants 61876209, 61936004 and U1913602. (Corresponding author: Xiaoping Wang; e-mail: wangxiaoping@hust.edu.cn)

J. Zheng, S. Zhang, Z. Wang, X. Wang and Z. Zeng are with the School of Artificial Intelligence and Automation and the Key Laboratory of Image Processing and Intelligent Control of Education Ministry of China, Huazhong University of Science and Technology, Wuhan 430074, China.

has been recently found that Convolutional Neural Networks (CNNs) have strong feature extraction capabilities [16], [17], [18]. Several works [19], [20] have successfully employed CNN for Speech Emotion Recognition (SER). Furthermore, the multi-layer structure of the Deep Convolutional Neural Network (DCNN) can learn more abstract and robust features [21], which is important for emotion recognition in complex environments. In [22], the authors converted the audio data into log Mel-spectrograms (static, delta and delta delta) which is similar to the representation of RGB image, and used temporal pooling strategy to fix the shape of audio feature with variable length in the time axis. In this way, 2D DCNN with excellent performance in visual tasks can also be used to extract features from 1D speech signals, and the same audio data processing method is also used in [23]. In addition, [23] used a DCNN named IR50 to extract facial expressions from video data and video features are concatenated with audio features for emotion recognition. In a word, the prior works have empirically confirmed that DCNN can effectively extract robust affective features.

The powerful ability of DCNN in feature extraction, the capability of attention mechanism in mining cross-modal dependencies, and the success of BDA in multimodal fusion motivate us to combine them for emotion recognition. But in order to realize speech emotion recognition system based on DCNN, attention mechanism and BDA, there are three issues that need to be addressed. First, when deconvolution is used in the up-sampling process and the regression loss is constructed by using the output of the decoder, the autoencoder structure based on convolution requires a fixed input size. In addition, the mini-batch gradient descent training strategy also requires a fixed shape of the input data. However, the dimension of speech emotion data in the time axis is usually not fixed, so appropriate data processing methods are required. The second is how to combine the output features of the convolutional encoder of each modality which is crucial for mining the interactions between multiple modalities. It is important to note that the highly heterogeneous nature and unaligned nature of different modal data is a daunting challenge to data processing and autoencoder structure design. Finally, the number of samples of the existing speech emotion dataset is very limited, so it is difficult for the network to learn all the functions at one time [24].

In this work, a Multi-channel Weight-sharing Autoencoder with Cascade Multi-head Attention (MCWSA-CMHA) algorithm is proposed to solve the above problems. As illustrated in Fig. 1, in the data preprocessing part, the frame images extracted from the video are sent to the ResNet [17] which is pre-trained on the large-scale ImageNet dataset [21]. For the audio data, the processing method in [22] is adopted to convert 1D audio data into a log Mel-spectrograms with three channels, so that it has image attributed to some extent. In order to meet the requirement of the autoencoder network for data dimension, Large-scale Time Pyramid Pooling (LTPP) is introduced in the data preprocessing part. Then, the data with fixed dimensions is sent to MCWSA-CMHA to perform unsupervised reconstruction training. Finally, after the reconstruction training, the output of each modal encoder is combined by CMHA and the coding results are sent to the prediction part to realize the emotion classification task. The effectiveness of the proposed novel algorithm in SER is demonstrated by extensive experiments on two large and widely used emotional datasets, i.e., the IEMOCAP [25] dataset and the MSP-IMRPOV dataset [26].

The contributions of this work are summarized as follows:

1) We propose LTPP to deal with multimodal emotional data with an unfixed time dimension. The dimensions of the data are fixed and time clues are retained as much as possible by applying LTPP to the data processing.
2) The proposed MCWSA-CMHA algorithm realizes the full fusion of multimodal features, and mines the interactive information between different modal data. Specifically, MCWSA inherits the modal fusion capability of the BDA model, meanwhile has scalability and strong feature extraction capability to cope with highly heterogeneous multiple modalities; while the proposed CMHA is capable of simultaneously modeling the interactions between multiple modal features and is also scalable.
3) Comprehensive experiments are designed to verify the effectiveness of the proposed model, and both qualitative and quantitative analyses show that the proposed model has a satisfactory performance.

The remainder of this work is structured as follows. The related work is reviewed in Section II. Section III describes in detail the model proposed in this work. The experimental setup and results are introduced in Section IV and section V respectively. The conclusion is presented in section VI.

II. RELATED WORK

Our framework focuses on the fusion of multimodal and the key of the fusion method is the autoencoder and attention mechanism. Therefore, in this section, we first briefly overview the previous works on multimodal fusion in SER. In addition, we review in detail the related works of autoencoder and attention mechanism in SER which are the core of the proposed algorithm.

A. Multimodal Fusion for Emotion Recognition

Due to the crucial role of modal fusion for emotion recognition, a number of studies have focused on modal fusion and various fusion methods have been proposed. They can be roughly divided into two categories, one is decision-level fusion, the other is feature-level fusion. Compared with decision-level fusion, feature-level fusion has the advantage of modeling inter-modal dynamics in an efficient manner. Therefore, we mainly focus on the feature-level fusion.

There are many conventional models for feature-level fusion. Ngiam *et al*. [12] extended the autoencoder which was proposed for data dimension reduction to Bi-channel model named bimodal deep autoencoder. In this model, two separated autoencoders with independent encoders and decoders share information by a common latent representation layer to capture the cross-modal dynamics. Similar to the work of [12], Wang *et al*. [27] imposed orthogonal regularization on the weighting matrices of the bimodal deep autoencoder model to

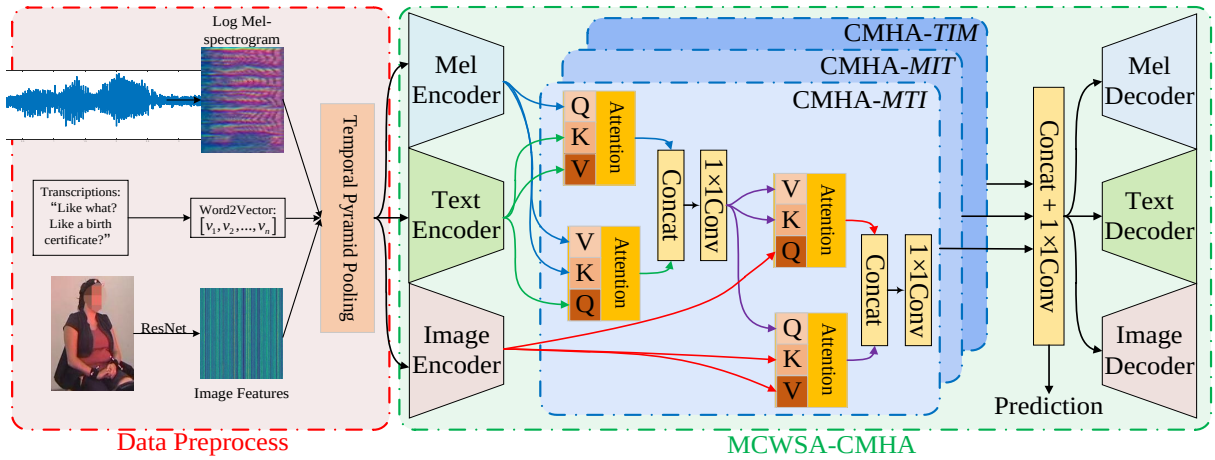

Fig. 1: Illustration of the proposed MCWSA-CMHA. The red dashed box represents the data preprocessing module. The blue dashed box represents the proposed CMHA. Different shades of blue dashed boxes represent different orders of modal fusion. Q, K, and V in attention module represent Query, Key, and Value respectively.

reduce redundant information. Silberer *et al*. [28] also adopted bimodal deep autoencoder to learn higher-level embeddings from textual and visual input. However, all of them did not apply this fusion method to multimodal emotion recognition task and most of their algorithms were limited to RBM. Our approach is motivated by the idea of autoencoder fusion model, as it can effectively integrate information of different modalities through reconstruction training. Apart from the autoencoder model, researchers applied Canonical Correlation Analysis (CCA) [29], which is originally used for measuring the correlation between two vectors, to multimodal fusion from another novel perspective. Specifically, Andrew *et al*. [30] extended CCA to deep CCA (DCCA) which can learn complex non-linear transformation of different modalities through DNN. Furthermore, literature [31], [32] introduced DCCA to multimodal emotion recognition by taking advantage of its characteristics of non-linear transformation and its ability to measure correlations of features.

In addition to these conventional models, another stream of research in feature-level fusion utilizes some newly developed technologies, such as attention mechanism and Generative Adversarial Networks (GANs) to mine the subtle correlation between modalities. For instance, Zadeh *et al*. [6] proposed Memory Fusion Network which modeled view-specific and cross-view interactions through time with a special attention mechanism and summarized through time with a Multi-view Gated Memory. The authors also proposed another model to discovering interactions between modalities through time using a neural component called Multi-attention Block and storing them in the hybrid memory of a recurrent component [33]. Sahu *et al*. [7] proposed a novel fusion model based on (GANs). In addition, Peng *et al*. proposed Cross-modal Generative Adversarial Networks (CM-GANs) [34] which can be exploited to model different modal data distribution for bridging the heterogeneity gap.

Both streams of research mentioned above show that there are correlations between different modal data, but due to the heterogeneous nature of the data, such correlation is difficult to discover, so the method of modal fusion is of vital importance.

B. Autoencoder in Emotion Recognition

Autoencoder is a popular unsupervised learning model. Due to its unsupervised nature, it has a wide range of applications in the field of emotion recognition which is extremely lacking in labeled data. Deng *et al*. [35] proposed a sparse autoencoder model which can discover knowledge in acoustic features from small target data to improve the performance of SER. The authors also proposed a semisupervised autoencoder to combine the task of reconstruction with the task of emotion recognition [36]. The authors' work shows that autoencoder can learn discriminative latent code that leads to significant performance improvement in an unsupervised manner.

Besides the above conventional autoencoder models, Du *et al*. [37] proposed a spatio-temporal autoencoder with the ability of capturing discriminative long-term dynamic to video based dimensional emotion recognition. Dumpala *et al*. [31] proposed a cross-modal autoencoder combined with DCCA for sentiment classification. In [38], the authors used three autoencoders to extract the features of three modalities and introduced cross-modality distribution matching to the latent representation. Variational autoencoders (VAE) with generative capability were also used in emotion recognition. For instance, Latif *et al*. [39] used VAE to learn the latent representation of speech signals and applied it to emotion classification. Xiao *et al*. [40] used VAE for multi-task learning with emotion recognition as primary task, and learning the input distribution and the common representations of different domains as auxiliary tasks. Apart from VAE, in [41], adversarial autoencoder (AAE) as a generation model was also used in multi-task learning for speech emotion recognition.

Although the above mentioned literature proved the effectiveness of autoencoders in the field of emotion recognition, most of them were only for single-modal data, or used multiple independent autoencoders to handle different modalities. Therefore, there is still a lack of in-depth research on multimodal fusion based on autoencoders.

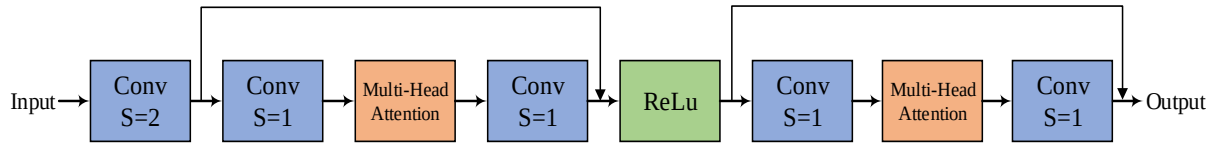

Fig. 2: Base Block of Encoder. The blue block represents the convolutional layer, the S in the blue block is the step size of the convolution, the orange block stands for the multi-head attention layer, the green block represents the activation layer and the arrows denote the flow of data.

C. Multi-head Attention in SER

The Attention model was originally used in machine translation [42]. In the field of artificial intelligence, attention has become an important part of the structure of neural networks and has a large number of applications in automatic speech recognition [43], object tracker [44], natural language processing [45]. The prior works have empirically demonstrated the effectiveness of attention mechanism in learning long term dependencies. Therefore, attention mechanism is very suitable for emotion recognition task with temporal nature. Specifically, Mirsamadi *et al*. [46] presented a feature pooling strategy with local attention, which can focus on specific regions that are more emotionally salient. Li *et al*. [47] used Multi-head Self-attention mechanism [48] to model the relative dependencies between elements. Furthermore, Tsai *et al*. [15] introduced transformer network [48], which has made great achievements in machine translation, to MER task. Our approach is also motivated by Tsai's work, as it introduced transformer attention mechanism with the ability to infer long term dependencies across modalities into MER. However, in Tsai's work, the interaction between pairs of three modalities is the main consideration, while our approach can mine the cross-modal dynamics between three or more modalities simultaneously.

III. METHOD

A. Base Block of Encoder and Decoder

The research in SER shows that deep networks can improve robustness to noisy data by learning more abstract contextual feature [49] [50], but a notorious problem of vanishing/exploding gradients arises [51] when more layers are stacked in CNN. Therefore, as depicted in Fig. 2, the residual structure is used as the basic module of the encoder. For the decoder, we use deconvolution with learnable parameters to implement up-sampling. According to [47] [48], Multi-head Self-attention mechanism can focus on attention-capturing factors in speech, thereby contributing to speech emotion recognition. Accordingly, the attention mechanism is also introduced into the encoder in this work. The input to the Self-attention module consists of d_k,d_k,d_v dimensional queries (Q), keys (K), values (V). The attention matrix of output is calculated as:

$$\text{Attention}\,(Q,K,V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{1}$$

In contrast to the traditional attention mechanism, the Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. Multi-head attention first performs h times different linear projections to queries, keys, and values, and then performs the attention function in parallel, yielding (d_v/h)-dimensional output values. These values are concatenated and once again projected, resulting in the final values [48]. The Multi-head Self-attention is calculated as:

$$\begin{aligned}\text{MultiHead}\,(Q,K,V) &= \text{Concat}\,(\,\text{head}_1,\ldots,\,\text{head}_h)\,W^0\\ \text{where head}_i &= \text{Attention}\,\Big(QW_i^Q,\\ &\qquad KW_i^K, VW_i^V\Big)\end{aligned} \tag{2}$$

where $W_i^Q \in \mathbb{R}^{d_k/h}$, $W_i^K \in \mathbb{R}^{d_k/h}$, $W_i^V \in \mathbb{R}^{d_v/h}$ are the weight matrices that implement the projections. W^O is the weight matrix in linear output function.

B. Cascade Multi-head Attention

In this section, we describe the proposed Cascade Multi-head Attention (CMHA) for multimodal feature fusion and affective interaction mining. As shown in the blue dashed box in Fig. 1, by connecting multi-head attention modules in series, CMHA has the ability to mine the interactions between three or more modal data with heterogeneous and non-aligned properties. Specifically, take the fusion of three modalities (video, audio, and text) as an example. Firstly, these three modalities are preprocessed to features: M for audio modality, I for visual modality and T for textual modality. Then, the features of each modality are fed into their respective encoders to get their respective latent code: $Z^M \in \mathbb{R}^{F\times C}$, $Z^I \in \mathbb{R}^{F\times C}$, and $Z^T \in \mathbb{R}^{F\times C}$ (Since different encoders are used for different modalities, the encoder can be designed so that the latent code of each modality has the same dimension).

Similar to Tsai's work [15], we adopt multi-head attention module to fuse cross-modal information. Take the fusion of modal $Z^M \in \mathbb{R}^{F\times C}$ and $Z^T \in \mathbb{R}^{F\times C}$ as an example, we define the parallel attention layers as h. Then the multi-head Query, Key, and Value can be defined as $Q_i^M = Z_i^M \times W_i^Q$, $K_i^T = Z_i^T \times W_i^K$, and $V_i^T = Z_i^T \times W_i^V$, respectively, and $i \in \{1,\ \cdots,\ h\}$. $Z_i^M \in \mathbb{R}^{F\times \frac{C}{h}}$ and $Z_i^T \in \mathbb{R}^{F\times \frac{C}{h}}$ are divided by Z^M and Z^T in the channel direction, respectively. $W_i^Q \in \mathbb{R}^{\frac{C}{h}\times\frac{d_k}{h}}$, $W_i^K \in \mathbb{R}^{\frac{C}{h}\times\frac{d_k}{h}}$, and $W_i^V \in \mathbb{R}^{\frac{C}{h}\times\frac{d_v}{h}}$ are projection weights. For each of these we use $d_k = d_v = C$. The i-th attention layer can be calculated by the following formula:

$$\begin{aligned}\text{head}_i &= \text{Attention}\left(Q_i^M, K_i^T, V_i^T\right) \\ &= \text{softmax}\left(\frac{Q_i^M\left(K_i^T\right)^\top}{\sqrt{d_k/h}}\right)V_i^T \\ &= \text{softmax}\left(\frac{Z_i^M W_i^Q\left(Z_i^T W_i^K\right)^\top}{\sqrt{d_k/h}}\right)Z_i^T W_i^V.\end{aligned} \tag{3}$$

Then, the cross-modal information $CI^{T\to M}$ between modal M and T can be calculated by the following formula:

$$CI^{T\to M} = \text{Concat}\left(\text{head}_1, \ ..., \ \text{head}_h\right)W^O, \tag{4}$$

where $W^O \in \mathbb{R}^{d_v\times C}$ is the weight matrix in linear output function, and the dimension of $CI^{T\to M}$ is $F\times C$. By inputting features Z^M and Z^T into the multi-head attention module, the interaction weight matrix of modality M and modality T can be obtained, and then this weight matrix is multiplied by Z^T to obtain the cross-modal feature. However, only the interaction of two modalities can be modeled in this way. In order to effectively model the interaction between three or more modalities, we extend the multi-head attention module. As shown in the blue dashed box in Fig. 1, through Equation (4), the cross-modal feature $CI^{T\to M}$, which represents the interactive information between modality M and T, can be obtained. In the same way, the $CI^{M\to T}$ can be obtained. Then, the joint representation CI^{MT} of the modal M and modal T can be obtained by the following formula:

$$CI^{MT} = \text{Conv}\left(\text{Concat}\left(CI^{M\to T}, CI^{T\to M}\right)\right), \tag{5}$$

where Concat represents the concatenation operation in the channel direction and Conv represents the 1×1 convolution operation. Due to the convolution layer with dimension $1\times 1\times 2C\times C$, the dimension of CI^{MT} is same as the input $CI^{M\to T}$. When CI^{MT} and another modal feature Z^I are used as the input of a new multi-head attention module, the interactive information $CI^{MT\to I}$ between the three modalities can be simultaneously modeled, and it can be represented by the following formula:

$$\begin{aligned}CI^{MT\to I} &= \text{MultiHead}\left(Z^I, CI^{MT}, CI^{MT}\right) \\ &= \text{Concat}\ (\ \text{head}_1, \ldots, \ \text{head}_h)\,W^O, \\ \text{where head}_i &= \text{Attention}\ \Big(Z_i^I W_i^Q, \\ &\qquad CI_i^{MT} W_i^K, CI_i^{MT} W_i^V\Big).\end{aligned} \tag{6}$$

It should be noted that the calculation process of Eq. (3) and Eq. (4) is the same as that of Eq. (6), the difference between them is only that the input data is different. In the Eq. (6), Z_i^I is divided by Z^I in the channel direction. In the same way, $CI^{I\to MT}$ can be obtained. Then, similar to formula (5), the joint representation CI^{MTI} of the modal M, modal T, and modal I can be calculated by the following formula:

$$CI^{MTI} = \text{Conv}\left(\text{Concat}\left(CI^{MT\to I}, CI^{I\to MT}\right)\right) \tag{7}$$

Taking into account the sequential order of modal fusion, similar to formulas (3) to (7), CI^{MIT} and CI^{TIM} can be obtained. It should be noted that in the three-modal fusion example, there is no sequential order between the two modalities that are first fused, so when considering the modal fusion order, there are only three situations mentioned above. By connecting attention modules in series and using different modal features as queries for different attention modules, the cross-modal interaction for multiple modalities can be obtained. As shown in Fig. 1, the three cross-modal features (CI^{MTI}, CI^{MIT}, and CI^{TIM}) obtained according to different fusion orders are concatenated together to represent the three-modal fusion features.

C. Loss Function of MCWSA-CMHA

The proposed model MCWSA-CMHA needs to go through two stages of training. The first stage is unsupervised reconstruction training, and the second stage is supervised classification training. For the first stage of training, different encoders are used to encode the data of different modalities to obtain the latent code of the respective modalities and then CMHA is used to fuse the latent code of each modality to obtain a multimodal latent code. Finally, the multimodal latent code is fed into different decoders to reconstruct the input of each channel. The loss function for reconstruction training can be defined as:

$$L_{MCWSA} = L_M + L_T + L_I, \tag{8}$$

here, L_M, L_T, and L_I are losses for audio modality, text modality, and visual modality, respectively. They can be computed by:

$$\begin{aligned}L_M &= \left\|M - \left(D_M\circ F_L\circ E_M\right)(M)\right\|^2, \\ L_T &= \left\|T - \left(D_T\circ F_L\circ E_T\right)(T)\right\|^2, \\ L_I &= \left\|I - \left(D_I\circ F_L\circ E_I\right)(I)\right\|^2,\end{aligned} \tag{9}$$

where $\circ$ represents function composition. M, T, and I are the pre-processed audio data, text data, and video data, respectively. E_M, E_T, and E_I respectively refer to the Mel Encoder, Text Encoder, and Image Encoder as shown in Fig. 1. F_L refers to the fusion layers, that is, the CMHA layer and the Concat layer in Fig. 1. The parameters of this part are shared between each modality. D_M, D_T, and D_I refer to the Mel Decoder, Text Decoder, and Image Decoder, respectively.

By and large, the parameters of MCWSA, which are presented in the green dashed box in Fig. 1, are trained to minimize the objective function as:

$$\begin{aligned}&E_M, E_T, E_I, F_L, D_M, D_T, D_I = \\ &\underset{E_M,E_I,F_L,D_M,D_I}{\text{argmin}} \Big(\left\|M - \left(D_M\circ F_L\circ E_M\right)(M)\right\|^2 \\ &\quad + \left\|T - \left(D_T\circ F_L\circ E_T\right)(T)\right\|^2 \\ &\quad + \left\|I - \left(D_I\circ F_L\circ E_I\right)(I)\right\|^2\Big).\end{aligned} \tag{10}$$

Through such a process, E_M, E_T, and E_I try to extract more abstract and robust features which are then fused through F_L, while D_M, D_T, and D_I try to recover single-modal data from the fused features. Finally, through the reconstruction training, the fused features obtained from the concatenate layer have the attributes of audio data, text data, and video data at

the same time. This training process achieves a high degree of fusion of different modal data that may span very different domains, thereby bridging the heterogeneous gap between different modal features. In addition, the second stage of supervised training is based on cross entropy loss for discrete emotion classification, which will not be repeated in this work.

IV. EXPERIMENTAL SETUP

In this section, we first introduce the dataset used in this work including the IEMOCAP and MSP-IMPROV datasets, then describe the model configuration and evaluation measures.

A. Datasets

This work evaluates the performance of the proposed model on two different datasets: IEMOCAP [25] and MSP-IMPROV [26], which are widely used for emotion recognition task.

IEMOCAP: This dataset contains 10039 turns audiovisual data including audio, video, facial motion information, and context. Ten professional actors (five males and five females) participated in dyadic interactions, they were asked to express clear emotions based on three selected scripts. In addition, they were also asked to improvise dialogs in hypothetical scenarios. Then the categorical labels (i.e., angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted, and others) of each utterance from either of the actors in the dyadic interactions were annotated by three different annotators and the participant. Following the prior study [39], four categorical emotions (angry, happy, sad and neutral) are considered in this work. Furthermore, the category "excited" is included in the category "happy". Finally, 5531 utterances (1103 angry, 1636 happy, 1084 sad and 1708 neutral) are used in experiments.

MSP-IMPROV: This dataset is a multimodal emotional dataset comprised of six sessions. In each session, two actors (one male and one female) performed dyadic interactions. Each session consists of 20 target sentences and was segmented into utterances. All utterances of this corpus were labeled by five annotators on one of five categorical emotions: angry, happy, neutral, sad and others. Finally, 7798 utterances (2644 for happy, 792 for angry, 885 for sad and 3477 for neutral) are used in experiments.

B. Data Preprocessing

The datasets used in this work contain three modalities of video, audio, and text. We first extract all frames from the video clip and save them to image files. Due to the large amount of original image data, it will consume a lot of computing resources to send the original image directly to the subsequent network. Therefore, in order to compress the data amount, ResNet pre-trained on the large-scale ImageNet dataset is used to extract the image features which are denoted as IF:

$$IF \in \mathbb{R}^{N\times 2048} \tag{11}$$

where N represents the number of frames produced by a video clip, 2048 represents the number of units in the penultimate layer of the ResNet network.

For audio data, we adopt the log Mel-spectrogram which is widely used for speech emotion recognition [22] [23], then the log Mel-spectrogram with 1-D shape is converted to MF:

$$MF \in \mathbb{R}^{F\times T\times C} \tag{12}$$

where F is the number of Mel-filter banks, T is the length along the time axis associated with the audio duration and the size of Haming window, and C=3 represents the static, delta and delta-delta coefficients of Mel-spectrogram. Then this kind of three channels of spectrogram which is similar to RGB image representation can be used as the input of traditional DCNN. In detail, the log Mel-spectrogram is computed with 64 Mel-filter banks from 20 to 8000Hz and a 25ms Haming window size with 10ms overlapping. In this case, $MF \in \mathbb{R}^{64\times T\times 3}$.

For textual data, people often express their emotions relying on a set of context-specific words [52]. Therefore, in order to capture contextual information in the textual data, a well-trained Word2Vec model [53] is used to convert each word in an utterance to a 300-dimensional vector, represneted as TF:

$$TF \in \mathbb{R}^{N\times 300} \tag{13}$$

where N represents the number of words in an utterance.

Due to the varying lengths of the data samples, the data time dimension is still not fixed after the above processing steps. So it is difficult to use mini-batch gradient descent algorithm and autoencoder algorithm. In order to fix the dimension of the data, we introduce Temporal Pyramid Pooling (TPP) [54] [55], which is widely used to process audio and video data of variable duration on three modal data. TPP is inspired by the Spatial Pyramid Pooling (SPP) and the difference between them is that SPP performs pooling operation on the spatial dimension, while TPP performs pooling operation on the time dimension. As illustrated in Fig. 3, the data input to TPP has the spatial dimension of F and the time dimension of T. In TPP, the data is first divided into different regions according to different scales along the time axis. Then a pooling operation is performed on each region. The final feature is obtained by concatenating the pooling results of each region and as shown in the Fig. 3, its dimension is $N \times F$, where N depends on the number of pooling regions. Through TPP, we can get data with fixed dimensions, but performing a pooling operation on the time dimension will discard the time clues which might be important for emotion recognition.

In order to retain as much as possible the time clues of the data, we do not divide the data into one to three regions according to the previous literature [22], but try to expand the number of divided regions. We can consider a limiting situation, that is, the number of divided regions is equal to the dimensionality of the original data, which is equivalent to not performing a pooling operation on the data, which means that the time clues will not be discarded. Therefore, we assume that the more regions are segmented, the more time clues are retained. Finally, we introduce Large-scale Temporal Pyramid Pooling (LTPP) to fix the time dimension of the data. Specifically, considering the amount of the data and the average dimension of the dataset, we set the maximum number of divided regions as 16 for IF and TF, while 64 for MF.

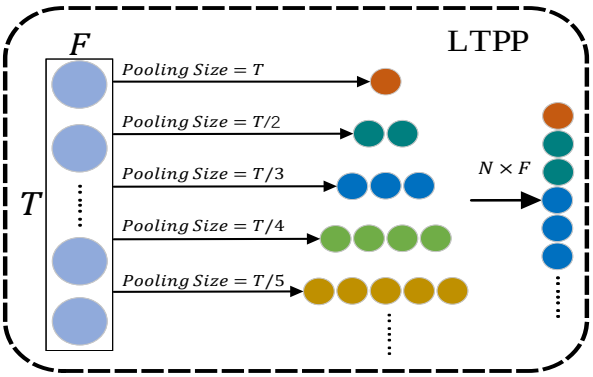

Fig. 3: Illustration of Large-scale Temporal Pyramid Pooling (LTPP). The dimension of the input feature is $F \times T$, and the dimension of the output feature is $N \times F$, where N depends on the level of the pyramid pooling.

Through the above data preprocessing steps, data with fixed dimensions is obtained, in which the video data, audio data and textual data are denoted as I, M and T respectively.

C. Model Configuration

In the data preprocessing phase, in order to fix the data dimension and retain the time clues as much as possible, LTPP is used. Specifically, for audio data, the pyramid level is $l = 0, 2, 4, ..., 64$; for visual data, the pyramid level is $l = 0, 2, 4, 8, 16$; for textual data, the pyramid level is $l = 0, 2, 4, 8, 16$. Then mean pooling is performed on each level. Furthermore, a global maximum pooling is used to ensure the data dimension is even, while retaining the most prominent information in the original data. Finally, in experiments, the dimensions of the preprocessed audio data, visual data, and textual data are $\mathbb{R}^{64\times128\times3}$, $\mathbb{R}^{32\times2048}$, and $\mathbb{R}^{32\times300}$, respectively. (For a small amount of data whose time dimension is less than the pyramid level, zero padding is used to make their dimensions meet the requirements). By designing encoders with different architectures, the encoders will output features of uniform dimension denoted as $EM \in \mathbb{R}^{32\times300}, EI \in \mathbb{R}^{32\times300}, ET \in \mathbb{R}^{32\times300}$. As depicted in Fig. 1, the outputs of encoders (EM, EI and ET) are fed into CMHA module. In each attention module of CMHA, there are three trainable matrices with dimension of 300×300 which are used for linear projection, in this process, the dimensions of features are not be changed. Subsequently, the concatenation operation changes the dimension of the features to 32×600, and the convolution operation changes it back to 32×300. Since there are three submodules of CMHA: CMHA-*MTI*, CMHA-*MIT* and CMHA-*TIM*, they all output a feature with a dimension of 32×300, through the concatenation operation and convolutional operation, these three features are then fused into a feature with a dimension of 32×300. Finally, the fusion feature is fed into three Decoder and the outputs of Decoders are used to construct reconstruction loss.

Audio encoder and visual encoder are stacked from 5 blocks and 2 blocks, respectively, while the textual encoder is stacked from 5 Multi-head attention layers. In each block, the Conv actually contains a convolutional layer, a batch normalization layer, an activation layer and a dropout layer. We set the parameters of the batch normalization layer $momentum = 0.9$, $epsilon = 1e-5$. Instead of adding a pooling operation in the block, the stride of the first convolution layer is set to 2 to achieve the down-sampling which can reduce computational consumption and increase the receptive field.

In the training and validation phase, we perform the proposed model on the Pytorch1.4.0 framework with an NVIDIA GTX 1080Ti GPU. The model is trained with mini-batch size of 64, Adam optimizer with a learning rate of 0.001. To avoid overfitting, L2 regularization with weight of 0.0001 and dropout with drop rate of 0.5 are used. The maximum number of epochs is set as 200 and validation is performed after every epoch on the training set. If the accuracy on the validation set is not improved, the trained model will not be saved. The step by step training is performed on the proposed model. Firstly, the MCWSA is trained in an unsupervised manner and the weights learned in the first stage are used in the prediction part. Finally, the prediction model is fine-tuned on the training set.

D. Evaluation Methods

Since the datasets IEMOCAP and MSP-IMPROV used in this work do not provide a standard test set and the available data samples are limited, we utilize the K-fold cross-validation validation strategy commonly used in the literature. Specifically, following the setting in the previous literature [41], we set K of the K-fold cross-validation strategy to 10, which means that the dataset is divided into ten folds and one of them is selected as the validation set during each training, eight folds are used for training and the remaining fold is used for testing. The final matrix is the average of the results obtained from these ten different test sets.

It should be noted that the datasets used in this work are unbalanced, the number of happy and neutral samples is far more than the number of angry and sad samples, so apart from the unweighted average which is widely used in speech emotion recognition task is adopted as our evaluator, we also adopted the F1-score measurement method commonly used in the previous literatures [47] [56].

V. EXPERIMENTS AND ANALYSIS

A. Comparison with Previous Study

In this section, the model proposed in this work is compared with previous models to demonstrate the advanced performance of the proposed model. Table I lists the results obtained by five existing methods and the methods proposed in this work on the IEMOCAP and MSP-IMPROV datasets. Take the IEMOCAP dataset as an exmple, Gideon and colleagues used progressive neural network and transfer learning to improve the SER performance, and achieved recognition accuracy of 65.7% [57]. Latif *et al*. [41] used a multi-task learning framework based on VAE and obtained a classification accuracy of 68.8%. Li *et al*. [47] proposed global context-aware LSTM model and obtained a classification accuracy of 79.2%. Tsai *et al*. [15] and Dai *et al*. [58] improved multimoal fusion model

and obtained accuracies of 74.7% and 82.7%, respectively. As can be seen from Table I, the proposed MCWSA-CMHA model achieves the best result of 86.3% among the six different methods.

TABLE I: Comparision of unweighted accuracy (UA %) of the proposed method with the previous works on the IEMOCAP and MSP-IMPROV dataset

Model	IEMOCAP	MSP-IMPROV
ProgNet [57]	65.7	60.5
Semi-supervised AAE [41]	68.8	63.6
MulT [15]	74.7	-
GCA-LSTM [47]	79.2	-
MTEM [58]	82.7	-
MCWSA-CMHA (Proposed)	**86.3**	**71.8**

Although the proposed model achieves the best performance on both commonly used datasets compared to previous methods, these still need to be interpreted with the necessary care. In this work, we carry out some special processing on the original datasets, including using the pre-training network to extract the features of the visual modality, and using the LTPP to fix the time dimension. Therefore, in order to compare the proposed model with the previous methods more fairly, we combine the previous methods (such as Concat [8], TFN [10] and MulT [15], etc.) into the basic model MCWSA, and then use the same features and measures to compare the performance of these models. Specifically, the fusion module CMHA of the proposed model is replaced by other previous fusion methods to measure the effectiveness of the fusion module. We implement these schemes and present the results in Table II. A, V and T in this table represent audio, video and textual modalities, respectively. NP represents that the MCWSA model is not pre-trained.

First of all, it can be observed that no matter which fusion method is used, the recognition accuracy is improved to a certain extent by using more modalities. This demonstrates that modal fusion is conducive to emotion recognition.

Secondly, through the horizontal comparison of the data in the table, it can be observed that the accuracy of using video modality (A+V, V+T) is higher than that of not using video modality (A+T) in the case of only using two modalities. We analyze that this is due to the fact that the ResNet network pre-trained on the large-scale ImageNet dataset is used to extract discriminative affective features from video modal data, which is also consistent with previous research finding [22]. In order to further explore the impact of using pre-trained networks to extract video modal features on model performance, two different pre-trained networks (ResNet [17] and VGG-Face [59]) are used to conduct experiments. The experimental result is shown in Table III and it is easy to see that there are differences in model performance corresponding to different video modal features extracted from different pre-training networks. Furthermore, the classification accuracy based on ResNet is higher than that based on VGG-Face. Therefore, in the subsequent experiments, the video modal features we used are extracted by ResNet.

Thirdly, through the comparison of the accuracy of models with and without unsupervised pre-training, we observe that, no matter which feature fusion method is used, the classification accuracy can be improved by performing unsupervised pre-training on models using different fusion methods. This suggests that it is effective to use the multi-channel autoencoder network with noise removal ability in emotion recognition task with noise and heterogeneous data.

Finally, the experimental results in Table II show that the proposed CMHA fusion method gets impressive recognition accuracies in comparison with the state-of-the-art fusion methods. We report an UA of 86.3% on the IEMOCAP dataset, on which outperforms all the three compared methods, i. e., 84.5% by [8], 81.1% by [10] and 84.8% by [15]. Furthermore, although some previous works [15] [60] [61] have proposed modal fusion methods based on attention models, as far as we know, this is an early effort to combine attention mechanism and autoencoder for MER, which can effectively fuse multiple modalities and is scalable.

To further investigate the recognition accuracy, we present the confusion matrix corresponding to the results of the proposed CMHA in Table II. Fig. 4 depicts the confusion matrices of the proposed CMHA model and the MulT model on IEMOCAP and MSP-IMPROV datasets. Fig. 4 (a) shows that the proposed model has an accuracy of 92.0% in identifying sad emotion and the other three emotions are classified with accuracies higher than 80%, on IEMOCAP dataset. We can also observe that the recognition accuracy of each emotion category of the proposed model is higher than that of the MulT model which is presented in Fig. 4 (b) and this suggests that the proposed model has better performance in multimodal fusion which makes emotion classification easier. Fig. 4 (c) and Fig. 4 (d) show the confusion matrix of the proposed model and MulT model on MSP-IMPROV dataset, respectively, and it can be observed that the proposed model is more accurate than MulT model in identifying anger, neutral and sadness. In general, the recognition accuracies of the two models on MSP-IMPROV dataset are lower than that on IEMOCAP dataset. Furthermore, we observe that the neutral is identified with the lowest accuracy which demonstrates the difficulty in recognizing neutral emotion.

In addition to quantitative experiments, the t-SNE algorithm [62] is utilized for qualitative experiments. Fig. 5 presents a visualization of the features obtained by different fusion methods. It can be easily observed that the features obtained by the proposed CMHA have a more compact distribution compared with those obtained by other fusion methods. Furthermore, Fig. 5 suggests that the proposed CMHA maps affective features to a hyperspace where the emotions are better presented, so that different categories of emotions can be distinguished more easily. Note that this is also consistent with the results of the quantitative experiments mentioned above.

B. Effectiveness of MCWSA

Since the emotion data after preprocessing does not have sufficient visibility, an approximate visualization experiment is designed to investigate the effectiveness of the proposed MCWSA method on multimodal fusion. Specifically, the MNIST dataset of handwritten digits is used to replace the

TABLE II: The unweighted accuracy (UA %) and F1 score of three previous multimodal fusion methods and CMHA for category emotion classification on IEMOCAP and MSP-IMPROV datasets.

Dataset	Method	Modality									
		A+V		A+T		V+T		A+V+T (NP[1])		A+V+T	
IEMOCAP		UA (%)	F1	UA (%)	F1	UA (%)	F1	UA (%)	F1	UA (%)	F1
	Concat [8]	73.0	0.731	69.5	0.695	74.6	0.742	82.4	0.826	84.5	0.833
	TFN [10]	-	-	-	-	-	-	80.4	0.806	81.1	0.813
	MulT [15]	75.0	0.754	71.4	0.712	75.9	0.762	83.0	0.832	84.8	0.850
	The proposed CMHA	**78.9**	**0.787**	**74.0**	**0.745**	**78.3**	**0.788**	**85.5**	**0.854**	**86.3**	**0.865**
MSP-IMPROV		UA (%)	F1	UA (%)	F1	UA (%)	F1	UA (%)	F1	UA (%)	F1
	Concat [8]	58.6	0.570	55.0	0.511	58.7	0.555	66.5	0.658	66.8	0.657
	TFN [10]	-	-	-	-	-	-	66.2	0.661	67.4	0.668
	MulT [15]	61.9	0.606	59.4	0.579	60.8	0.594	67.6	0.660	68.6	0.679
	The proposed CMHA	**63.1**	**0.609**	**60.1**	**0.584**	**64.4**	**0.614**	**70.9**	**0.688**	**71.8**	**0.718**

[1] NP represents that the MCWSA model is not pre-trained.

TABLE III: The unweighted accuracy (UA %) based on the video modal features which are extract by pre-trained ResNet and VGG-Face.

Pre-trained Network	Dataset	Modality				
		A+V	A+T	V+T	A+V+T (NP[1])	A+V+T
ResNet [17]	IEMOCAP	78.9	74.0	78.3	85.5	86.3
	MSP-IMPROV	63.1	60.1	64.4	70.9	71.8
VGG-Face [59]	IEMOCAP	76.4	73.2	76.7	83.8	84.5
	MSP-IMPROV	61.8	57.9	61.2	68.5	70.4

[1] The meaning of NP is same as that in Table II.

emotion dataset for the approximate visualization experiment. In order to be as consistent as possible with the settings using multimodal emotion data, the data from 0 to 2, 3 to 5 and 6 to 8 in MNIST dataset are fed into channel 1, channel 2 and channel 3 of MCWSA respectively for training. In short, different images with different content are used to replace the different emotion modalities (in essence, images with different content cannot completely replace multimodal data, but the fusion performance of the model can be indirectly verified by testing the regression performance of the model on images with different content). As shown in Fig. 6, the "Input" on the vertical axis represents the input data of the model, and the "Output" represents the output data of the model corresponding to the "Input", while the channel 1, channel 2, and channel 3 on the horizontal axis represent the different channels of the multi-channel autoencoder. Different numbers are randomly selected and fed into the trained MCWSA model, and then it can be observed from Fig. 6 that each channel of the MCWSA model after training on the training set can relatively fully recover the input data on the test set which indirectly demonstrates that the MCWSA has potential in multimodal feature fusion.

In addition to the visualization experiment based on the MNIST dataset, the t-SNE algorithm is also utilized to verify the validity of the MCWSA model. Fig. 7 presents a visualization of three different features. The first column presents the original features, the second column depicts the fused features of the unsupervised training model, and the last column shows the fused features of the supervised training model. From the first column of Fig. 7, it can be observed that the distribution is chaotic when the original high-dimensional features are mapped to a two-dimensional plane. However, unlike the first column, the second column depicts the distribution of the fused features becomes orderly after reconstruction training. Although there is no obvious separation between the samples of different emotion categories, the samples of four emotion categories are clustered in similar positions on the plane which can potentially reduce the heterogeneity gap. Therefore, the implementation of reconstruction training on the MCWSA model can promote the classification in MER. In addition, the third column of Fig. 7 presents the visualization results of the fused features of the two-stage training model.

C. Effectiveness of CMHA

In this section, we present quantitative and qualitative experiments on the proposed CMHA. Previous works [6] [10] [15] [61] have shown that the interaction between modalities is essential for emotion recognition. We hypothesize that the higher the correlation between the fused features, the better the effect of modal fusion and the better the effect of interactive information mining between modalities. Therefore, we present different experiments which can calculate the correlation of features.

At first, we introduce mutual information which is a measure of the mutual dependence between the two variables to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3144885, IEEE Transactions on Multimedia

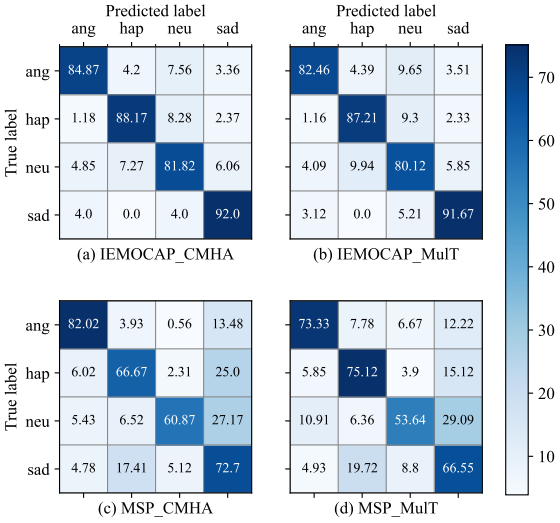

Fig. 4: Comparison of the confusion matrices of the proposed CMHA model and MulT model on IEMOCAP and MSP-IMPROV datasets. Subfigures (a) and (c) are the confusion matrices of the proposed CMHA model. Subfigures (b) and (d) are the confusion matrices of MulT model from [15].

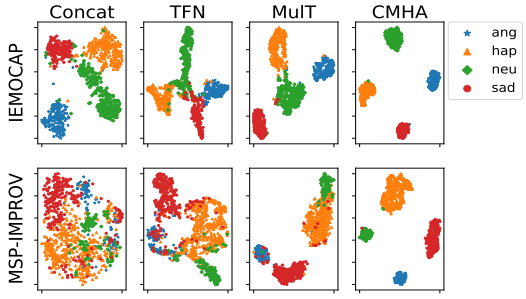

Fig. 5: Fused Feature distribution visualization by the t-SNE algorithm. The fused features of three previous multimodal fusion methods and the proposed CMHA method, on IEMOCAP and MSP-IMPROV datasets, are presented. The different colors stand for different emotions.

verify the effectiveness of the proposed CMHA. The mutual information between two features can be approximated by mutual information neural estimated (MINE) [63]. Fig. 8 shows the mutual information of three different combinations of modal features. The green curve and the red curve represent the mutual information of original features and transformed features, respectively. It should be noted that the transformed features in Fig. 8 refer to the output features of different modal encoders. Since three modalities (video, audio and text) are used in this work, they can be combined in three different ways: video and text, audio and text, audio and video. The mutual information of these three combinations is presented in Fig. 8 (a), Fig. 8 (b) and Fig. 8 (c), respectively. It can be observed that the transformed features have more mutual information than the original features, indicating that

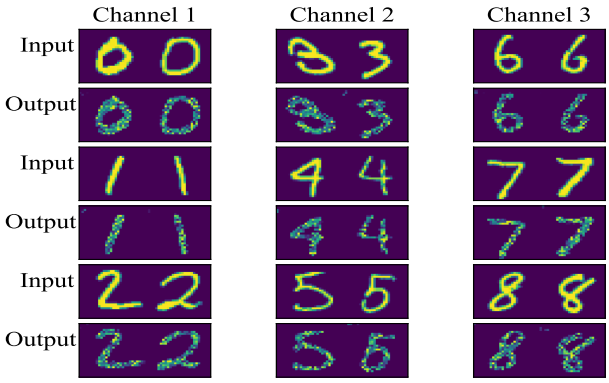

Fig. 6: Visualization experiment of the proposed MCWSA on MNIST dataset of handwritten digits. Different range of numbers are fed into different channels of the MCWSA.

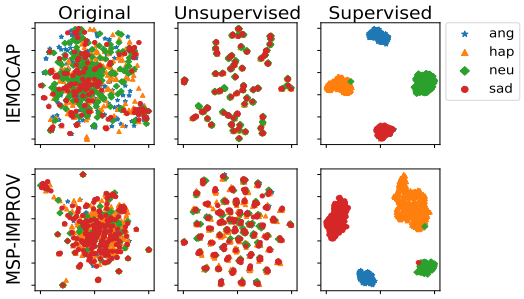

Fig. 7: Visualization of the distribution of different features by t-SNE algorithm. The original features, the fused features of the unsupervised training model, and the fused features of the supervised training model, on IEMOCAP and MSP-IMPROV datasets, are presented. The different colors stand for different emotions.

the correlation between the transformed features of different modalities is stronger, which is more conducive to emotion classification. In fact, the results in Fig. 8 also demonstrates the effectiveness of the proposed CMHA model in modal fusion.

Second, DCCA [30] is used to analyze the correlation between features obtained by different fusion methods. In order to ensure the fairness of the experiment, we keep the dimensions of the features input to the DCCA network consistent, while ensuring that the same DCCA network and the same training hyperparameters are used in the experiment. The experimental results are shown in Fig. 9, we can observe that the correlation of features generated based on the proposed CMHA method is similar to that generated based on the MulT fusion method, and significantly higher than that generated based on the TFN and CONCAT fusion methods which support the quantitative experimental results shown in Table II.

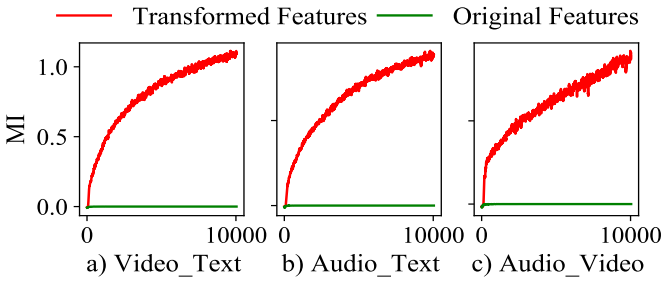

Fig. 8: Mutual information estimated by MINE. The green curve shows the estimated mutual information (MI) of the original features. The red curve represents the MI of the transformed features. The horizontal axis represents the training steps of the DNN used to estimate MI, and the vertical axis represents the estimated MI. Moving average smoothing is used to smooth the curves.

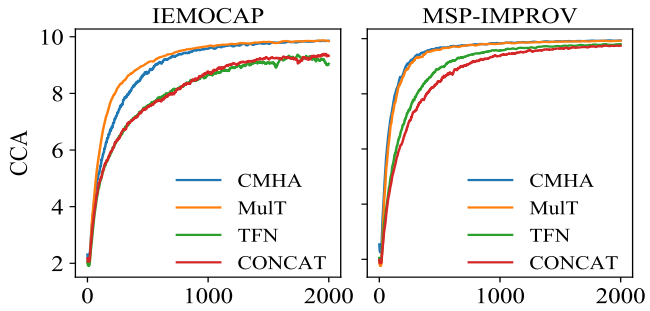

Fig. 9: Correlation between features calculated by DCCA. Different colored curves represent models using different fusion methods. Moving average smoothing is used to smooth the curves.

VI. CONCLUSION

This paper is motivated by how to combine bimodal deep autoencoder [12] and multi-head attention [48] for multimodal emotion recognition. In contrast with previous work focusing on utilizing bimodal autoencoder based on Boltzmann machine, this paper proposes a novel Multi-channel Weight-sharing Autoencoder model for multimodal fusion task, and uses an encoder network based on convolution to improve network feature extraction capabilities. Furthermore, unlike the previous work concatenating the latent code of autoencoder directly, we present a scalable feature fusion method based on multi-head attention mechanism, namely CMHA, to fuse the output features of multiple encoders. We evaluate our proposed model using the popular IEMOCAP and MSP-IMPROV emotion datasets and demonstrated that it yields promising performance in comparison with the state-of-the-arts. The proposed approach can overcome the problem of high heterogeneity among different modalities and has outstanding extensibility which is a significant contribution to the emerging multimodal tasks.

Our analysis shows that: 1) MCWSA has excellent encoding and decoding performance on multimodal emotion datasets; 2) through unsupervised pre-training, MCWSA can make the hyperspace of data orderly, and remove noise; 3) the features transformed by the CMHA fusion method have higher mutual information, indicating that through the transformation of CMHA, the correlation between features increases; 4) compared with the other fusion methods, the features transformed by the proposed CMHA can get a higher canonical correlation, which suggests the superiority of the fusion method.

Due to the lack of datasets with emotion labels and the existence of a large amount of unlabeled audio and video data on the Internet, future work will pay more attention to how to use unlabeled data to further assist emotion recognition.

REFERENCES

[1] J. Tao and T. Tan, *Affective information processing*. Springer London, 2009.
[2] P. Sarkar and A. Etemad, "Self-supervised ecg representation learning for emotion recognition," *IEEE Transactions on Affective Computing*, 2020.
[3] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson, "Detecting anger in automated voice portal dialogs," in *Ninth International Conference on Spoken Language Processing*, 2006, pp. 1053–1056.
[4] Z. Huang, J. Epps, and D. Joachim, "Speech landmark bigrams for depression detection from naturalistic smartphone speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 5856–5860.
[5] M.-J. Han, C.-H. Lin, and K.-T. Song, "Robotic emotional expression generation based on mood transition and personality model," *IEEE transactions on cybernetics*, vol. 43, no. 4, pp. 1290–1303, 2012.
[6] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, "Memory fusion network for multi-view sequential learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[7] G. Sahu and O. Vechtomova, "Dynamic fusion for multimodal data," *arXiv preprint arXiv:1911.03821*, 2019.
[8] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.
[9] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*, 2016, pp. 439–448.
[10] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.
[11] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," *arXiv preprint arXiv:1806.00064*, 2018.
[12] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 689–696.
[13] H. Lee, C. Ekanadham, and A. Ng, "Sparse deep belief net model for visual area v2," in *Advances in Neural Information Processing Systems*, 2008, pp. 873–880.
[14] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[15] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019, pp. 6558–6569.
[16] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[19] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," *arXiv preprint arXiv:1904.03833*, 2019.

[20] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[22] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, 2017.

[23] H. Zhou, D. Meng, Y. Zhang, X. Peng, J. Du, K. Wang, and Y. Qiao, "Exploring emotion features and fusion strategies for audio-video emotion recognition," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 562–566.

[24] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," *arXiv preprint arXiv:1709.03842*, 2017.

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[26] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, 2016.

[27] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proceedings of the 24th International Conference on Artificial Intelligence*, 2015, pp. 2291–2297.

[28] C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 721–732.

[29] H. Hotelling, *Relations Between Two Sets of Variates*. New York, NY: Springer New York, 1992, pp. 162–190.

[30] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, 2013, pp. 1247–1255.

[31] S. H. Dumpala, I. Sheikh, R. Chakraborty, and S. K. Kopparapu, "Audio-visual fusion for sentiment classification using cross-modal autoencoder," in *Proc. Neural Inf. Process. Syst.(NIPS)*, 2019, pp. 1–4.

[32] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep canonical correlation analysis," *arXiv preprint arXiv:1908.05349*, 2019.

[33] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, and L.-P. Morency, "Multi-attention recurrent network for human communication comprehension," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018.

[34] Y. Peng and J. Qi, "Cm-gans: Cross-modal generative adversarial networks for common representation learning," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 15, no. 1, pp. 1–24, 2019.

[35] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 511–516.

[36] J. Deng, X. Xu, Z. Zhang, S. Frühholz, and B. Schuller, "Semisupervised autoencoders for speech emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 31–43, 2018.

[37] Z. Du, S. Wu, D. Huang, W. Li, and Y. Wang, "Spatio-temporal encoder-decoder fully convolutional network for video-based dimensional emotion recognition," *IEEE Transactions on Affective Computing*, 2019, doi:10.1109/TAFFC.2019.2940224.

[38] J. Liang, R. Li, and Q. Jin, "Semi-supervised multi-modal emotion recognition with cross-modal distribution matching," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2852–2861.

[39] S. Latif, R. Rana, J. Qadir, and J. Epps, "Variational autoencoders for learning latent representations of speech emotion: A preliminary study," *arXiv preprint arXiv:1712.08708*, 2017.

[40] Y. Xiao, H. Zhao, and T. Li, "Learning class-aligned and generalized domain-invariant representations for speech emotion recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 4, pp. 480–489, 2020.

[41] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-task semi-supervised adversarial autoencoding for speech emotion recognition," *IEEE Transactions on Affective Computing*, 2020, doi:10.1109/TAFFC.2020.2983669.

[42] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[43] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[44] Q. Chu, W. Ouyang, H. Li, X. Wang, B. Liu, and N. Yu, "Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4836–4845.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[46] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[47] R. Li, Z. Wu, J. Jia, Y. Bu, S. Zhao, and H. Meng, "Towards discriminative representation learning for speech emotion recognition," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 2019, pp. 5060–5066.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[49] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, 2016.

[50] T. Tan, Y. Qian, H. Hu, Y. Zhou, W. Ding, and K. Yu, "Adaptive very deep convolutional residual network for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1393–1405, 2018.

[51] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[52] M. Giatsoglou, M. G. Vozalis, K. Diamantaras, A. Vakali, G. Sarigiannidis, and K. C. Chatzisavvas, "Sentiment analysis leveraging emotions and word embeddings," *Expert Systems with Applications*, vol. 69, pp. 214–224, 2017.

[53] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[54] Z. Yu, X. Xu, X. Chen, and D. Yang, "Temporal pyramid pooling convolutional neural network for cover song identification," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 2019, pp. 4846–4852.

[55] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen, "Temporal pyramid pooling-based convolutional neural network for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2613–2622, 2016.

[56] H. Pham, P. P. Liang, T. Manzini, L.-P. Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6892–6899.

[57] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *arXiv preprint arXiv:1706.03256*, 2017.

[58] W. Dai, Z. Liu, T. Yu, and P. Fung, "Modality-transferable emotion embeddings for low-resource multimodal emotion recognition," *arXiv preprint arXiv:2009.09629*, 2020.

[59] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition." British Machine Vision Association, 2015, pp. 1–12.

[60] J. Huang, J. Tao, B. Liu, Z. Lian, and M. Niu, "Multimodal transformer fusion for continuous emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 3507–3511.

[61] Z. Wang, Z. Wan, and X. Wan, "Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis," in *Proceedings of The Web Conference 2020*, 2020, pp. 2514–2520.

[62] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[63] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. PMLR, 2018, pp. 531–540.

Jiahao Zheng received his B.E. degree from Northeastern University, Shenyang, China, in 2019. He is currently working toward the M.S. degree in systems engineering from the Huazhong University of Science and Technology, Wuhan, China.

His current research interests include DNN-based emotion computing, natural language processing, video understanding, and chaos-based image encryption and secure communication.

Sen Zhang received his B.Sc. degree from Zhengzhou University of Light Industry in 2015 and received the M.S. degree in school of Physics and Opotoelectric Engineering from Xiangtan University, Xiangtan, China, in 2019. He is currently working toward the Ph.D. degree at the School of Automation, Huazhong University of Science and Technology, Wuhan, China.

His current research interests include memristive systems and circuits, chaos and fractional-order chaotic systems and circuits, and design and implementation of brain-like intelligent computing circuit based on memristor.

Zilu Wang (M'21) received the Ph.D. degree in Control Science and Engineering from Huazhong University of Science and Technology, Wuhan, China, in 2021, received the B.E. degree in Automation from Central South University, Changsha, China, in 2016. She is currently a postdoctoral fellow in Department of Computer Science and Engineering, Southern University of Science and Technology. Her current research interests include memristive brain-inspired circuit design, neuromorphic computing, and evolutionary neural network.

Xiaoping Wang (SM'19) received the B.S. and M.S. degrees in automation from Chongqing University, Chongqing, China, in 1997 and 2000, respectively, and the Ph.D. degree in systems engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2003.

Since 2011, she has been a Professor at the School of Automation, Huazhong University of Science and Technology. Her current research interests include memristors and its applications to memory storage, modeling, and simulation.

Zhigang Zeng (F'20) received the Ph.D. degree in systems analysis and integration from the Huazhong University of Science and Technology, Wuhan, China, in 2003.

He is currently a Professor at the School of Automation, Huazhong University of Science and Technology, and also at the Key Laboratory of Image Processing and Intelligent Control of the Education Ministry of China, Huazhong University of Science and Technology. He has authored or coauthored more than 100 international journal papers. His current research interests include theory of functional differential equations and differential equations with discontinuous right-hand sides, and their applications to dynamics of neural networks, memristive systems, and control systems.

Dr. Zeng has been a member of the Editorial Board of $Neural\ Networks$ since 2012, Cognitive Computation since 2010, and Applied Soft Computing since 2013. He was an Associate Editor of the $IEEE\ Transactions\ on\ Neural\ Networks$ from 2010 to 2011. He has been an Associate Editor for the $IEEE\ Transactions\ on\ Cybernetics$ since 2014 and the $IEEE\ Transactions\ on\ Fuzzy\ Systems$ since 2016.