Speech Emotion Recognition Via CNN-Transforemr and Multidimensional Attention Mechanism

Xiaoyu Tang, Yixin Lin, Ting Dang, Yuanfang Zhang, Jintao Cheng

***Abstract*—Speech Emotion Recognition (SER) is crucial in human-machine interactions. Mainstream approaches utilize Convolutional Neural Networks or Recurrent Neural Networks to learn local energy feature representations of speech segments from speech information, but struggle with capturing global information such as the duration of energy in speech. Some use Transformers to capture global information, but there is room for improvement in terms of parameter count and performance. Furthermore, existing attention mechanisms focus on spatial or channel dimensions, hindering learning of important temporal information in speech. In this paper, to model local and global information at different levels of granularity in speech and capture temporal, spatial and channel dependencies in speech signals, we propose a Speech Emotion Recognition network based on CNN-Transformer and multi-dimensional attention mechanisms. Specifically, a stack of CNN blocks is dedicated to capturing local information in speech from a time-frequency perspective. In addition, a time-channel-space attention mechanism is used to enhance features across three dimensions. Moreover, we model local and global dependencies of feature sequences using large convolutional kernels with depthwise separable convolutions and lightweight Transformer modules. We evaluate the proposed method on IEMOCAP and Emo-DB datasets and show our approach significantly improves the performance over the state-of-the-art methods[1].**

***Index Terms*—Speech emotion recognition, temporal-channel-spatial attention, lightweight convolution transformer, local global feature fusion.**

I. Introduction

EMOTION recognition has significant importance in various fields, especially in increasingly common human-computer interaction systems [1], and speech emotion recognition (SER) has promising applications in areas such as mental health monitoring [2], educational assistance, personalized content recommendation, and customer service quality monitoring. Speech contains rich emotional information, and as one of the most basic human communication methods, speech emotion recognition is particularly important for computers to analyze and respond to the emotional state of human users and respond to them accordingly. With the rapid development of artificial intelligence, speech emotion recognition has received extensive research attention. Human speech contains a wealth of information, including not only the language content but also attributes such as gender and emotions of the speaker. It is of great significance to accurately identify emotional information from speech signals.

Feature extraction of speech is a rather important and challenging task in speech emotion recognition, and the extraction of features directly affects the effectiveness of subsequent model training and the accuracy of the final algorithm for emotion recognition. Speech features can be categorized as acoustic-based features and deep learning-based features where acoustic-based features can be broadly classified into rhythmic features [3], spectral-based correlation features [4], and phonetic features [5]. Among them, spectral-based correlation features reflect the characteristics of the signal in the frequency domain, where there are differences in the performance of different emotions in the frequency domain. Based on the spectral correlation features include linear spectrum [6] and inverse spectrum [7], Linear Prediction Cofficients (LPC), Log Frequency Power Coefficients (LFPC), etc.; Inverse spectrum includes Mel-Frequency Cepstrum Coefficients (MFCC), Linear Prediction Cepstrum Cofficients (LPCC), etc. Among them, MFCC is regarded as a low-level feature based on human knowledge, which is widely used in the field of speech.

Early SER algorithms mainly used acoustic-based features and combined with traditional machine learning algorithms, including hidden Markov models [8], Gaussian mixture models [9], and support vector machines [10]. In recent years, deep learning-based neural networks have gradually become active in the field of speech emotion recognition [11], [12], and compared with traditional models, deep learning-based models have shown better performance in speech emotion recognition. Deep learning-based features use neural networks to learn more advanced features from the original signal of speech or some low-dimensional acoustic features. Convolutional neural networks (CNNs) are effective in capturing local acoustic details in speech, while long short-term memory networks (LSTMs) are widely used in speech emotion recognition for modeling dynamic information and temporal dependencies in speech. Additionally, attention mechanisms are also a key

Corresponding author: Xiaoyu Tang. E-mail address: tangxy@scnu.edu.cn.

Xiaoyu Tang is with the School of Electronic and Information Engineering, Faculty of Engineering, South China Normal University, Foshan, Guangdong 528225, China, and also with the School of Physics and Telecommunications Engineering, South China Normal University, Guangzhou, Guangdong 510000, China.

Yixin Lin is with the School of Electronic and Information Engineering, Faculty of Engineering, South China Normal University, Foshan, Guangdong 528225, China.

Ting Dang is with the Nokia Bell Labs, Cambridge, UK.

Yuanfang Zhang is with the Autocity (Shenzhen) Autonomous Driving Co.,ltd.

Jintao Cheng is with the School of Physics and Telecommunications Engineering, South China Normal University, Guangzhou, Guangdong 510000, China.

[1]Our code is available on https://github.com/SCNU-RISLAB/CNN-Transforemr-and-Multidimensional-Attention-Mechanism

factor in improving model performance, as they can adaptively focus on the importance of different features to obtain better speech features at the discourse level. For example, Qi *et al.* [13] proposed a hierarchical network based on static and dynamic features, which uses LSTM to encode dynamic and static features of speech and designs a gating model to fuse the features, an attention mechanism is used to acquire the discourse-level speech features. Liu *et al.* [14] proposed a local-global perceptual depth representation learning system. One module contains a multiscale CNN and a time-frequency CNN (TFCNN) to learn the local representation, and in the other module, a Capsule network with an improved routing algorithm is utilized to design a multi-block dense connection structure, which can learn both shallow and deep global information.

Although speech emotion recognition (SER) models composed of CNNs exhibit better performance than traditional models, these networks can only extract local information in speech, such as the energy and rhythm of a particular segment of speech, while struggling to learn global information in speech features, such as the overall volume and speaking rate of the speaker, and the duration of energy, thus neglecting the global correlation of features [15]. In recent years, the transformer [16] based on the self-attentive mechanism has been widely used in major deep learning tasks. Tarantino *et al.* [17] used transformer in combination with global windowing for speech emotion recognition and achieved better performance, but transformer is weak for local feature extraction. Some recent work has attempted to combine CNN and transformer to alleviate the limitations of using CNN and transformer alone. Wang *et al.* [18] stacked the transformer blocks after the CNN model to improve the global features of aggregation. A model combining the transformer and CNN is proposed in [19], enabling it to learn local information while capturing global dependencies. The original transformer tends to have a high number of parameters for computing multi-headed self-attention, which requires a lot of resources and poses some difficulties in the training of the network. In addition, this kind of work tends to stack transformers in the last part of the model or a simple combination of transformer and CNN, which makes it hard to obtain better local information of the speech.

In addition, attention mechanisms improve the effectiveness of task processing by selectively attending to features that are most relevant to the current task, and have received widespread attention in major fields. In recent years, several researchers have utilized deep learning methods for feature extraction and used attention mechanisms to improve performance. A lightweight self-attention module is proposed in [20], which uses MLP to extract channel information and a large perceptual field extended CNN to extract spatial contextual information. Guo *et al.* [21] proposed an attention mechanism based on time, frequency, and CNN channels to improve representation learning ability. However, temporal information is often embedded in speech, which reflects the dynamic changes of speech, such as pitch and energy variations over time. Temporal features can reflect the temporal context and evolution of emotion expression in speech. However, it falls short in capturing the temporal information present in speech, which represents the dynamics of speech and plays a crucial role in emotion recognition. This limitation is a common issue in both MLPs and CNNs.

In this paper, we investigate how to effectively combine transformer and CNN and apply them to SER to characterize local features and global features in speech signals, and propose the temporal-channel-space attention mechanism in the model for multiple dimensions of feature enhancement. Specifically, we first use a set of stacked CNNs to capture local information in speech and learn shallow features of speech for the transformer module for better training of the transformer module. In the stacked CNN module, two sets of convolutional filters of different shapes are used to capture both temporal and frequency domain contextual information. Specifically, after stacking the CNN modules, we introduced a temporal-channel-space attention mechanism that models the contextual emotional expression of features over time, and efficiently fuses the attention of the spatial and channel dimensions of the speech feature map through the Shuffle unit. Furthermore, a combination of transformer and CNN is used to model the local and global dependencies of feature sequences by a deep separable convolution with residuals and a lightweight transformer module. The main contributions of this work are summarized as follows:

- A framework based on CNN and transformer is proposed for speech emotion recognition. Our framework uses time-frequency domain convolution and stacked convolution blocks to extract initial local features of speech and stacked CNN and transformer blocks are used to enhance local and global features.
- To enhance the finiteness of the feature map and model the temporal information of speech, a temporal-channel-space attention mechanism called Time-Shuffle Attention (T-Sa) is used in our model. T-Sa enhances the feature map in multi-dimensions.
- We propose a module based on deeply separable convolution and a lightweight transformer called Lightweight convolution transformer(LCT). This model employs lightweight convolutional blocks to efficiently extract local information from features, and incorporates Coordinate Attention (CA) into the multi-head self-attention mechanism to capture long-range dependencies among features while enhancing their temporal and spectral information.
- Extensive experiments of our proposed model on IEMOCAP and EMO-DB datasets demonstrate the effectiveness of the model in SER tasks.

The rest of the paper is organized as follows. Section II briefly reviews related work. Details of the system are presented in Section II. In Section IV,we present experimental results to showcase the effectiveness of our model on two widely-used benchmark datasets. Section V concludes this work.

II. RELATED WORK

In this section, we will briefly review the algorithms related to speech emotion recognition, namely convolutional and

recurrent neural networks, attention mechanism, and transformer.

A. Convolutional neural networks and recurrent neural networks

Speech is a continuous time-series signal, and CNN and RNN have been two main network structures for SER. Motivated by the studies of CNN in computer vision, AlexNet [22] and ResNet [23] show promising results in image classification tasks and therefore have been studies in SER. Zhu *et al.* [24] has proposed a new Global Aware Multi-scale (GLAM) neural network that utilizes a global perception fusion module to learn multi-scale feature representations, with a focus on emotional information. The multi-time-scale (MTS) method was introduced in [25], which extends the CNNs by scaling and resampling the convolutional kernel along the time axis to increase temporal flexibility. Liu *et al.* [14] proposed a local global-aware deep representation learning system that uses CNNs and Capsule Networks to learn local and deep global information.

RNN can model the temporal information in speech and capture long-term dependencies in the speech signal more effectively. A new layered network HNSD was proposed [13] that can efficiently integrate static and dynamic features of SER, which uses LSTM to encode static and dynamic features and gated multi-features unit (GMU) for frame level feature fusion for the emotional intermediate representation. Xu *et al.* [26] proposed a hierarchical grained and feature model (HGFM) that uses recurrent neural networks to process both discourse-level and frame-level information of the speech. Since convolutional neural networks can capture local information of features, while recurrent neural networks take advantage of modeling temporal information, many works have combined these two approaches and achieved outstanding results. Li [27] extracted location information from MFCC features and VGGish features by bi-direction long short time memory (BiLSTM) neural network, and then fused these two features to predict emotions. Liu *et al.* [28] combined triplet loss and CNN-LSTM models to obtain more discriminative sentiment information, and the proposed framework yielded excellent results in experiments. Zou [29] *et al.* used CNN, BiLSTM, and wav2vec2 to extract different levels of speech information, including MFCC, spectrogram, and acoustic information, and fused these three features by an attention mechanism. Zhang [30] *et al.* used CNNs to learn segment-level features in spectrograms, using a deep LSTM model to capture temporal dependencies between speech segments.

B. Attention mechanism

In recent years, attention mechanisms have received a lot of attention in major fields to improve the effectiveness of task processing by focusing on information that is more critical to the current task among the many inputs. A channel attention mechanism called Squeeze-and-Excitation (SE) was proposed in [31], which assigns weights to individual channels and adaptively recalibrates the feature responses of the channels. Woo *et al.* [32] proposed a convolutional block attention module that combines both spatial and channel dimensions to obtain attention with better results. In addition, some researchers have used deep learning methods for feature extraction of speech and enhancement of feature maps using attention mechanisms. An attention pooling-based approach was proposed in [33], compared to existing average and maximum pooling, it can combine both class-agnostic bottom-up attention maps and class-specific top-down attention maps in an effective manner. Mustaqeem *et al.* [20] proposed a self-attentive module (SAM) for SER systems,which uses a multilayer perceptron (MLP) to recognize global information of the channels and identifies spatial information using a special dilated CNN to generate an attentional map for both channels. SAM significantly reduces the computational and parametric overhead. A spectro-temporal-channel (STC) attention mechanism was proposed in [21], which acquires attention feature maps along three dimensions: time, frequency, and channel. Xi *et al.* [34] employed an attention mechanism based on the time and frequency domain to introduce long-distance contextual information.

The current attention mechanisms typically focus more on spatial or channel information in feature maps, often neglecting the temporal characteristics in speech. However, temporal features in speech are equally important for emotion recognition. Therefore, it is necessary to pay more attention to temporal information in attention mechanisms to better explore and utilize the temporal characteristics in speech signals.

C. Transformer

Transformer has been rapidly developing in the field of natural language processing (NLP) in recent years and has achieved great success. Due to its powerful ability to obtain global information, the transformer has been gradually extended to the fields of speech. An end-to-end speech emotion recognition model [18] was proposed to enhance the global feature representation of the model by using stacked transformer blocks at the end of the model. Hu *et al.* [19] took advantage of multiple models, improved the learning ability of the model by residual BLSTM, and proposed a convolutional neural network and E-transformer module to learn both local and global information. Recently, transformer-based self-supervised methods have also been applied to speech, and some transformer-based models have achieved great success in automatic speech recognition (ASR), including wav2vec [35], VQ-wav2vec [36], and wav2vec2.0 [37]. There is also some work in speech emotion recognition that employs these models for migration learning. A pre-trained wav2vec2.0 model [38] is used as the input to the network and the outputs of multiple network layers of the pre-trained model are combined to produce a richer representation of speech features. Cai *et al.* [39] proposed a multi-task learning (MTL) framework that uses the wav2vec2.0 model for feature extraction and simultaneously training for speech emotion classification and text recognition. Among computer vision tasks, ViT [40] first applied transformer directly to image patch sequences which is groundbreaking in applying transformer structure to computer vision. ViT has a superior structure and reduced computational resource consumption compared to convolutional neural

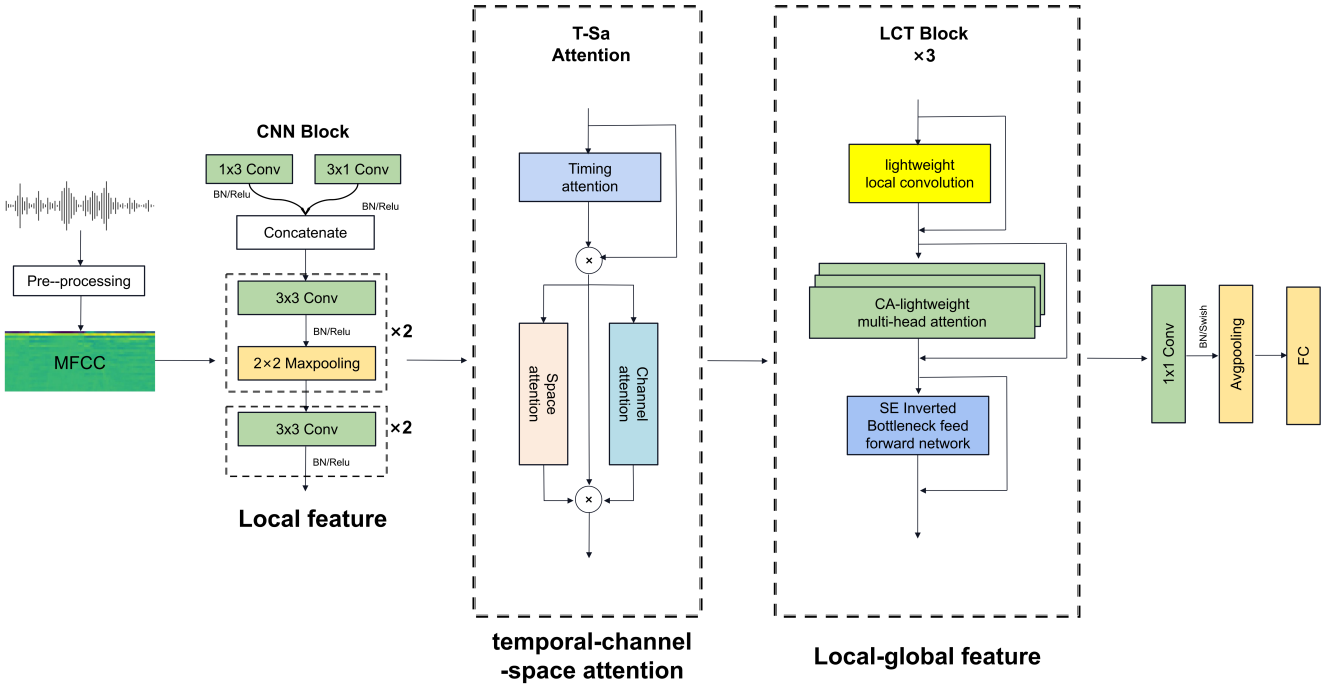

Fig. 1. An illustration of our proposed speech emotion recognition framework consisting of three modules: i) CNN Block is used to extract local information in speech; ii) T-Sa attention module enhances speech information in three dimensions: time-space-channel; iii) LCT Block combines local and global information in speech.

networks. There are many similar approaches in the field of speech. ViT was introduced to speech and improved based on the properties of spectrograms in [41], which proposed a separable transformer (SepTr) that uses the transformer to process tokens at the same time and the same frequency interval, respectively. In [42], a method to improve ViT was applied to infant cry recognition by combining the original log-Mel spectrogram, first-order time-frame and frequency-bin differential log-Mels 3D features into ViT for infant cry recognition.

III. METHOD

In this section, we describe the proposed model in detail. Our proposed model is shown in Fig. 1, which consists of three parts, namely CNN Block, T-Sa attention mechanism module, and LCT Block, these three modules will be introduced in detail next.

A. Overview of the model

To take full advantage of convolutional neural networks and transformer to model the speech sequences and use attention mechanism to enhance the features in time, space and channel, based on which our model is designed. As shown in Fig. 1, for a given input speech sequence, a series of preprocessing steps are performed. Specifically, we uniformly process different lengths of speech sequences into 1.8s, the longer sequences will be cropped into subsegments, and for shorter sequences, we process them using loop filling, after which MFCC features are extracted of speech as the input to the model. The local features of the speech first are obtained by a CNN block, where the irregular-sized time-frequency domain convolution is used to obtain the features in the time and frequency domains of the speech. The features are then enhanceed using a T-Sa attention mechanism block, which contains a bilstm attention module to model the features in the temporal order, followed by a spatial-channel attention mechanism to focus on the spatial and channel information. Finally, the global and local information of speech is learned interactively by an LCT Block, which enables the model to learn information at different scales. The three parts of the model in this paper are described in detail in the following sections.

B. CNN Block

For a large model such as transformer, if MFCC features are directly input into transformer, it will bring a large number of parameters. And since the dataset of speech emotion recognition is generally small, using transformer directly for feature learning will make the model difficult to converge. Therefore, we introduce a CNN Block to pre-learn the local features in speech. As shown in Figure 1, the CNN Block consists of a series of convolutional and pooling layers. For the input feature MFCC, the two dimensions of MFCC correspond to two dimensions in the temporal and frequency domains, respectively, for which we first use a pair of irregular convolutions to obtain the perceptual field in a specific range. For a convolution of size 3×1, we set the perceptual field in the time domain to 1, thus minimizing the effect in the time domain to learn information in the frequency domain, and for a convolution of size 1×3 which is a similar process, which in turn allows capturing a multi-scale representation in

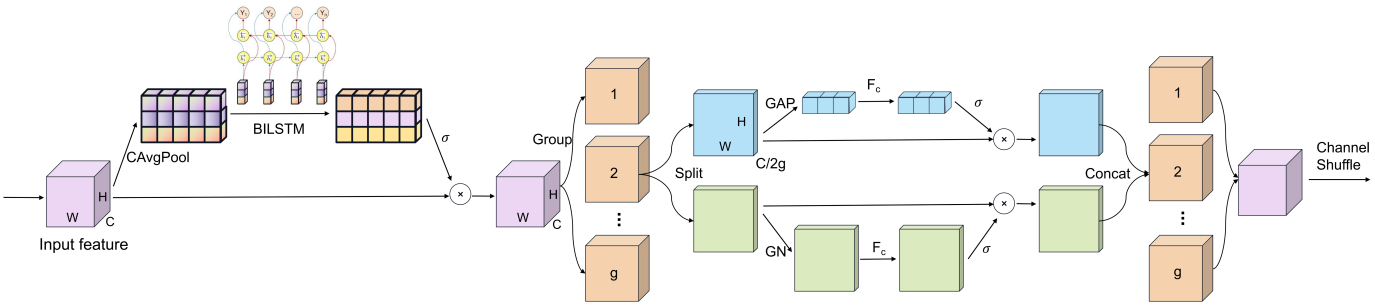

Fig. 2. An illustration of our proposed T-Sa attention module consisting of two modules: Timing attention which enhances the current features temporally through BiLSTM, and Space-channel attention which enhances the features from the spatial and channel dimensions.

the temporal-frequency domain. The results are then fed to successive convolutional layers and maxpooling layers, which are used to further capture the local representation in speech, the batch normalization (BN), and relu activation function are applied after each convolutional layer.

C. T-Sa attention module

In this paper, inspired by [43], we adopt a novel attention module to combine temporal attention with spatial-channel attention, focusing on the temporal dynamics of speech features and spatial-channel information in the feature map, as shown in Fig. 2, which is divided into two parts before and after. In the former part, the attention model enhances the temporal information in the features by Bilstm to model the current features temporally, based on the fact that speech information is temporal information and the order of features in time is of importance. The latter part follows the spatial and channel attention, which is a common concern in existing attention, and efficiently combines the attention of the spatial and channel dimensions of speech feature maps through the Shuffle unit. T-Sa attention module enhances the features in the model through three dimensions and with a small number of parameters and achieves better results in SER.

1) Timing attention: After Pre-processing and CNN Block for feature extraction of speech, giving the feature map size of $X \in R^{C\times H\times W}$ for the input T-Sa attention module, where C, H and W denote the number of channels, spatial height and width, respectively. To model the temporal attention in speech features, a recurrent neural network is used to model the speech information which is bilstm. The input of bilstm is a two-dimensional sequence while our speech feature map is three-dimensional, so what we need is to process the feature map X. If we directly reshape the feature map, the input bilstm will bring a great number of parameters number. Therefore, average pooling is used to reduce the dimensionality of the channels, and the specific calculation is:

$$X^{CAvgPool} = \frac{1}{C}\sum_{i=1}^{C} X(i) \tag{1}$$

After the feature passes through average pooling, the size of the feature map is $X^{CAvgPool} \in R^{H\times W}$. Feature map is adjusted to $X^{CAvgPool} \in R^{W\times H}$ by reshaping operation and then feeding it to the bilstm layer. By encoding long distances from front to back and from back to front, bilstm can better capture bidirectional feature dependencies. $X^{CAvgPool}$ is encoded by bilstm as follows:

$$\overrightarrow{h}^{bilstm} = \overrightarrow{BILSTM}(X^{CAvgPool}) \tag{2}$$

$$\overleftarrow{h}^{bilstm} = \overleftarrow{BILSTM}(X^{CAvgPool}) \tag{3}$$

Two LSTMs in bilstm process the sequence forward and backward, respectively, and then concate the outputs of the two LSTMs together:

$$H^{bilstm} = \text{Concatenate}\left(\overrightarrow{h}^{bilstm}; \overleftarrow{h}^{bilstm}\right) \tag{4}$$

H^{bilstm} is then applied sigmoid activation and multiplied with the input feature map using the residual scheme to output temporal attention:

$$X^{time} = \sigma\left(H^{bilstm}\right) \cdot X \tag{5}$$

2) Space-channel attention: Spatial attention and channel attention are widely used in computer vision. Most methods transform and aggregate features in these two directions, such as SE [31], CBAM [32], BAM [44], GCNet [45], but these methods do not make full use of the correlation between space and channel which are not efficient. Therefore, we adopted SA-Net [43] as our spatial-channel attention module.

Given the output X^{time} of the temporal attention module, the input is first divided into G groups, and the size of each sub-feature map is $X^{time'} \in R^{C/G\times H\times W}$. Then, each group is split into two sub-branches in the channel dimension., $X^{spatial}$ and $X^{channel}$, one of which obtains spatial attention and the other obtains channel attention.

For the channel attention branch, firstly, the global average pooling (GAP) operation is performed on the input of the branch to embed the global information. Then, a simple gating mechanism and sigmoid activation are used to perform adaptive learning of spatial features. A residual scheme is used to multiply the input channel branch feature map. The specific operation of the spatial attention module is as follows:

$$X^{channel'} = \sigma\left(W_1 \cdot GAP(X^{channel}) + b_1\right) \cdot X^{channel} \tag{6}$$

W_1 and b_1 are learnable parameters.

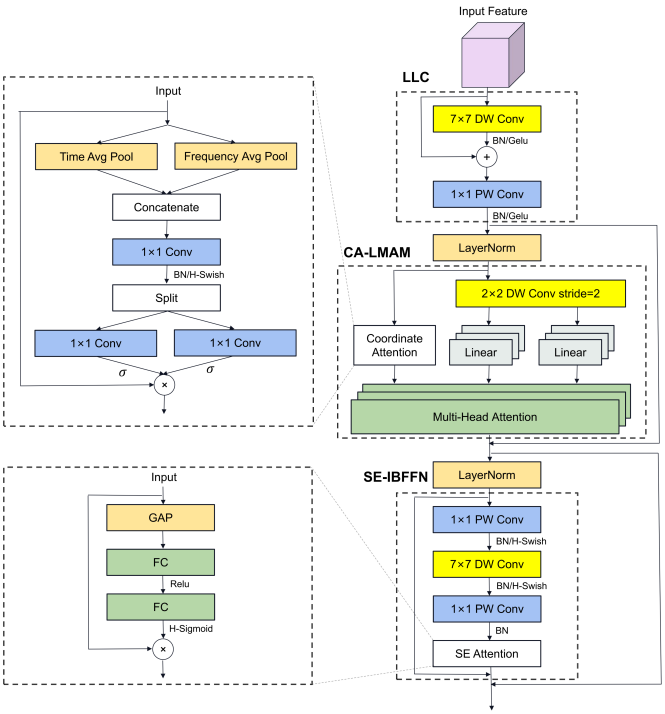

Fig. 3. An illustration of our proposed LCT Block attention module consisting of three modules: i) LLC, which efficiently captures the local information of the features through a lightweight convolution module; ii) CA-LMAM, which captures the long-range dependencies in the features and enhance the time-frequency domain features of speech through a CA module; iii) SE-IBFFN, which introduces a nonlinear part through a feedforward network with inverse residuals to enhance the model the expressiveness of the model.

For the spatial attention branch, GN [46] operation is first performed on the input of the branch to embed spatial statistical information, and then a simple gating mechanism and sigmoid activation are used to perform adaptive learning of features on the channel, and the residual scheme is used to multiply the input channel branch feature map. The specific operation of the channel attention module is as follows:

$$X^{spatial'} = \sigma\left(W_2 \cdot GN(X^{spatial}) + b_2\right) \cdot X^{spatial} \tag{7}$$

Both W_2 and b_2 are learnable parameters, and then the outputs of the two branches are merged by concatenating them along the channel dimension:

$$X^{attention} = \text{Concatenate}\left(X^{channel'}; X^{spatial'}\right) \tag{8}$$

The attention weights on space and channel are learned separately through two branches, and the corresponding residual scheme is multiplied with the respective input feature map to enhance the representations of space and channel. Finally, the channel shuffle [47] is employed to facilitate communication of information between different groups along the channel dimension. The information of the features interacts in the channel dimension, and the size of the output feature map is the same as the initial input.

D. LCT Block

Inspired by [48], we proposed an LCT Block shown in Fig. 3, which consists of three modules: lightweight local convolution (LLC), coordinate attention-lightweight multi-head attention mechanism (CA-LMAM) and SE Inverted Bottleneck feed forward network (SE-IBFFN). Through the combination of convolution and transformer, LCT Block obtains the information of features from the local and the whole, which is more efficient and has fewer parameters than the traditional transformer. LLC is a lightweight convolution module, which efficiently obtains local information of features. The CA-LMAM module can capture the long-distance dependencies of features and enhance the information in the time-frequency domain through the Coordinate Attention (CA) [49] module. SE-IBFFN introduces a nonlinear part through a feedforward network with inverted residuals, which enhances the performance of the model and can further capture the local information of features. These three modules will be introduced in detail below

1) LLC: In order to make up for the lack of local information in trasnformer, a convolution module is used to obtain local information in speech which called LLC. Some work such as CMT [48] also adopted a similar architecture, but the convolution module in CMT is relatively simple, only using a Depthwise convolution with residuals, which can not obtain effective local information. Inspired by [50], we used a large convolution kernel Depthwise convolution with residuals and a Pointwise convolution in the local feature extraction module, given an input feature $X \in R^{C \times H \times W}$ as follows:

$$LLC\left(X\right) = PWConv\left(DWConv\left(X\right) + X\right) \tag{9}$$

The activation layer and batch normalization are omitted, and the same omission will be in subsequent formulas. $PWConv$

represents Pointwise convolution and $DWConv$ represents Depthwise. In $PWConv$ we used a large 7×7 convolution kernel is used to get a larger receptive field than a 3×3 convolution kernel. In addition, $PWConv$ has smaller parameters than $DWConv$ and ordinary convolution. Additionally, a residual structure is incorporated to address the issue of gradient dispersion.

2) CA-LMAM: In transformer, multi-head attention is usually used to make the model pay more attention to the more noteworthy part of itself, which can obtain long-distance dependence in speech features. Given the output $X \in R^{C \times H \times W}$ in the LLC module, the input of multi-head attention is query Q, key K, and value V, respectively. If the original features are directly input into multi-head attention, it will often bring a large amount of calculation. The amount of calculation is related to the size of the input features. Therefore, 2×2 $DWConv$ is used to downsample the parts of K and V, as shown below:

$$K = DWConv\left(X\right) \tag{10}$$

$$V = DWConv\left(X\right) \tag{11}$$

Where $K \in R^{C \times \frac{H}{2} \times \frac{W}{2}}$, $V \in R^{C \times \frac{H}{2} \times \frac{W}{2}}$, then the H and W dimensions are merged and input a linear layer respectively. Finally get $K' \in R^{N' \times C}$, $V' \in R^{N' \times C}$, where $N' = \frac{H}{2} \times \frac{W}{2}$, C is the dimension of the linear layer output.

For Q, a CA module is used to enhance the time-frequency domain representation of speech features. This attention mechanism can obtain long-distance feature dependence along one direction and spatial dependence information in another direction. However, in speech, speech features have more special significance in the spatial dimension. The abscissa of speech features is the time axis, which represents the time domain information of speech while the ordinate of the speech feature represents the frequency information of the speech, so CA can better aggregate the dependent information in the time domain and frequency domain for speech features, which is more suitable for SER tasks.

The specific operation of CA is shown in Fig. 3. Given the input $X \in R^{C \times H \times W}$, pooling is performed in the time domain and frequency domain respectively:

$$X^{TAvgPool} = \frac{1}{W} \sum_{i=1}^{W} X(i) \tag{12}$$

$$X^{FAvgPool} = \frac{1}{H} \sum_{j=1}^{H} X(j) \tag{13}$$

$X^{TAvgPool} \in R^{C \times H \times 1}$ and $X^{FAvgPool} \in R^{C \times 1 \times W}$ are the feature maps after aggregation in two directions of time-frequency domain. Then, they concatenate and perform convolution operations to encode the information:

$$X^{tf} = Conv\left(\text{concatenate}\left(X^{TAvgPool}; X^{FAvgPool}\right)\right) \tag{14}$$

Where $X^{tf} \in R^{C/r \times 1 \times (W+H)}$, r is the reduction rate of the channel, and then separated into two separate features X^t, X^f along the spatial direction, where $X^t \in R^{C/r \times H \times 1}$, $X^f \in R^{C/r \times 1 \times W}$. The extraction of features is then performed by two separate convolutions, followed by activation using the σ activation function, and then multiplied by the initial output to learn the more critical information in the feature:

$$s^t = \sigma\left(Conv\left(X^t\right)\right) \tag{15}$$

$$s^f = \sigma\left(Conv\left(X^f\right)\right) \tag{16}$$

$$Q = X \cdot s^t \cdot s^f \tag{17}$$

We adjust the dimension of Q, eventually $Q' \in R^{N \times C}$, where $N = H \times W$.

Eventually learning information about the model itself through multi-headed self-attention:

$$\text{LMAM}(Q, K, V) = \text{Softmax}\left(\frac{Q'K'^T}{\sqrt{C}} + B\right)V' \tag{18}$$

Where B is a learnable parameter, representing the relative position coding of multi-head self-attention, which is used to characterize the relative position relationship between tokens. It is more flexible than the traditional absolute position coding, making transformer better model the relative position information of speech features.

3) SE-IBFFN: In the original transformer, FFN is generally composed of two linear layers. In this paper, inspired by [51], a series of Pointwise convolution and Depthwise convolution make up our SE-IBFFN. Compared with the Transformer model traditionally composed of linear layers, this module can capture local information while learning channel information, and has a smaller number of convolution parameters than ordinary ones.

The structure of SE-IBFFN is shown in Fig. 3. Given the input $X \in R^{C \times H \times W}$, the specific calculation process is as follows:

$$X^{conv} = PWConv\left(DWConv\left(PWConv\left(X\right)\right)\right) \tag{19}$$

$$SE\text{-}IBFFN\left(X\right) = SE\left(X^{conv}\right) + X \tag{20}$$

Firstly, the Pointwise convolution operation is performed on the input, and the number of channels is expanded to 4 times of the original to increase the feature size of the channels. Then a 7×7 large convolution kernel $DWConv$ is used to obtain the local information in the feature, and the large convolution kernel can provide a larger receptive field without increasing too many parameters. The feature map is then restored to its original size using $PWConv$.

Then, the SE module is to obtain the attention of the channel dimension of the feature map. Given the input $X \in R^{C \times H \times W}$ of SE, the specific calculation process is as follows:

$$SE\left(X\right) = \sigma\left(GAP\left(FC\left(FC\left(X\right)\right)\right)\right) \cdot X \tag{21}$$

Compared with [51], we put the SE module after $PWConv$, which makes the parameters of the model smaller. A residual structure is used to solve the problem of gradient dispersion.

IV. EXPERIMENTS SETUP

In this section, we will introduce the dataset and experimental details used in our study, as well as the evaluation metrics used to assess the performance of our algorithm.

A. Corpus Description

To verify the performance of our proposed algorithm, performance tests on two benchmark databases are conducted, on which we will evaluate our algorithm.

Actually, Interactive Emotional Dyadic Motion Capture (IEMOCAP) [52] is an action, multimodal, and multimodal database that contains data from 10 actors and actresses during an emotional binary interaction, with two speakers (one male and one female) speaking in each session. The IEMOCAP database has been annotated by several annotators with categorical labels and dimensional labels. The database combines discrete and dimensional sentiment models. In our work, the method used improvisational and scripted data, choose anger, happiness, neutral, sadness and excitement as basic emotions, and merge happy and excited into happy. We partitioned the IEMOCAP dataset into training and testing sets by randomly selecting 80% and 20% of the data, respectively.

To valid our method robustness, we tested our method in other datasets. The Berlin Emotional Database (Emo-DB) [53] is a German emotional speech database recorded by the Technical University of Berlin. The database includes recordings of ten actors, comprising of five male and five female, who simulate seven emotions, including neutral, anger, fear, joy, sadness, disgust, and boredom, on ten sentences (five short and five long), resulting in a total of 535 speech recordings (233 male and 302 female). It has high emotional freedom, adopts 16 kHz sampling, and 16-bit quantization, and saves files in WAV format. It is a discrete emotional language database, and the excitation method is performance type.

B. Implementation Details

In the feature generation phase, the method uses the mel-frequency cepstrum coefficients (MFCCs) extracted by 26 Mel filters as the feature input. Meanwhile it divided each input speech into 1.8 seconds of speech segments, and the overlap between segments is 1.6 seconds, which can generate a large number of speech samples to solve the problem of scarcity of data set samples in SER. To obtain the prediction result for a sentence in the test set, we take the average of the prediction results of all speech segments within that sentence. Our model trained a total of 150 epochs, used the cross entropy criterion as the objective function, and used the Adam optimizer. The weight decay rate is 10^{-6}, the learning rate and the mini-batch size are set to 0.001 and 128, respectively, and the multiplication factor 0.95 is exponentially decayed until the value reaches 10^{-6}. Our experiment is carried out on Ubuntu 18.04 with a GeForce RTX 2080ti GPU, and we utilized Pytorch 1.7 as the training framework.

In addition, we use the method of mixup [54] to train, so as to improve the generalization ability of the system. This method constructs new training samples and labels by linear interpolation, and effectively smoothes the discrete data space into continuous space. In our proposed model, we set α to 0.2 for best performance.

C. Evaluation Metrics

In this section, we'll describe in detail the criteria we use to evaluate the performance of our algorithms. For various categories of performance in the dataset, Precision, Recall, and F1-score are the general metrics to measure their performance. First of all, four concepts will be introduced: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), where TP means actual positive and predicted positive, FP means actual positive and predicted negative, TN means actual negative and predicted positive, and FN means actual negative and predicted negative. The Precision, Recall, and F1-score can be expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$Recall = \frac{TP}{TP + FN} \tag{23}$$

$$F1 - score = \frac{precisio \times recall \times 2}{precision + recall} \tag{24}$$

To evaluate the overall performance of our model, weighted average accuracy (WA) and unweighted average accuracy (UA) will be used as evaluation metrics, where WA is the weighted average accuracy of different sentiment categories, and its weight is related to the number of sentiment categories, and UA is the average accuracy of different sentiment categories. The validity of the model is better evaluated in the context of unbalanced SER dataset samples. The calculation methods for these two metrics are as follows:

$$Acc_i = \frac{TP_i}{TP_i + FP_i} \tag{25}$$

$$WA = \frac{\sum_{i=1}^{C} N_i \times Acc_i}{\sum_{i=1}^{C} N_i} \tag{26}$$

$$UA = \frac{1}{C} \sum_{i=1}^{C} Acc_i \tag{27}$$

Among them, C represents the number of emotional categories, and N_i represents the number of samples of class i.

V. RESULTS

In this section, we conduct extensive experiments to evaluate the performance of our method on two datasets. The section mainly compares our proposed model with state-of-the-art baselines, and then verify the effectiveness of our proposed modules through ablation experiments.

A. Comparison with State-of-the-Art

To compare with our proposed model, we evaluate the performance of the algorithm from the WA and UA perspective with a series of existing methods in IEMOCAP and Emo-DB datasets, as shown in Table I and Table II. The proposed model is compared with several commonly used speech emotion recognition models, including algorithms based on the combination of CNN and transformer [19], CNN-based algorithms [14] [55], LSTM-based algorithms [56] [57], attention-based methods [21] [58], and some other algorithms.

TABLE I
COMPARISON ON IEMOCAP

Comparative Methods	WA	UA
Latif et al. [59]	-	68.8
Hu et al. [19]	69.73	70.11
Guo et al. [21]	61.32	60.43
Gao et al. [60]	70.34	70.82
Wang et al. [57]	69.40	69.50
Latif et al. [56]	-	64.10
Liu et al. [14]	70.34	70.78
Dai et al. [61]	65.40	66.90
Proposed	**71.64**	**72.72**

TABLE II
COMPARISON ON EMO-DB

Comparative Methods	WA	UA
Tuncer et al. [62]	90.09	89.47
Li et al. [58]	83.30	82.10
Zhong et al. [63]	85.76	86.12
Kerkeni et al. [64]	-	86.22
Suganya et al. [55]	-	85.62
Proposed	**90.65**	**89.51**

To verify the performance of our proposed method, we compare it with other speech emotion recognition algorithms in IEMOCAP and Emo-DB datasets. The results are shown in Table I and Table II.

For IEMOCAP dataset, as shown in Table I, our proposed method outperforms other methods. In addition, compared with the simple splicing of transformer and CNN [19], our method achieves the best results through the ingenious design of local and global feature extraction. Compared to traditional spatial and channel attention [21], temporal attention information makes it more competitive. In addition, for the CNN or LSTM-based method [14] [57] [56] [61], our method introduces global information through a lightweight transformer module to bring more comprehensive features to the system, and also shows that our transformer module has a stronger ability to obtain global information than some capsule networks. Finally, our method is as competitive as the method using semi-supervised methods [59] and pre-trained models [60].

For Emo-DB dataset, as shown in Table II, our proposed method outperforms several methods. Similar to the performance in IEMOCAP dataset, our attention mechanism and local-global model have significant advantages over traditional attention and CNN-based models [58] [63] [55]. In addition, Emo-DB dataset is a smaller dataset. Compared with non-deep learning feature extraction and feature learning methods [62] [64], our overall end-to-end network based on deep learning also has a relatively better performance on this small dataset, indicating that our method still has excellent robustness on small datasets.

In summary, combining the performance of these two datasets proves the superiority of our proposed method.

B. Results and Analysis

In this section, Table III and Table IV list the Precision, Recall, F1-score, and overall WA and UA for each sentiment category in IEMOCAP dataset and Emo-DB dataset, respectively. In addition, the confusion matrices are visualized of the two datasets in Fig. 4 and Fig. 5, where the diagonal indicates that the sentiment is correctly classified, and other locations indicate that the sentiment is misclassified as other sentiments. The darker the color in the grid while higher the accuracy.

TABLE III
CONFUSION MATRIX OF THE PROPOSED MODEL ON IEMOCAP

Emotion	Precision	Recall	F1-score
Neural	72.88	62.32	67.19
Sad	70.51	75.00	72.68
Angry	74.43	79.13	76.71
Happy	69.68	74.43	71.98
WA		71.64	
UA		72.72	

TABLE IV
CONFUSION MATRIX OF THE PROPOSED MODEL ON EMO-DB

Emotion	Precision	Recall	F1-score
Neural	86.67	100.00	92.86
Sad	94.44	100.00	97.14
Angry	88.00	95.65	91.67
Happy	83.33	83.33	83.33
Boredom	100.00	93.75	96.77
Disgust	100.00	76.92	86.96
Fear	83.33	76.92	80.00
WA		90.65	
UA		89.51	

For IEMOCAP dataset, as shown in Table IV and Fig. 4, our proposed model achieves good results on this dataset and has good accuracy for all four emotions, especially sadness, anger and happiness. Among these four emotions, anger has the highest recognition accuracy, while neutral has the lowest recognition accuracy. Sadness, anger and happiness will be misjudged as neutral in some cases. The result is similar to that in [65], which also verifies that neutral emotion is a defect of expression. This emotion is easily expressed as other emotions, which makes the model misjudged. Therefore, neutral emotions are easily confused with other emotions.

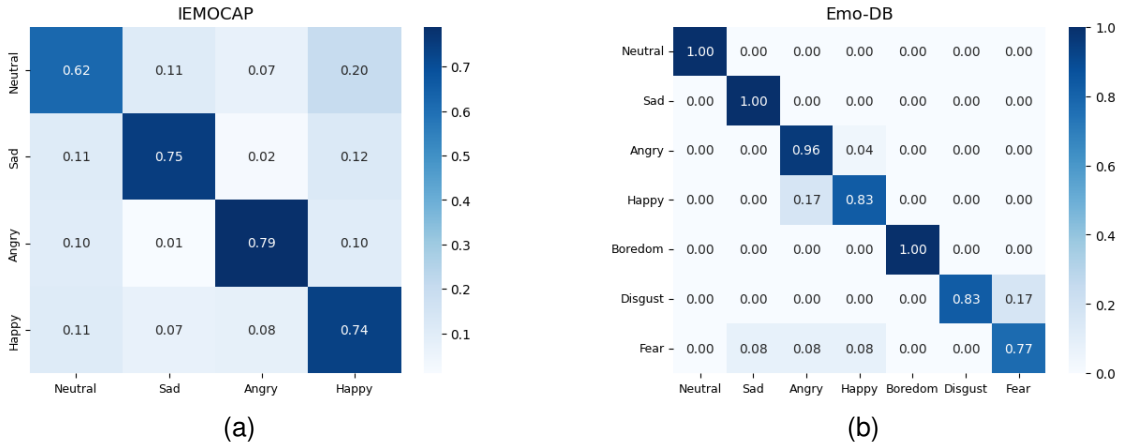

Fig. 4. Visualization the confusion matrices of the proposed method: (a) Confusion matrix on IEMOCAP; (b) Confusion matrix on Emo-DB.

For Emo-DB dataset, as shown in Table IV and Fig. 5, our proposed model also achieved good results on this dataset, and has good accuracy for seven emotions. The three emotions achieved 100% accuracy, and Angry also achieved 96% accuracy. In addition, the wrong judgment in happiness is anger. We believe that this is because these two emotions have similar arousal, while the wrong judgment in disgust is fear. Finally, the accuracy of fear is lower than that of other emotion categories, but it also achieves relatively good results. In some cases, fear is easily misjudged as sad, angry and happy, this is because these four emotions have a high degree of arousal.

TABLE V
PERFORMANCE WITH DIFFERENT EXPERIMENTAL SETTINGS

Models	WA	UA
W/o T-Sa	69.92	71.18
W/o lstm attention	67.84	69.62
W/o LCT	67.12	68.93
W/ Conv-LCT	68.29	69.98
W/o CA	69.29	70.12
W/o SE	69.92	71.31
Proposed	**71.64**	**72.72**

C. Ablation Study

To explore the role of each part of our proposed model, Table V shows the results of a series of ablation experiments. The following are the modules for comparison.

- W/o T-Sa: This module removes the T-Sa module from our model, using only the CNN Block and LCT Block sections.
- W/o lstm attention: This module removes the temporal attention lstm attention part in our T-Sa module, and the other parts are retained.
- W/o LCT: This module removes the LCT module in our model and only uses the CNN Block and T-Sa parts.
- W/ Conv-LCT: This module replaces the LLC module in our LCT module with a 3×3 convolution, and the rest is preserved.
- W/o CA: This module removes the CA portion of our LCT module and preserves the rest.
- W/o SE: This module removes the SE Attention part of our LCT module, and the other parts are retained.

It can be seen from the table that when using the T-Sa module, our method has an absolute improvement of 1.2% and 1.54% in WA and UA, indicating that our attention mechanism module has a very significant effect and can aggregate the noteworthy parts of the features. In addition, when the temporal attention is removed, the model effect is also greatly reduced, indicating that temporal attention plays a non-important role in our model to enhance the temporal information in speech. We also directly removed the LCT module, which caused the performance of the model to be reduced by 4.52% and 3.79% on WA and UA, it clearly shows the importance of introducing global information into our LCT module.

For our LCT part, our experiments also verified the role of different modules within the LCT. Firstly, the LLC part is replaced with a 3×3 ordinary convolution, which reduces the performance of the model by 3.35% and 2.74% on WA and UA, and the number of parameters has also been improved. This shows that the wider receptive field brought by our large convolution kernel LLC module is very important, and it does not bring a larger number of parameters. In order to verify the role of our CA module, we removed the CA module, which caused the performance of the model to decrease by 2.35% and 2.6% on WA and UA, indicating that the time-frequency domain representation of the speech features enhanced by the CA module is of vital importance. Finally, the SE Attention part in our SE-IBFFN is removed, which reduces the performance of the model by 1.72% and 1.41% on WA and UA. It also proves that using the SE module to obtain the attention of the feature map channel dimension can improve the performance of the model to recognize emotions.

D. Model Efficiency Analysis

In order to explore the size and efficiency of the proposed model, Table VI presents a comparison of the parameter

TABLE VI
COMPARISON OF MODEL PARAMETERS AND ACCURACY

Models	Params	WA	UA
W/ Conv-LCT	1,404,922	68.29	69.98
W/ConvFFN-LCT-L	10,390,522	70.46	72.14
W/ConvFFN-LCT-S	2,526,202	68.83	69.71
W/Transformer	1,357,810	69.38	70.43
W/MobileVitv1 block-depth2 [66]	2,036,458	70.01	70.98
W/MobileVitv1 block-depth3 [66]	2,726,506	66.58	67.30
W/MobileVitv2 block-depth2 [67]	937,898	68.29	69.68
W/MobileVitv2 block-depth3 [67]	1,086,634	67.12	69.79
W/MobileVitv3 block-depth2 [68]	**661,162**	69.20	70.14
W/MobileVitv3 block-depth3 [68]	884,010	68.47	69.73
Proposed	1,031,674	**71.64**	**72.72**

counts and accuracy between our model and other models. The following are the modules we designed for comparison.

- W/ Conv-LCT: This module replaces the LLC module in our LCT module with a 3×3 convolution and the other parts are retained.
- w/ ConvFFN-LCT-L: This module replaces the first PW convolution and DW convolution in the SE-IBFFN module of our LCT module with a 7×7 convolution and the other parts are retained.
- W/ ConvFFN-LCT-S: This module replaces the first PW convolution and DW convolution in the SE-IBFFN module of our LCT module with a 3×3 convolution and the other parts are retained.
- W/Transformer: This module replaces CA-LMAM and SE-IBFFN with traditional transformer, which is the same as the setting of LCT.
- W/ MobileVit block: This module replaces our LCT module with the MobileVit block of various series of Mobile Vit, with an input channel size of 128 and a depth of 2 or 3. The intermediate layer size of the MLP is set to 4 times of the input channel size for all versions. For v1 and v2, the number of intermediate layers is 192, and for v3, the number of intermediate layers is 128.

From the Table VI, it can be seen that when the LLC module in LCT is replaced with a regular convolution, the number of parameters increases and the accuracy is lower than that of the LLC module. This indicates that LLC has better accuracy with fewer parameters. Additionally, when the first PW and DW convolutions in SE-IBFFN are replaced with regular convolutions, the number of parameters increases significantly, and the accuracy is still lower than the original model. We also reduced the size of the convolution kernel to 3×3, which improved the number of parameters, but it is still larger than the original SE-IBFFN, and the accuracy decreased. This suggests that the SE-IBFFN module did not bring a huge number of parameters while using a larger convolution kernel to achieve a larger receptive field, and the accuracy is still excellent.

Additionally, we attempted to replace CA-LMAM and SE-IBFFN in LCT with traditional transformers, and experimental results showed that the traditional transformer has a larger number of parameters and a decrease in accuracy compared to LCT, with a decrease of 2.23% and 2.29%, respectively. The entire LCT module was also replaced with other transformer models, including the MobileVit series which combines CNN and Vit and is a lightweight transformer model. The MobileVit block in each series was used to replace LCT, and the results showed that the parameter size of the MobileVitv3 block is smaller than LCT, but there is still a gap in accuracy. The performance of the two-layer MobileVitv1 block was the best, but it still did not reach the accuracy of LCT. Additionally, we found that models with a depth of 2 performed better than those with a depth of 3, due to the difficulty of fitting to small SER datasets as the number of parameters increases.

VI. CONCLUSION

In this study, to better model local and global features of speech signals at different levels of granularity in SER and capture temporal, spatial and channel dependencies in speech signals, we propose a Speech Emotion Recognition network based on CNN-Transformer and multi-dimensional attention mechanisms. The network consists of three modules. First, a CNN block is used to model time-frequency domain information in speech, capturing preliminary local information in speech. Second, we propose a T-Sa network to model the emotional expression context of features over time and efficiently fuse the spatial and channel dimensions of speech feature maps through Shuffle units. Finally, to efficiently fuse local information and long-distance dependencies in speech, we propose an LCT module that uses lightweight convolutional modules and introduces Coordinate Attention into multi-head self-attention. This allows for the fusion of features at different levels of granularity while enhancing information in the time-frequency domain of features without introducing a high number of parameters.

In future work, in addition to MFCC features, we will try more hierarchical speech features and combine current CNN and transformer structures to improve the performance of Speech Emotion Recognition from multiple feature dimensions.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 62001173), the Project of Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds) (Grant No. pdjh2022a0131, pdjh2023b0141).

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[2] L.-S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2010.

[3] A. Mahdhaoui, M. Chetouani, and C. Zong, "Motherese detection based on segmental and supra-segmental features," in *2008 19th International Conference on Pattern Recognition*. IEEE, 2008, pp. 1–4.

[4] S. E. Bou-Ghazale and J. H. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Transactions on speech and audio processing*, vol. 8, no. 4, pp. 429–442, 2000.

[5] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[6] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 80–84, 1997.

[7] S. S. Barpanda, B. Majhi, P. K. Sa, A. K. Sangaiah, and S. Bakshi, "Iris feature extraction through wavelet mel-frequency cepstrum coefficients," *Optics & Laser Technology*, vol. 110, pp. 13–23, 2019.

[8] Y. Q. Qin and X. Y. Zhang, "Hmm-based speaker emotional recognition technology for speech signal," in *Advanced Materials Research*, vol. 230. Trans Tech Publ, 2011, pp. 261–265.

[9] J. Pribil, A. Pribilova, and J. Matousek, "Artefact determination by gmm-based continuous detection of emotional changes in synthetic speech," in *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2019, pp. 45–48.

[10] B. Schuller, B. Vlasenko, F. Eyben, M. Wöllmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[11] B. Gupta and S. Dhawan, "Deep learning research: Scientometric assessment of global publications output during 2004-17," *Emerging Science Journal*, vol. 3, no. 1, pp. 23–32, 2019.

[12] S. Kumar, T. Roshni, and D. Himayoun, "A comparison of emotional neural network (enn) and artificial neural network (ann) approach for rainfall-runoff modelling," *Civil Engineering Journal*, vol. 5, no. 10, pp. 2120–2130, 2019.

[13] Q. Cao, M. Hou, B. Chen, Z. Zhang, and G. Lu, "Hierarchical network based on the fusion of static and dynamic features for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6334–6338.

[14] J. Liu, Z. Liu, L. Wang, L. Guo, and J. Dang, "Speech emotion recognition with local-global aware deep representation learning," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7174–7178.

[15] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] L. Tarantino, P. N. Garner, A. Lazaridis *et al.*, "Self-attention for speech emotion recognition," in *Interspeech*, 2019, pp. 2578–2582.

[18] X. Wang, M. Wang, W. Qi, W. Su, X. Wang, and H. Zhou, "A novel end-to-end speech emotion recognition network with stacked transformer layers," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6289–6293.

[19] D. Hu, X. Hu, and X. Xu, "Multiple enhancements to lstm for learning emotion-salient features in speech emotion recognition," *Proc. Interspeech 2022*, pp. 4720–4724, 2022.

[20] S. Kwon *et al.*, "Att-net: Enhanced emotion recognition system using lightweight self-attention module," *Applied Soft Computing*, vol. 102, p. 107101, 2021.

[21] L. Guo, L. Wang, C. Xu, J. Dang, E. S. Chng, and H. Li, "Representation learning with spectro-temporal-channel attention for speech emotion recognition," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6304–6308.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[23] 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.

[24] W. Zhu and X. Li, "Speech emotion recognition with global-aware fusion on multi-scale feature representation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6437–6441.

[25] E. Guizzo, T. Weyde, and J. B. Leveson, "Multi-time-scale convolution for emotion recognition from speech audio signals," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6489–6493.

[26] Y. Xu, H. Xu, and J. Zou, "Hgfm: A hierarchical grained and feature model for acoustic emotion recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6499–6503.

[27] C. Li, "Robotic emotion recognition using two-level features fusion in audio signals of speech," *IEEE Sensors Journal*, 2021.

[28] J. Liu and H. Wang, "A speech emotion recognition framework for better discrimination of confusions," in *Interspeech*, 2021, pp. 4483–4487.

[29] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech emotion recognition with co-attention based multi-level acoustic information," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7367–7371.

[30] S. Zhang, X. Zhao, and Q. Tian, "Spontaneous speech emotion recognition using multiscale deep convolutional lstm," *IEEE Transactions on Affective Computing*, 2019.

[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," 2018.

[32] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[33] P. Li, Y. Song, I. V. McLoughlin, W. Guo, and L.-R. Dai, "An attention pooling based representation learning method for speech emotion recognition," 2018.

[34] Y.-X. Xi, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "Frontend attributes disentanglement for speech emotion recognition," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7712–7716.

[35] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[36] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

[37] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[38] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[39] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech emotion recognition with multi-task learning," in *Interspeech*, vol. 2021, 2021, pp. 4508–4512.

[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[41] N. Ristea, R. Ionescu, and F. Khan, "Septr: Separable transformer for audio spectrogram processing. arxiv 2022," *arXiv preprint arXiv:2203.09581*.

[42] H.-t. Xu, J. Zhang, and L.-r. Dai, "Differential time-frequency log-mel spectrogram features for vision transformer based infant cry recognition," *Proc. Interspeech 2022*, pp. 1963–1967, 2022.

[43] Q.-L. Zhang and Y.-B. Yang, "Sa-net: Shuffle attention for deep convolutional neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 2235–2239.

[44] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[45] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[46] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[47] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[48] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "Cmt: Convolutional neural networks meet vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 175–12 185.

[49] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[50] A. Trockman and J. Z. Kolter, "Patches are all you need?" *arXiv preprint arXiv:2201.09792*, 2022.

[51] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1314–1324.

[52] 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, pp. 335–359, 2008.

[53] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss *et al.*, "A database of german emotional speech." in *Interspeech*, vol. 5, 2005, pp. 1517–1520.

[54] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[55] S. Suganya and E. Charles, "Speech emotion recognition using deep learning on audio recordings," in *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, vol. 250. IEEE, 2019, pp. 1–6.

[56] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and B. W. Schuller, "Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition," *arXiv preprint arXiv:2005.08453*, 2020.

[57] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6474–6478.

[58] S. Li, X. Xing, W. Fan, B. Cai, P. Fordson, and X. Xu, "Spatiotemporal and frequential cascaded attention networks for speech emotion recognition," *Neurocomputing*, vol. 448, pp. 238–248, 2021.

[59] 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*, vol. 13, no. 2, pp. 992–1004, 2022.

[60] Y. Gao, J. Liu, L. Wang, and J. Dang, "Metric learning based feature representation with gated fusion model for speech emotion recognition." in *Interspeech*, 2021, pp. 4503–4507.

[61] D. Dai, Z. Wu, R. Li, X. Wu, J. Jia, and H. Meng, "Learning discriminative features from spectrograms using center loss for speech emotion recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7405–7409.

[62] T. Tuncer, S. Dogan, and U. R. Acharya, "Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques," *Knowledge-Based Systems*, vol. 211, p. 106547, 2021.

[63] S. Zhong, B. Yu, and H. Zhang, "Exploration of an independent training framework for speech emotion recognition," *IEEE Access*, vol. 8, pp. 222 533–222 543, 2020.

[64] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using an optimal combination of features based on emd-tkeo," *Speech Communication*, vol. 114, pp. 22–35, 2019.

[65] M. Hou, Z. Zhang, Q. Cao, D. Zhang, and G. Lu, "Multi-view speech emotion recognition via collective relation construction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 218–229, 2021.

[66] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.

[67] Mehta, Sachin and Rastegari, Mohammad, "Separable self-attention for mobile vision transformers," *arXiv preprint arXiv:2206.02680*, 2022.

[68] S. N. Wadekar and A. Chaurasia, "Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features," *arXiv preprint arXiv:2209.15159*, 2022.

Xiaoyu Tang (Member, IEEE) received the B.S. degree from South China Normal University in 2003 and the M.S. degree from Sun Yat-sen University in 2011. He is currently pursuing the Ph.D. degree with South China Normal University. He is working with the School of Physics and Telecommunication Engineering, South China Normal University, where he engaged in information system development. His research interests include machine vision, intelligent control, and the Internet of Things. He is a member of the IEEE ICICSP Technical Committee.

Yixin Lin received the B.Eng. degree from the School of Physics and Telecommunication Engineering, South China Normal University, in 2021, where he is currently pursuing the M.E. degree with the Department of Electronics and Information Engineering. His research interests include artificial intelligence and speech emotion recognition.

Ting Dang is a Senior Research Scientist at Nokia Bell Labs, Cambridge, UK. Before joining Nokia, she worked as a Senior Research Associate at the University of Cambridge. Dr. Dang earned her Ph.D. degree from the University of New South Wales in Sydney, Australia, and holds MEng and BEng degrees in Signal Processing from the Northwestern Polytechnical University in China. Her primary research interests are on exploring the potential of audio signals (e.g., speech) for mobile health, i.e., automatic disease and mental state prediction and monitoring (e.g., COVID-19, emotion) via mobile and wearable audio sensing. Her work aims to develop generalised and interpretable machine learning models to improve healthcare delivery. She served as the (senior) program committee and invited reviewer for more than 30 top-tier conferences and journals, such as AAAI, NeurIPS, ICASSP, IEEE TAC, IEEE TASL etc. She has won the Asian Dean's Forum (ADF) Rising Star Women in Engineering Award 2022, IEEE Early Career Writing Retreat Grant 2019 and ISCA Grant 2017.

Yuanfang Zhang is currently principal researcher at Autocity (Shenzhen) autonomous driving Co.,ltd. He got his dual Ph.D. degrees with the School of Computer Science at Northwestern Polytechnical University, Shaanxi Province, China and Faculty of Engineering and IT, University of Technology Sydney, Australia. His current research focuses on unmanned vehicle sensing technology, multimodal learning methodology, comprehensive computer vision.

Jintao Cheng received his bachelor's degree from the school of Physics and Telecommunications Engineering, South China Normal University in 2021. His research is computer vision, SLAM and deep learning.