IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 July 2024, accepted 9 August 2024, date of publication 22 August 2024, date of current version 19 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3447770

RESEARCH ARTICLE

Accuracy Enhancement Method for Speech Emotion Recognition From Spectrogram Using Temporal Frequency Correlation and Positional Information Learning Through Knowledge Transfer

JEONG-YOON KIM AND SEUNG-HO LEE
Department of Electronic Engineering, Hanbat National University, Daejeon 34158, Republic of Korea

Corresponding author: Seung-Ho Lee (shlee@cad.hanbat.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant by the Ministry of Science and ICT (MSIT) under Grant NRF-2022R1F1A1066371 and in part by the Regional Innovation Strategy (RIS) through the National Research Foundation of Korea (NRF) Grant by the Ministry of Education (MOE) under Grant 2021RIS-004.

ABSTRACT In this paper, we propose a method to improve the accuracy of speech emotion recognition (SER) by using vision transformer (ViT) to attend to the correlation of frequency (y-axis) with time (x-axis) in spectrogram and transferring positional information between ViT through knowledge transfer. The proposed method has the following originality i) We use vertically segmented patches of log-Mel spectrogram to analyze the correlation of frequencies over time. This type of patch allows us to correlate the most relevant frequencies for a particular emotion with the time they were uttered. ii) We propose the use of image coordinate encoding, an absolute positional encoding suitable for ViT. By normalizing the x, y coordinates of the image to -1 to 1 and concatenating them to the image, we can effectively provide valid absolute positional information for ViT. iii) Through feature map matching, the locality and location information of the teacher network is effectively transmitted to the student network. Teacher network is a ViT that contains locality of convolutional stem and absolute position information through image coordinate encoding, and student network is a structure that lacks positional encoding in the basic ViT structure. In feature map matching stage, we train through the mean absolute error (L1 loss) to minimize the difference between the feature maps of the two networks. To validate the proposed method, three emotion datasets (SAVEE, EmoDB, and CREMA-D) consisting of speech were converted into log-Mel spectrograms for comparison experiments. As a result of the experiment, the proposed method achieved 99.47%, 99.76%, and 95.24% accuracy, significantly exceeding the state-of-the-art methods, with much less computational complexity on three emotion datasets.

INDEX TERMS Speech emotion recognition (SER), positional encoding, transfer learning, vision transformer (ViT), temporal frequency correlation.

I. INTRODUCTION

Speech-related tasks, such as speech to text (STT) and speech emotion recognition (SER), are becoming increasingly important in our daily lives, with applications in intelligent speakers, voice assistants, and more. Early studies utilized only the magnitude of speech samples [1], [2], [3], [4], [5], but recently, methods used in visual tasks have been introduced to analyze spectrogram, which convert frequency information of speech into images [6], [7], [8], [9], [10], [11], [12],

The associate editor coordinating the review of this manuscript and approving it for publication was Shadi Alawneh.

[13], [14], [15]. The most popular method for converting speech samples into spectrogram images is the short-time Fourier transform (STFT). The STFT is a Fourier-related transform used to determine the frequency components of a small receptive field of a time-varying signal. Derives the frequency components from the magnitude of the speech sample using STFT, which are then mapped to the mel scale to more closely match human hearing characteristics, creating a log-Mel spectrogram. In recent speech-related research, the log-Mel spectrogram is used for analysis and classification.

We focuses on SER among various speech-related tasks, and for this purpose, we compare pros and cons of convolutional neural networks (CNNs) [16] and vision transformer (ViT) [17], which are commonly used for vision tasks. CNNs are characterized by including and analyzing pixels near the input data and have been used in various fields for a long time. In particular, the locality obtained by analyzing pixels near the input data plays a big role in improving the accuracy of vision tasks. However, CNNs are prone to overfitting due to excessive locality, are not suitable for large datasets, and can lead to inaccurate results when there is a large difference from the input data. To deal with these issues, ViT have emerged as an alternative to CNNs. Generally, ViT segment the input image into multiple square patches and then perform global processing through multi-head self-attention, which has the potential to significantly improve the performance of vision tasks. However, global processing is characterized by slow convergence, need large datasets, and difficult optimization. Furthermore, ViT is sensitive to optimizers, dataset-appropriate learning hyperparameters, training schedules, and network depth, requiring many experiments to empirically select values. In recent study has proposed a method that combines the advantages of CNNs with the advantages of ViT [18]. This method replaces ViT's patchify process with several convolutional layers to initially extract local features and then analyze them globally using ViT. The multiple convolutional layers used in the patchify process are called convolutional stem and serve as a robust improvement over ViT's slow convergence speed and variation in learning hyperparameters. However, the application of convolutional stem slightly but clearly increases the number of trainable parameters, which increases the required resources. Also, if the convolutional stem does not perform size reduction of the image and retains the rich information, it will require more resources.

Therefore, in this paper, we propose ViT, which uses knowledge transfer to learn the advantages of the convolution without increasing resources, and vertically partitioned patches to analyze the frequency correlation over time in the log-Mel spectrum. We also propose the use of image coordinate encoding, which is an absolute positional encoding method suitable for ViT, assuming that the locality inference ability of convolutional stem is insufficient. We design a teacher network and a student network for knowledge transfer. Teacher network is a ViT with convolutional stem and image coordinate encoding, and student network is a basic ViT with vertically partitioned patches and no location encoding. Figure 1 shows the overall process of the proposed method in this paper.

The main contributions of this paper are as follows. i: For temporal frequency correlation analysis in Mel spectrogram images, we analyze through comparison of attention masks how vertically long patches, rather than the existing square-shaped patches, can affect SER accuracy; ii: Propose the use of image coordinate encoding methods suitable for ViT; iii: Increase resources by transferring knowledge from relatively large-scale ViT, which uses a vertically long patch shape and includes convolutional stem and positional encoding, to relatively small-scale ViT, which uses a vertically long patch shape without convolutional stem and positional encoding. We empirically prove through accuracy and cosine similarity that performance improvement is possible.

II. RELATED WORKS

In the early days of speech signal processing, automatic speech recognition (ASR) [19], [20], [21] and speech to text (STT) [22] were important due to the development of voice assistant technology to provide hands-free help. However, recent speech signal processing research has placed considerable emphasis on SERs, with efforts to develop more accurate emotion recognition models using log-Mel spectrogram-based features. While early approaches focused primarily on more classical machine learning-based methods, there has been a recent shift toward deep neural network-based models. Among them, methods that utilize the attention mechanism have seen a noticeable increase.

Dosovitskiy et al. [17] proposed ViT, an application of attention-only transformers to image classification. They applied a transformer encoder to image classification and used patches of 16×16 segmented input images instead of word token embeddings as the input sequence. ViT performed poorly when trained on small datasets, but excelled when trained on large datasets. These results suggested that transformers could replace much of the work done by CNNs, often with higher accuracy, and many speech processing methods based on ViT have emerged.

Gong et al. [14] proposed an audio spectrogram transformer (AST) that analyzes log-Mel spectrogram by dividing them into square patches using ViT. They further trained on log-Mel spectrogram using DeiT [23], which is a knowledge distillation [24] from CNNs trained on the ImageNet dataset [25]. In the patchify process, they partitioned a 16×16 patch with 6 pixel overlap to enable fine-grained analysis of log-Mel spectrogram through self-attention.

Ristea et al. [15] proposed a separable transformer (SepTr) that analyzes a spectrogram twice, horizontally and vertically. SepTr consists of a horizontal transformer that analyzes the log-Mel spectrogram in the horizontal direction and a vertical transformer that analyzes it in the vertical direction. In the patchify process, split the log-Mel spectrogram into 1×1 patches to the analysis.

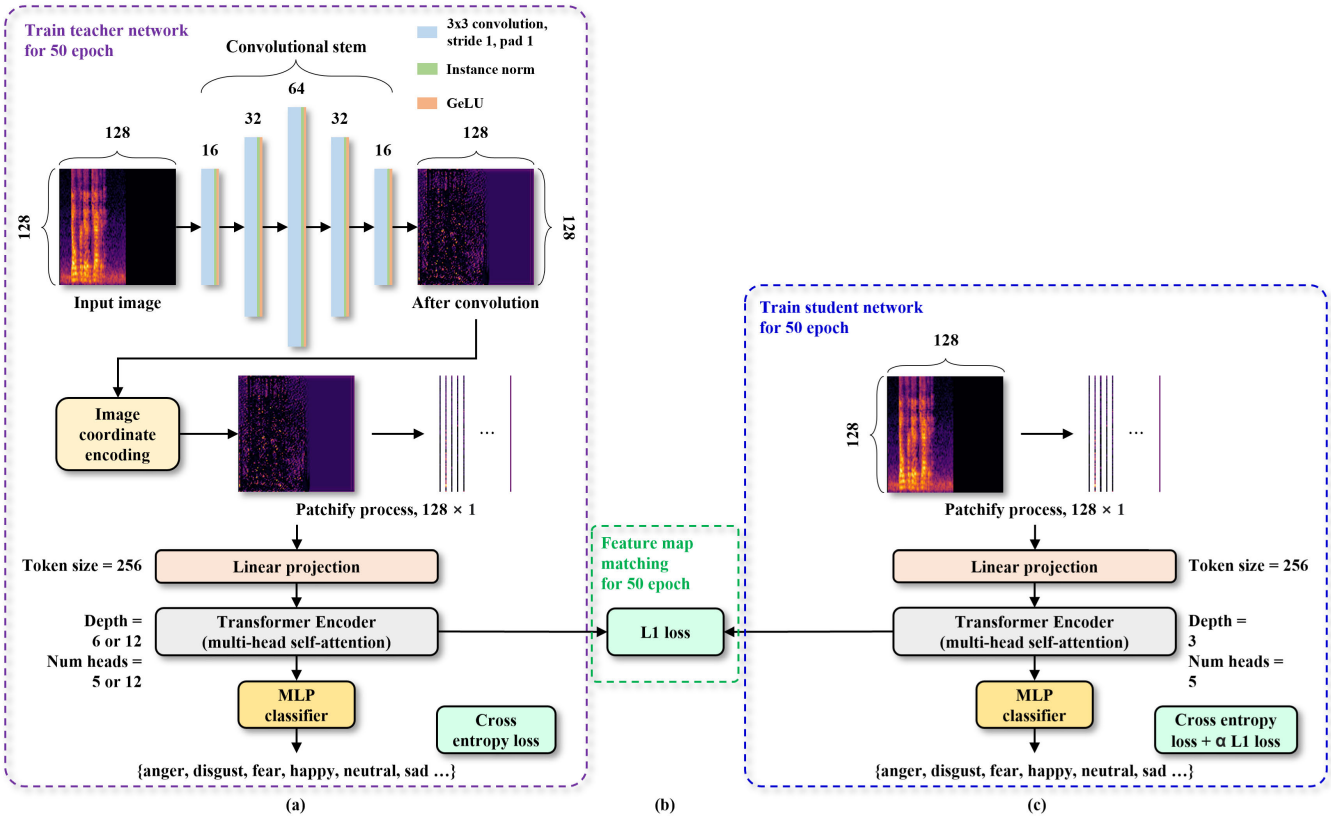

FIGURE 1. The overall process of proposed method in this paper. i: (a) Train the teacher network using cross entropy loss. ii: (b) Match the feature maps of the student network and the trained teacher network using Mean absolute error(L1 loss). iii: (c) Train the student network with the feature map matching performed using cross entropy loss + α * L1 loss. Each stage is performed repeatedly with 50 epochs.

Zhang et al. [26] proposed a new approach for speech emotion recognition using the F-Emotion algorithm for feature clustering and a parallel deep learning model. The F-Emotion algorithm calculates the weight of each voice emotional feature and determines the optimal feature combination for emotion recognition. Parallel deep learning models process these optimal features and generate separate recognition results for each emotion category. Their method achieved accuracy of 88.80% in the evaluation on EmoDB by decision fusion of the outputs of all parallel channels through a voting mechanism.

Ong et al. [27] proposed an analysis method using an enhanced multiscale vision transformer (MViTv2) as a voice emotion recognition method using log-Mel spectrogram. Log-Mel spectrogram provides a representation that can capture the complex frequency spectrum and temporal characteristics of speech signals. MViTv2's multiscale attention mechanism and transformer-based structure achieved 90.57% accuracy in an evaluation conducted on EmoDB by capturing complex relationships between features at various scales.

Kakuba et al. [28] proposed a method to improve speech emotion recognition accuracy in log-Mel spectrogram by combining dilated convolution and attention mechanism. This method uses the characteristics of dilated convolution, which allows the receptive field to be widened while maintaining the amount of calculation. The wider receptive field achieved accuracy of 93.75% and 95.93% in SAVEE and EmoDB, respectively by capturing long-term dependencies through wider analysis of the spectrogram.

Saleem et al. [29] proposed DeepCNN, which consists of a parallel CNN and a convolutional transformer, to extract spectro-temporal features from Mel frequency cepstral coefficients (MFCCs). DeepCNN consists of a framework consisting of two parallel CNNs for low-level feature extraction, a convolutional transformer for spectro-temporal feature extraction, high-level feature representation using an attention mechanism, and a 4-layer CNNs for emotion classification. As a result, DeepCNN achieved 94.20% accuracy in an evaluation conducted on EmoDB by capturing spectro-temporal features well using a convolutional transformer.

Mishra et al. [30] proposed a multi-resolution variational mode decomposition (MRVMD) method. Their proposed method uses MRVMD to decompose the speech signal into sub-signals known as intrinsic mode functions (IMFs), such as multi-resolution variational mode Mel-frequency cepstral coefficient (MRVMMFCC) and multi-resolution variational mode approximate entropy (MRVMAE). Afterwards, these features are combined and used to classify emotions using

a deep neural network (DNN) classifier method achieved an accuracy of 83.40% and 90.51% for SAVEE and EmoDB.

Zhang et al. [31] proposed a method to comprehensively explore and integrate pre-training, self-training, and model size expansion. By leveraging these advanced techniques, they have significantly improved data efficiency. Through the use of a pre-trained conformer model with 8 billion parameters and a combination of pre-training, self-training, and scaling up model size, the accuracy of SAVEE and CREMA-D was significantly improved by 92.50% and 88.20% even with limited label data. Additionally, demonstrate the general benefits of utilizing large-scale pre-training and self-training models across a variety of speech domains and dataset sizes.

However, transformer-based methods still suffer from the problem that they are sensitive to hyperparameters during training and difficult to optimize. Xiao et al. [18] determined that the problem with transformer-based methods is the lack of locality, which can be achieved through hyperparameter-robust and relatively easy-to-optimize CNNs. They replaced ViT's patchify process with a convolutional stem consisting of multiple layers of CNNs to achieve locality while improving on the problems of transformer-based methods. Chen et al. [32] found that freezing the weights of the patchify process in ViT's self-supervised learning to a random initialization improves stability. They also found that removing positional embeddings only slightly reduced accuracy.

Islam et al. [33] explored the extent to which CNNs encode spatial positional information and its impact on vision tasks such as semantic segmentation and salient object detection. They found that CNNs inferred information about spatial positioning from zero-padding located at the boundaries of the image and that they relied on and learned from positional information to a much greater extent than expected. In particular, positional information was more prominent in the deeper layers of the CNNs. Baevski et al. [22] used the inference of positional information using these CNNs as a substitute for positional encoding in the transformer, achieving high accuracy for speech recognition using the magnitude of the speech sample.

We determined that for 1D speech magnitude data such as [22], it is easy to infer absolute positional information from a small number of layers of CNNs. However, for 2D spectrogram, a small number of CNNs is insufficient and absolute positional encoding is required separately at the input of the transformer. Therefore, as a method suitable for ViT, we perform image coordinate encoding by normalizing the coordinates of 2D images to -1 to 1 and concatenating them to the image. In addition, it is common for existing spectrogram analysis methods to use a square-shaped receptive field. However, we assume that it would be more effective to correlate the frequency (y axis) corresponding to a specific emotion with the time (x axis) at which it was uttered, so we use a receptive field that can completely segment the spectrogram vertically.

III. BACKGROUND

A. SHORT-TIME FOURIER TRANSFORM

We convert each audio sample into a 2D time-frequency matrix to get an image-like representation. To do this, we compute the discrete short-time Fourier transform (STFT) as follows:

$$STFT(m,k) = \sum_{n=-\infty}^{\infty} x[n] \cdot w[n - aH] \cdot e^{-j\frac{2\pi}{N}kn} \quad (1)$$

Given input discrete signal $x[n]$, $w[n]$ is a window function (in this paper, Hamming) of length L, H is the hop size, and N is the total number of discrete Fourier transform (DFT) points (frequency bins). $STFT(m,k)$ is STFT coefficient for the k^{th} frequency bin and the m^{th} time-frame

B. LOG-MEL SPECTROGRAM

The Mel scale is a scale of frequencies that a listener determine to be the same distance from each other. The human ear easily distinguish the difference between frequencies up to 1000 Hz, but cannot distinguish the difference between 10 kHz and 10.5 kHz. Therefore, it is converted to Mel scale to match human hearing characteristics. The spectrogram obtained by STFT is log-transformed and Mel-scaled to become a log-Mel spectrogram. The Mel scale is linear up to 1 kHz and logarithmic for frequencies above that. To convert the spectrogram to Mel scale, the computed spectrogram is passed through a Mel-filter bank.

$$Mel(f) = 2595 \log(1 + \frac{f}{700}) \quad (2)$$

IV. METHODS

A. CONVOLUTIONAL STEM

In this paper, we design a convolutional stem consisting of six 3×3 convolutional layers to verify that the locality of the convolution can be learn to ViT through knowledge transfer. The output channels of the convolutional stem are [16, 32, 64, 32, 16, 1], respectively, and are output with stride $= 1$ and zero-padding $= 1$ to retain the input image size. We retain size to avoid large differences in downsizing and to make the student network and input sequence sizes similar in order to analyze the impact of convolutional stem. Batch normalization is a more common method, but since we use 4 mini-batches, similar to the experiment in [15], we use instance normalization, which is more advantageous with fewer mini-batches. For the activation function, we use gaussian error linear unit (GELU). GELU is bounded differently than the more common rectified linear unit (ReLU), LeakyReLU, and is therefore more immune to gradient vanishing. Figure 2 shows the structure of the convolutional stem in this paper.

B. IMAGE COORDINATE ENCODING

Relative position encoding, absolute position encoding, and position embedding are methods that are often used in the context of sequence data, such as in natural language

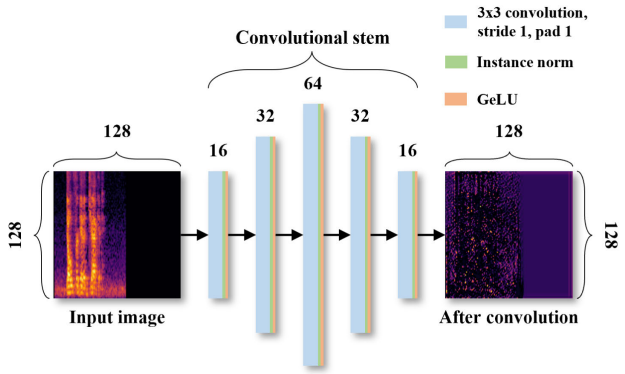

FIGURE 2. Structure of convolutional stem in this paper. The numbers above each block represent the output channels. To ensure that no information is lost, the after convolution is equal to the input size of 128 × 128.

processing or computer vision tasks. These methods aim to provide the network with information about the positional relationships between elements. On the other hand, if a transformer is used after a network that can infer positional information, such as CNNs, good performance can be obtained without performing a separate positional encoding[]. However, since only relative positional information can be inferred initially through border recognition through zero-padding, more trains are needed to infer absolute position from this information, convergence is slow, and overfitting occurs before global minima. Therefore, in this paper, image coordinate encoding is performed by normalizing and concatenating the coordinates of log-Mel spectrogram through convolutional stem to values from −1 to 1 (see figure 3).

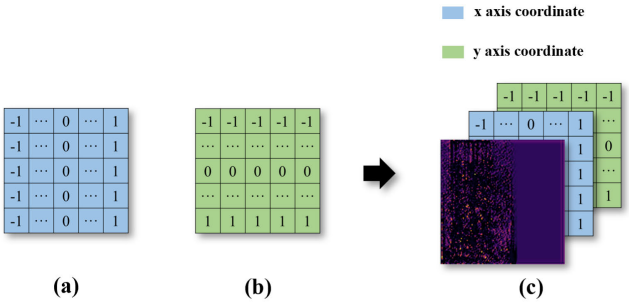

FIGURE 3. Image coordinate encoding. The x, y coordinates of the image are normalized to the range −1 to 1, (a) and (b) concatenated to a log-Mel spectrogram after convolutional stem, (c).

C. MULTI-HEAD SELF-ATTENTION

Transformer's encoder and decoder are both based on the attention mechanism. The attention mechanism is designed to compensate for the weakness of the seq2seq model, which processes each word in a sentence. The seq2seq model can only refer to words near the output layer. With attention, however, all words can be referenced in the calculation of the result without being bound by short- or long-term dependencies. In vision tasks, attention calculates how much attend to each input sequence (i.e., patch) to determine how much each input should be reflected in the resulting computation, i.e., in tasks such as object detection and image classification, more weight is given to important patches when they are segmented from the entire image. Self-attention can learn correlations between input sequences by attend to itself. The self-attention layer consists of three trainable weight matrices that are used to derive a query Q, key K, and value V from an input sequence X. The output attention(Q, K, V) is the weight of the input sequence. The output attention(Q, K, V) as follows:

$$Attention(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \qquad (3)$$

K^T represents the transpose of K, and d_k represents the dimension of K. Multi-head self-attention is a module for executing the self-attention mechanism multiple times in parallel. Intuitively, multi-head self-attention allows for more accurate analysis because it can attend to multiple subspaces differently.

D. NETWORK ARCHITECTURE

In this paper, we design two networks to verify that features with convolutional locality and absolute positional encoding can be implemented through basic ViT without positional encoding. The teacher network with convolutional stem and image coordinate encoding, and the student network without positional encoding in a basic structure ViT are designed for knowledge transfer. Knowledge transfer is performed in the direction of minimizing the difference between the features immediately before each multi-layer perceptron (MLP) classifier.

The teacher network consists of a convolutional stem and image coordinate encoding, ViT with token size = 256, and an MLP classifier. ViT is tuned to depth = 6 or 12 and num heads = 5 or 12 to compare performance with methods such as [14] and [15]. In general, the deeper the depth and the more heads, the more difficult it is to optimize for small-sized datasets. However, higher performance can be expected when sufficiently large data is available. Therefore, in this paper, several sizes of ViT are created and used in the experiments for objective comparison. The patch size is 128 × 1 (height, width), which can completely split a 128 × 128 image vertically. Figure 4 shows the structure of the teacher network.

The student network is a ViT consisting purely of multi-head self-attention, without convolutional stem and positional encoding. The patch size is 128 × 1, which is the same as the teacher network, because the input sequence between the teacher network and the student network must be similar to extract similar features, and knowledge transfer through feature map matching and verify implementation for locality and positional encoding are possible. The student network has

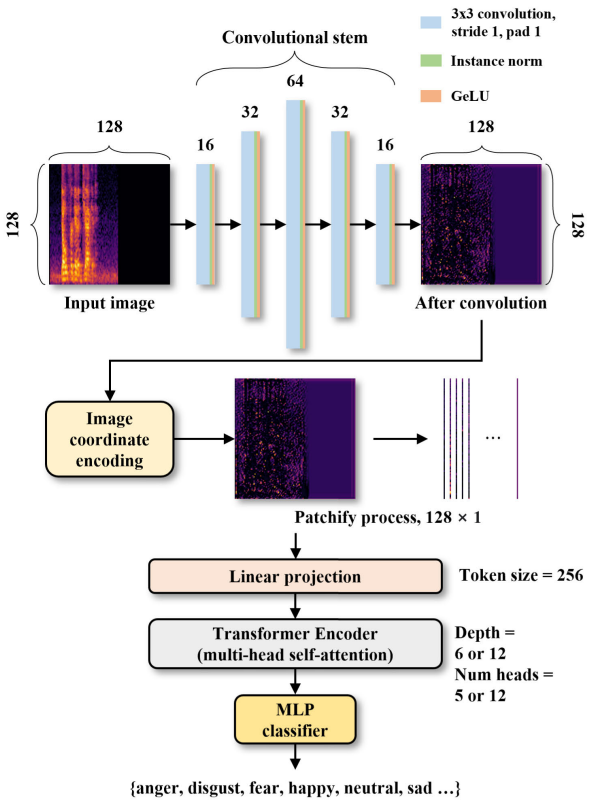

FIGURE 4. The structure of the teacher network. It contains a convolutional stem to obtain locality and an image coordinate encoding to disambiguate the location information.

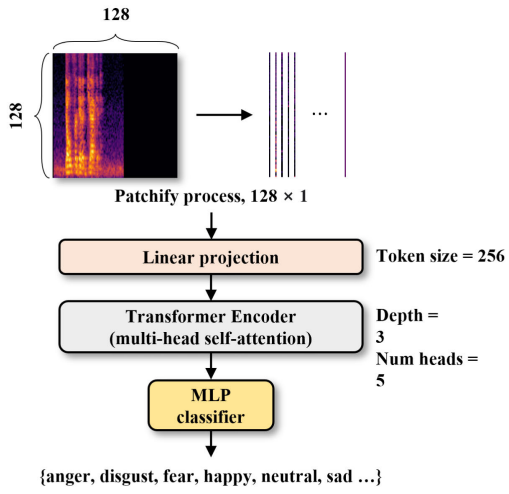

FIGURE 5. Structure of the student network. To verify that locality inference and positional encoding is possible through transfer learning without convolutional stem and image coordinate encoding, we construct a basic ViT.

token size = 256, depth = 3, and num heads = 5. Figure 5 shows the structure of the student network.

E. FEATURE MAP MATCHING

We perform feature map matching to effectively transfer the knowledge of the teacher network to the student network. In the feature map matching step, after freezing the weights of the trained teacher network, we calculate the mean absolute error (L1 loss) between features to make the features of the student network similar to the features of the teacher network, and train to minimize it.

V. EXPERIMENTS

A. DATASETS

The **SAVEE** dataset [34] contains 480 acted English utterances recorded by four male actors and consists of seven emotion categories: anger, fear, disgust, happiness, neutral, sadness, and surprise.

The **EmoDB** dataset [35] consists of 535 German utterances from 10 actors (5 female, 5 male) and includes seven emotion categories: anger, anxiety, boredom, disgust, happiness, neutral, and sadness.

The **CREMA-D** dataset [36] consists of 7,442 videos of 91 actors (48 male and 43 female) from various ethnic groups. Actors portray a variety of emotions by uttering 12 specific sentences that correspond to one of six emotional categories: anger, fear, disgust, happiness, neutrality, and sadness.

For each dataset, the data consists of speech waveforms collected from all participants, and was divided into trainset and testset at a ratio of 80%:20%. A sampling frequency of 16kHz was applied to all datasets, the duration was limited to 4 seconds, and zero padding was added to the empty space of samples less than 4 seconds. Additionally, we apply STFT with N = 1024, H = 64, window size = 512 and hamming window to convert them into log-Mel spectrogram images. In this paper, we perform data augmentation of noise perturbation, time shifting, and speed perturbation. The size of the generated log-Mel spectrogram image is 128 × 128.

B. EXPERIMENTAL SETUP

In our experiments, similar to [15], we train with 128 × 1 patch size for teacher network, feature map matching, and student network, respectively, with hyperparameters of batch size = 4, epoch = 50, and learning rate = 1e-4 (halves every 10 epochs). It is optimized with Adam. Afterwards, we train ViT with a 16 × 16 patch size for comparative experiments with different patch shapes. Classification loss calculates cross entropy loss and feature map matching loss calculates L1 loss. We use cross entropy loss for training the teacher network, L1 loss for training the feature map matching, and cross entropy + α * L1 loss for training the student network. (in this case $\alpha = 10$)

All experiments were performed on a Windows 10 workstation using an AMD Ryzen 5 7600X 6-Core Processor (4.70 GHz), 32 GB of RAM, and Nvidia RTX 4090 GPU (24GB of RAM, CUDA Cores: 16384). The entire workflow was implemented using CUDA library version 11.8 and cuDNN 8.9.3 in Pytorch version 2.0.1.

C. EVALUATION

The attention mask is a visual representation of which parts of the input image are most important to the attention weight matrix as a result of performing self-attention. We first

compare the difference between the attention masks of square patches and vertically elongated patches. If the attention mechanism leads to higher weighting of features that are important for the classification task, then similar attention masks should be derived for different augmented log-Mel spectrogram images. In this paper, for the attention mask comparison, we extracted attention masks from ViT using vertically elongated patches of size 128 × 1 and ViT using square patches of size 16 × 16. Both networks include convolutional stem and image coordinate encoding for comparison under the same conditions. The extracted attention masks were subjected to gaussian smoothing for ease of analysis. Figure 5 visually illustrates the experimental results of the difference in attention masks based on patch shape using randomly selected images from EmoDB. In case (a), which uses the method proposed in this paper, there is a common distinct line at the beginning and end of the signal, which indicates that similar images are recognized and attended to. However, in the case of (b), which uses a conventional square patch, we can see that it is difficult to find a common point between the attention masks. Furthermore, the time-shifted image shows confusing results with the highest attendance to the wrong zero padding region. As a result, we show that it is more robust and valid to analyze the association of frequency (y axis) with time (x axis) in a spectrogram for SER.

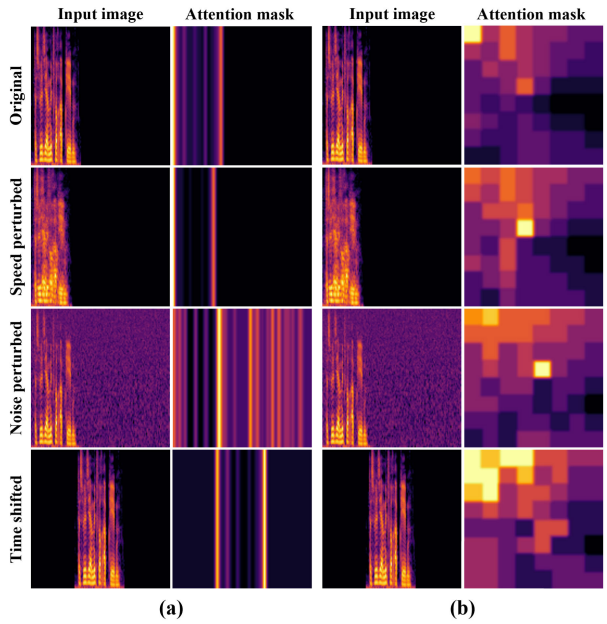

FIGURE 6. Comparison of attention masks from our proposed method (a) and using patches of size 16 × 16, (b). In case (a), where we analyzed the temporal frequencies correlation with vertically segmented patches, there is a common bright line (high attention weight) at the beginning and end of the signal. However, in case (b), the commonalities between the images are not well found and when time-shifted, the results are chaotic, with the highly attended in the zero padding region.

To verify the impact of the image coordinate encoding proposed in this paper, we compared the accuracy of the teacher network with depth = 6 and num heads = 5 with different positional encoding.

In Table 1, $teacher_{nope}$ is a teacher network that does not perform positional encoding, and $teacher_{ice}$ is a teacher network that performs image coordinate encoding. We can see that $teacher_{nope}$ performs poorly on all three datasets. This difference is more pronounced on CREMA-D than on SAVEE and EmoDB, which are relatively small datasets. The accuracy in Table 1 is the result of discarding the second decimal place.

TABLE 1. Accuracy comparison between without positional encoding ($teacher_{nope}$) vs. image coordinate encoding ($teacher_{ice}$).

Method	SAVEE↑	EmoDB↑	CREMA-D↑
$teacher_{nope}$	95.57%	98.83%	91.85%
$teacher_{ice}$	97.39%	99.06%	95.07%

Additionally, in this paper, to prove that the proposed knowledge transfer method through feature map matching is effective, we compared its objective performance with state-of-the-art methods through accuracy.

We trained the proposed method using the Adam optimizer with batch size = 4, epoch = 50, learning rate = 1e-4 (halves every 10 epochs). In addition, to examine the exact impact of the convolutional stem and image coordinate encoding, two methods consisting only of multi-head self-attention, AST [14] proposed by Gong et al. and SepTr [15] proposed by Ristea et al., were trained under the same conditions. In Table 2, ImageNet pretrain is applied to AST for an environment such as an experiment setup, and the conditions of patch size = 16 × 16, token size = 768, depth = 12, num heads = 12, etc. are used as is. The method was applied. In addition, for SepTr, for an environment similar to the experiments setup, conditions such as patch size = 1 × 1, token size = 256, vertical SepTr depth = 3, horizontal SepTr depth = 3, num heads = 5, etc. The method of this paper was applied. As a result of comparing Table 2 and Table 3, the method we proposed showed significantly improved performance by more than 5 10% over the previous state-of-the-art methods in all datasets used for evaluation, producing excellent results. Meanwhile, the highest performance was achieved in the smallest student network, and we assume that this result was achieved by using L1 loss. We believe that the log boundary problem of cross entropy loss has been solved by knowledge transfer through feature matching using L1 loss.

We displayed a confusion matrix to identify robustness or weakness categories for the test set randomly split 8:2. SAVEE and EmoDB are composed of 7 emotion categories, and CREMA-D is composed of 6 categories. As a result, SAVEE had the problem of incorrectly classifying sadness and surprise as neutral and happiness, respectively. EmoDB had the problem of misclassifying neutral as sadness, and the large-scale CREMA-D showed a significant number of errors, but showed common characteristics with other data

TABLE 2. Accuracy comparison with state-of-art-methods.

Method	SAVEE↑	EmoDB↑	CREMA-D↑
Zhang et al. [26] 2023	-	88.80%	-
Ong et al. [27] 2023	-	90.57%	-
Kakuba et al. [28] 2023	93.75%	95.93%	-
Saleem et al. [29] 2023	-	94.20%	-
Mishra et al. [30] 2024	83.40%	90.51%	-
Zhang et al. [31] 2022	92.50%	-	88.20%
AST [14] 2021	89.32%	96.02%	88.79%
SepTr [15] 2022	87.23%	92.75%	79.94%
Teacher(Ours)	98.17%	96.96%	92.64%
Student(Ours)	**98.95%**	**98.83%**	**94.07%**

that were confused between sadness and neutral. Figure 7 (a), (b), and (c) show the confusion matrices of SAVEE, EmoDB, and CREMA-D performed using the student network.

We analyzed the cosine similarity between the features of the teacher network and the student network to analyze the effectiveness of knowledge transfer through feature matching using L1 loss. As a result, the proposed method, which uses a patch form of the same shape to capture temporal frequency correlation but performs feature matching using L1 loss, showed high cosine similarity. This means that meaningful information such as locality and positional information could be reproduced even with a shallow layer transformer encoder through knowledge transfer through feature matching. In other words, if the training of ViT can be induced in some way to identify the latent features that can be obtained from the convolutional stem and the absolute coordinate information of the image from the beginning, a lighter and more generalized model can be made.

D. COMPUTATIONAL COMPLEXITY

To more objectively measure the computational complexity of the teacher network and student network proposed in this paper, we compared and measured ViT-based AST, SepTr, and trainable parameters, multiply–accumulate operation (MACs). As in the previous evaluation subsection, AST used the following conditions: patch size = 16 × 16, token size = 768, depth = 12, and num heads = 12. SepTr used conditions such as patch size = 1 In the case of AST, since imagenet pretrain is applied, it has an input parameter of 128 The teacher network and student network proposed in this paper were set to patch size = 128 × 1, depth = 6, num heads = 5, depth = 3, and num heads = 5, respectively. As a result of the measurement, the trainable parameter of AST was 21.4M and SepTr was 8.7M, while our teacher network was 2.9M and the student network was lower at 1.5M. In addition, while MACs' AST is 74.9G and SepTr is 143.2G and 0.6G, respectively, depending on the patch size, the teacher network is 1.1G, which is slightly higher than SepTr using a patch size of 16 × 16, and the student network is the lowest at 0.2G. appear. As a result, our method showed the lowest computational complexity. These results demonstrate that

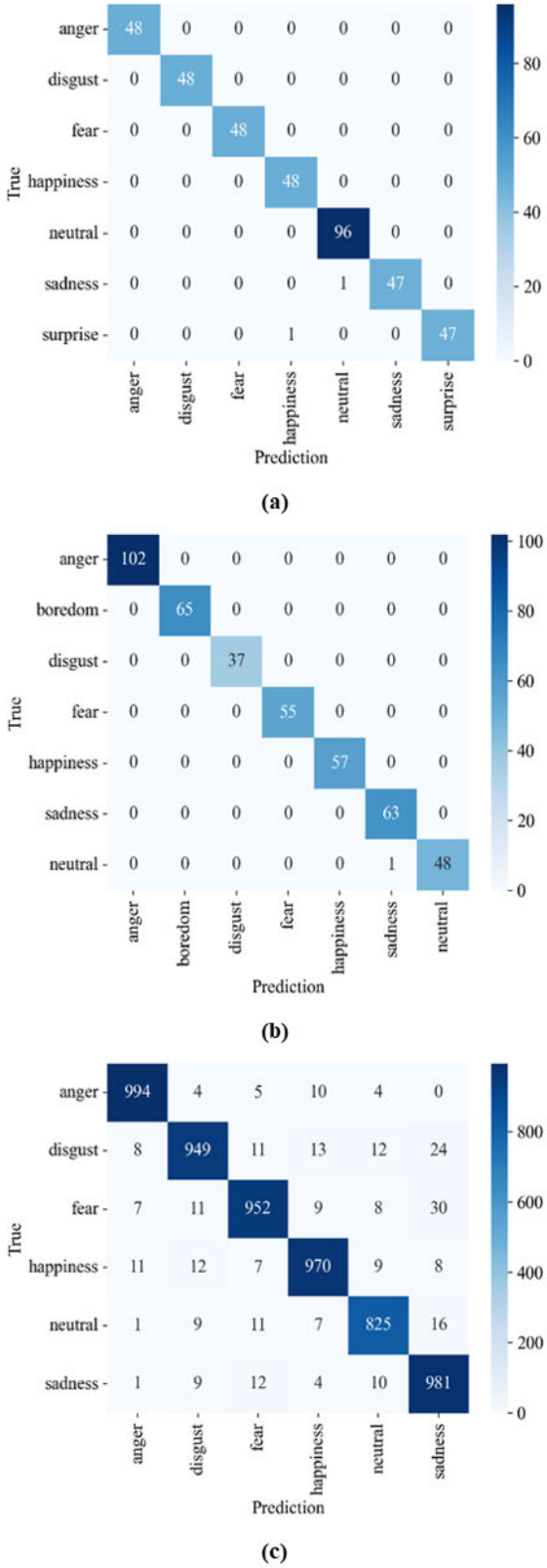

FIGURE 7. Confusion matrix of SAVEE, EmoDB, and CREMA-D performed with student network. (a) represents correct and incorrect SAVEE emotions. (b) represents correct and incorrect EmoDB emotions. (c) represents CREMA-D's correct and incorrect emotions.

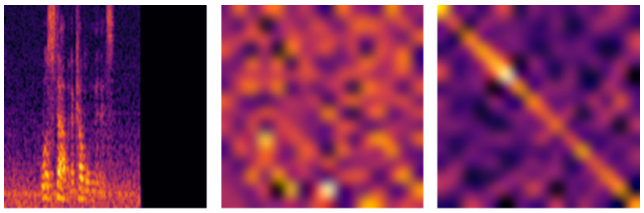

FIGURE 8. The difference between (b), which did not perform feature matching on the input spectrogram (a), and (c), which performed feature matching. We compared the cosine similarity between the features of the teacher network and the features of the student network.

our method of capturing temporal frequency correlation by splitting it into smaller (height, width) patches of size 128 × 1 has more input parameters than the method of analyzing log-Mel spectrograms with patches of size 16 × 16. Despite this, it is proven that higher accuracy can be achieved with lower computational complexity.

TABLE 3. Computational complexity comparison with state-of-art-methods.

Method	MACs↓	Params↓
AST	74.9G	21.4M
SepTr 1 x 1	143.2G	8.7M
SepTr 16 x 16	0.6G	8.7M
Teacher(Ours)	1.1G	2.9M
Student(Ours)	**0.2G**	**1.5M**

VI. CONCLUSION

In this paper, we proposed a new approach in the field of SER, accuracy enhancement method for SER from spectrogram using temporal frequency correlation and positional information learning through knowledge transfer. The proposed method showed that applying convolutional stem, image coordinate encoding, and vertically segmented patches to ViT can achieve significant performance improvement in SER using log-Mel spectrogram. Additionally, we verified that implementation is possible even in a simple structure through knowledge transfer through feature map matching from the teacher network to the student network. The proposed method achieved excellent results on various emotional datasets. The use of image coordinate encoding strengthened positional information, and vertically divided patches improved the performance of the network by allowing ViT's attending to operate on frequency corresponding to time. Moreover, knowledge transfer through feature map matching enabled effective information transfer between the teacher network and student network. In experiments, the results showed that the proposed method was more efficient and higher performance than state-of-the-art methods. This suggests that high accuracy can be achieved with less computational cost, which suggests the possibility of efficient and economical use in practical applications.

The method proposed in this study proves the effectiveness of changing the square-shaped patch of ViT, which was taken for granted, into a vertically divided shape suitable for log-Mel spectrogram images. In addition, it presents a new perspective in the SER field by providing excellent performance by utilizing the combination of convolutional stem and ViT, image coordinate encoding, and feature map matching. As a future research direction, research should be conducted on using patches of various shapes to enable appropriate attendance for various types of data for ViT, which has been studied while adhering to square patch segmentation in fields other than object detection and image classification. We speculate that if we can transform the image using appropriate padding and segment it into cross-shaped patches for analysis, will be able to capture the object's boundary more accurately. It is believed that additional research is needed to analyze using various patch types in semi-supervised and unsupervised network training methods guided by the attention mechanism.

REFERENCES

[1] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020.

[2] F. Daneshfar, S. J. Kabudian, and A. Neekabadi, "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier," *Appl. Acoust.*, vol. 166, Sep. 2020, Art. no. 107360.

[3] V. Garg, H. Kumar, and R. Sinha, "Speech based emotion recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2013, pp. 1–5.

[4] M. Hou, J. Li, and G. Lu, "A supervised non-negative matrix factorization model for speech emotion recognition," *Speech Commun.*, vol. 124, pp. 13–20, Nov. 2020.

[5] L. Tarantino, P. N. Garner, and A. Lazaridis, "Self-attention for speech emotion recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2578–2582.

[6] B. Maji, M. Swain, and M. Mustaqeem, "Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and Bi-GRU features," *Electronics*, vol. 11, no. 9, p. 1328, Apr. 2022.

[7] F. Andayani, L. B. Theng, M. T. Tsun, and C. Chua, "Hybrid LSTM-transformer model for emotion recognition from speech audio files," *IEEE Access*, vol. 10, pp. 36018–36027, 2022.

[8] S. Nagarajan, S. S. S. Nettimi, L. S. Kumar, M. K. Nath, and A. Kanhe, "Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales," *Digit. Signal Process.*, vol. 104, Sep. 2020, Art. no. 102763.

[9] S. Chen, M. Zhang, X. Yang, Z. Zhao, T. Zou, and X. Sun, "The impact of attention mechanisms on speech emotion recognition," *Sensors*, vol. 21, no. 22, p. 7530, Nov. 2021.

[10] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021.

[11] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[12] P. Jiang, H. Fu, H. Tao, P. Lei, and L. Zhao, "Parallelized convolutional recurrent neural network with spectral features for speech emotion recognition," *IEEE Access*, vol. 7, pp. 90368–90377, 2019.

[13] D. Bertero and P. Fung, "A first look into a convolutional neural network for speech emotion detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5115–5119.

[14] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Proc. Interspeech*, Aug. 2021, pp. 571–575.

[15] N.-C. Ristea, R. Tudor Ionescu, and F. Shahbaz Khan, "SepTr: Separable transformer for audio spectrogram processing," 2022, *arXiv:2203.09581*.

[16] E. Akleman, "Deep learning," *Computer*, vol. 53, no. 9, pp. 1–17, Sep. 2020.
[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, 2021, pp. 1–26.
[18] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," in *Proc. NIPS*, Dec. 2021, pp. 30392–30400.
[19] N. Kanda, G. Ye, Y. Gaur, X. Wang, Z. Meng, Z. Chen, and T. Yoshioka, "End-to-end speaker-attributed ASR with transformer," in *Proc. Interspeech*, Aug. 2021, pp. 4413–4417.
[20] T. Lohrenz, Z. Li, and T. Fingscheidt, "Multi-encoder learning and stream fusion for transformer-based end-to-end automatic speech recognition," in *Proc. Interspeech*, Aug. 2021, pp. 2846–2850.
[21] C.-H. Leong, Y.-H. Huang, and J.-T. Chien, "Online compressive transformer for end-to-end speech recognition," in *Proc. Interspeech*, Aug. 2021, pp. 1500–1504.
[22] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2020, pp. 12449–12460.
[23] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[26] L.-M. Zhang, G. W. Ng, Y.-B. Leau, and H. Yan, "A parallel-model speech emotion recognition network based on feature clustering," *IEEE Access*, vol. 11, pp. 71224–71234, 2023.
[27] K. L. Ong, C. P. Lee, H. S. Lim, K. M. Lim, and A. Alqahtani, "Mel-MViTv2: Enhanced speech emotion recognition with mel spectrogram and improved multiscale vision transformers," *IEEE Access*, vol. 11, pp. 108571–108579, 2023.
[28] S. Kakuba, A. Poulose, and D. S. Han, "Attention-based multi-learning approach for speech emotion recognition with dilated convolution," *IEEE Access*, vol. 10, pp. 122302–122313, 2022.
[29] N. Saleem, J. Gao, R. Irfan, A. Almadhor, H. T. Rauf, Y. Zhang, and S. Kadry, "DeepCNN: Spectro-temporal feature representation for speech emotion recognition," *CAAI Trans. Intell. Technol.*, vol. 8, no. 2, pp. 401–417, Jun. 2023, doi: 10.1049/CIT2.12233.
[30] S. P. Mishra, P. Warule, and S. Deb, "Improvement of emotion classification performance using multi-resolution variational mode decomposition method," *Biomed. Signal Process. Control*, vol. 89, Mar. 2024, Art. no. 105708.
[31] Y. Zhang et al., "BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1519–1532, Oct. 2022.
[32] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9620–9629.
[33] M. A. Islam, S. Jia, and N. D. B. Bruce, "How much position information do convolutional neural networks encode?" in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–24.
[34] S. Haq and P. J. B. Jackson, *Machine Audition: Principles, Algorithms and Systems, Chapter Multimodal Emotion Recognition*. Hershey, PA, USA: IGI Global, Aug. 2010, pp. 398–423.
[35] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1517–1520.
[36] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.

JEONG-YOON KIM received the Associate degree in electrical and electronics engineering from Chungnam State University and the bachelor's and master's degrees in electronic engineering from Hanbat National University, Daejeon, in 2020 and 2022, respectively, where he is currently pursuing the Ph.D. degree. His current research interests include positional encoding method of vision transformer and attention mechanism-based speech emotion recognition.

SEUNG-HO LEE received the bachelor's, master's, and Ph.D. degrees from the Department of Electronic Engineering, Hanyang University. He is currently a Professor with the Department of Electronic Engineering, Hanbat National University. While operating an image processing/deep learning/augmented reality laboratory, he has published 76 articles, registered 52 patents, and performed 82 research projects. He received the Hanbat National University Outstanding Research Award, in 2016, 2018, and 2021, the Hanbat National University Industry-Academic Cooperation Award, in 2019, and the Hanbat National University President's Award, in 2017, for contributing to the vitalization of industry-university cooperation. In 2018, he was selected as an excellent industry-university-research cooperation expert implemented by the Ministry of SMEs and Startups and was awarded the Minister of SMEs and Startups Award. The Institute of Electronics and Information Engineers Excellent Paper Award, in 2014, 2015, 2017, 2018, 2019, 2020, 2022, and 2023, were awarded the Excellence Paper Award by The Institute of Korean Electrical and Electronics Engineers. Since November 2022, he has been serving as the Vice President for Hanbat National University.