IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 July 2025, accepted 11 August 2025, date of publication 15 August 2025, date of current version 27 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3599218

RESEARCH ARTICLE

Self-Attention-Based Masked Spectrogram Generation and Self-Supervised Learning Method for Improving Speech Emotion Recognition

JEONG-YOON KIM AND SEUNG-HO LEE
Department of Electronic Engineering, Hanbat National University, Daejeon 34158, Republic of Korea
Corresponding author: Seung-Ho Lee (shyolee@hanbat.ac.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) ICAN (ICT Challenge and Advanced Network of HRD) Program funded by Korean Government (MSIT) under Grant IITP-2025-RS-2022-00156212, 100%.

ABSTRACT In this paper, we propose the Self-Attention-based Masked Spectrogram Generation (SAMSG) method to address the problem of model overfitting and improve generalization performance in speech emotion recognition under limited data conditions. A key challenge in many emotional speech datasets is that a small set of fixed sentences is repeatedly uttered with different emotional expressions, which can cause models to overfit to sentence-specific acoustic patterns rather than learn generalizable emotion-related features. To overcome this limitation, the proposed SAMSG method utilizes a pure self-attention-based model (DeiT) to obtain attention maps and applies the attention rollout technique to extract regions of high importance from time-frequency spectrograms. It then selectively masks only the regions that are important for emotion recognition, encouraging the model to learn complementary emotional information from less attended areas. This approach addresses the learning bias commonly seen in self-attention models, which tend to over-focus on localized regions of the input. The originality of the SAMSG method lies in its use of self-attention-driven masking, which—unlike conventional random masking—removes regions the model itself considers important, thereby promoting the learning of more diverse and robust emotional features. Our method alleviates overfitting without requiring external data or large-scale datasets, and achieves strong generalization even in data-constrained environments. Experiments conducted on the SAVEE, EmoDB, and CREMA-D datasets show that the proposed SAMSG method outperforms existing self-attention-based models, achieving accuracies of 94.44%, 96.30%, and 85.94%, respectively. It also attains macro-averaged F1-scores of 0.9401, 0.9692, and 0.8595, demonstrating consistent robustness across diverse emotional speech corpora.

INDEX TERMS Speech emotion recognition, self-attention-based, masked spectrogram generation, self-supervised learning, vision transformer.

I. INTRODUCTION

Speech emotion recognition task aims to identify the emotional state of a speaker from a speech signal, which is important for human-computer interaction, medical, and affective computing.

Some emotional speech datasets exhibit structural characteristics that can inadvertently lead to overfitting. Specific dataset [1] consist of a small set of fixed sentences that are repeatedly uttered with varying emotional states and intensities, such as anger, fear, and happiness. This design can lead models to rely too heavily on sentence-specific acoustic patterns rather than learning robust, emotion-related features distributed across time and frequency. This problem is

The associate editor coordinating the review of this manuscript and approving it for publication was Jolanta Mizera-Pietraszko.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

common in the real world. However, when people lie to hide their emotions, but we can still detect them by the tremors and pitch of their voices that are not hidden. This problem often leads to overfitting, which limits the generalization performance of models, and several previous studies [2], [3], [4] have achieved low accuracies (e.g., 68.81–70.47%) when trained on datasets with these characteristics. These results suggest that existing approaches may struggle to identify subtle emotional cues in situations where the repetitive sentence structure of the dataset makes it easy for the model to remember sentence-level artifacts. Based on these limitations, this study explores whether actively masking regions that the model has carefully observed can mitigate overfitting and facilitate learning complementary emotional information beyond repetitive sentence cues. We use self-attention as a tool for this exploration. Self-attention has demonstrated impressive performance in many fields, including natural language processing (NLP) task [5], [6], [7]. In image classification tasks [8], [9], [10], [11], [12] self-attention allows models to assign higher attention scores to spatial regions that contain semantically significant objects, thereby enabling them to focus on visually meaningful features [13]. In speech-related tasks such as speech recognition [14], [15], [16], [17] and speech emotion recognition [18], [19], [20], [21], self-attention captures the relative importance and relevance of features across both time and frequency. These properties are particularly valuable in speech emotion recognition, where emotional cues are unevenly distributed throughout an audio signal. Self-attention's ability to assign different levels of importance across the spectrogram allows the model to focus dynamically on prosodic variation, timbre, and energy contours—features that are strongly associated with emotion but may occur at any point in the signal. This makes self-attention a powerful tool for detecting subtle yet significant emotional indicators throughout an utterance. These properties of self-attention can also be used as guidance. For instance, Self-Attention Guidance (SAG) enhances attention during denoising processes, guiding the model to focus on informative areas while avoiding irrelevant or noisy information [22]. SAG is particularly effective in uncertain or low-confidence data, helping the model suppress overly confident focus on incorrect regions or avoid completely ignoring relevant but uncertain areas. However, the accuracy of SAG depends on the base model's ability to generate reliable attention maps; if the model misidentifies key regions, the guidance may become inaccurate. We adopt SAG to solve the problem that relies too much on sentence-by-sentence acoustic patterns mentioned above, and propose a new training strategy called Self-Attention-based Masked Spectrogram Generation (SAMSG). The SAMSG leverages self-attention guidance to identify the most important time-frequency regions within the spectrogram and selectively masks them during training. Inspired by masked self-supervised learning, the SAMSG applies masking in a supervised setting using only labeled data, encouraging the model to learn robust representations from less dominant regions and improving generalization. Unlike traditional pretraining-based masking approaches that rely on large amounts of unlabeled data, the SAMSG achieves effectiveness using only the available labeled datasets. Moreover, self-attention models typically require large-scale training data to avoid overfitting or underfitting. SAMSG mitigates this limitation by masking highly attended regions and compelling the model to learn from less salient areas, improving its ability to generalize. As a result, the proposed method supports robust emotion recognition even in data-constrained environments.

The main contributions of this study are summarized as follows:

- We propose the SAMSG, a novel masking strategy guided by self-attention, which identifies and selectively masks emotionally salient time-frequency regions during training.
- Our method applies masking in a purely supervised setting using only labeled data, avoiding the need for large-scale unlabeled datasets typically required by pretraining-based masking approaches.
- By encouraging the model to learn from less dominant regions of the spectrogram, SAMSG improves generalization performance and reduces overfitting in limited-data.
- The proposed SAMSG framework enables robust speech emotion recognition, even when emotional speech data is scarce or imbalanced.

The topics covered in each section of this paper are as follows: Section II reviews related works on recent deep learning methods for speech emotion recognition and related tasks. Section III details the proposed SAMSG method, including subsections on converting speech to log-Mel spectrograms, multi-head self-attention, model architecture, and the proposed SAMSG framework. Section IV presents experimental settings, covering the datasets used, evaluation setup, and ablation studies; it also demonstrates that the proposed SAMSG method is effective not only on datasets characterized by fixed sentences repeatedly uttered with varying emotional states and intensities, but also on general speech emotion recognition datasets. Finally, Section V summarizes the findings and concludes the paper.

II. RELATED WORKS

Self-attention was originally introduced NLP and other sequence modeling tasks. However, the introduction of the Vision Transformer (ViT) [23], which tokenizes images into patches and processes them using self-attention, marked a significant advancement in computer vision. ViT achieved competitive, and in some cases superior, performance compared to Convolutional Neural Networks (CNN) [24]. Nevertheless, Dosovitskiy et al. [23] observed that training ViT on medium-sized datasets such as ImageNet-1k [25] led to suboptimal performance unless it was pretrained on a substantially larger dataset like JFT-300M [26].

To address this, Touvron et al. [27] introduced a knowledge distillation approach in which a ViT (student) learns from a CNN (teacher) via a dedicated distillation token. This method enabled ViT to reach a top-1 accuracy of 83.1% on ImageNet-1k without requiring external data beyond the training set.

Owing to the data-intensive nature of ViT, many audio-related tasks—including automatic speech recognition and speech emotion recognition have adopted ViT pretrained on large-scale image datasets. For instance, Gong et al. [2] applied a pretrained ViT to keyword spotting and sound classification. They found that while a patch size of 128×2 better preserves time–frequency structure for spectrograms, the lack of pretrained models for such configurations forced them to use the standard 16×16 patch size.

To better adapt self-attention to the spectral structure of audio, Ristea et al. [3] proposed the Separable Transformer, which applies distinct attention mechanisms along the time and frequency axes via horizontal and vertical transformer branches. This approach achieved state-of-the-art performance in keyword spotting, sound classification, and speech emotion recognition tasks without requiring external unlabeled data, while using relatively small token sizes, fewer attention heads, and shallower depths compared to standard ViT models.

Other approaches have explored feature engineering and hybrid architectures to improve speech emotion recognition. Kanwal and Asghar [28] introduced a clustering-based genetic algorithm (GA) to select optimal feature subsets from grouped acoustic features, achieving better classification performance than traditional feature sets. Kakuba et al. [29] proposed an attention-based multi-learning architecture combining dilated convolutions and self-attention to effectively capture long-range emotional dependencies in speech. Their model uses spectral and voice quality features extracted from raw speech signals. The design enables parallel processing with a large receptive field while maintaining a relatively low parameter count, and it outperforms conventional CNN and Long Short-Term Memory (LSTM)-based models [30]. Ong et al. [31] extended multiscale ViT by introducing Mel-MViTv2, a ViT-based model tailored for speech emotion recognition with improved stability and accuracy through multiscale mel-spectrogram processing and self-attention.

Zhang et al. [32] proposed PM-SERNet, a parallel-branch speech emotion recognition model that learns from multiple feature spaces and integrates them via attention-weighted aggregation, improving robustness especially for emotionally ambiguous samples. Dal Rì et al. [4] conducted a large-scale benchmark of CNN-based architectures, analyzing their strengths, limitations, and the impact of preprocessing and augmentation strategies on speech emotion recognition performance. Jothimani and Premalatha [33] proposed MFF-SAug, a multi-feature fusion framework with spectrogram augmentation for speech emotion recognition that extracts Mel-Frequency Cepstral Coefficients (MFCC), Zero-Crossing Rate (ZCR), and Root Mean Square (RMS) features, applies silence removal, white-noise injection, and pitch-tuning augmentations, and feeds them into a lightweight 1D CNN, achieving improved result. Mishra et al. [34] explored a multi-resolution variational mode decomposition technique to extract diverse emotional features, such as MFCC and entropy-based descriptors, achieving superior results in speech emotion recognition. Ong et al. [35] proposed MaxMViT-MLP, a dual-path transformer architecture that fuses Constant-Q Transform (CQT) and Mel-STFT spectrograms using MaxViT and MViTv2 backbones, respectively. The outputs are combined via a lightweight multilayer perceptron (MLP), capturing both local and global patterns in time–frequency space for improved emotion recognition.

Radoi et al. [36] proposed an end-to-end temporally aggregated audio-visual network that randomly samples asynchronous audio–video windows across fixed segments to both fuse modality-specific features via lightweight CNN branches and naturally augment limited labeled data, achieving real-time, outstanding result in emotion recognition. Radoi et al. [37] proposed a lightweight CNN-based multimodal emotion recognition framework that reuses the same audio and visual networks across multiple temporal segments and employs an uncertainty-based iterative learning procedure—selecting and annotating the most uncertain samples each epoch—to minimize annotated data requirements; Their proposed method achieves improved result in audio-visual emotion recognition with only 2.7M parameters and real-time inference capability. Goncalves et al. [38] proposed Versatile Audio-Visual Learning (VAVL), a VAVL framework that unifies multimodal and unimodal emotion recognition in a single architecture. An auxiliary unimodal reconstruction task encourages the model to retain modality-specific characteristics while the shared layers learn cross-modal representations. VAVL can be trained on audio-only, visual-only, or paired data and seamlessly switches between categorical classification and continuous attribute regression. Goncalves et al. [39] proposed a joint learning framework that leverages both unimodal (voice-only, face-only) and multimodal (audio–visual) annotations within a single architecture: modality-specific branches are first pre-trained on their respective labels, then frozen and fused via shared conformer layers trained on audio–visual labels, yielding improved F1 scores, better calibration, and reduced bias.

Despite these advancements, deep learning models trained on limited emotional speech data often suffer from overfitting. We observed that such models tend to focus excessively on a small set of highly salient regions in the spectrogram, ignoring more subtle yet semantically important cues. To mitigate this, we propose a new training strategy that leverages attention maps to identify the most critical regions of a spectrogram and mask them during training. Inspired by masked language modeling and masked image modeling

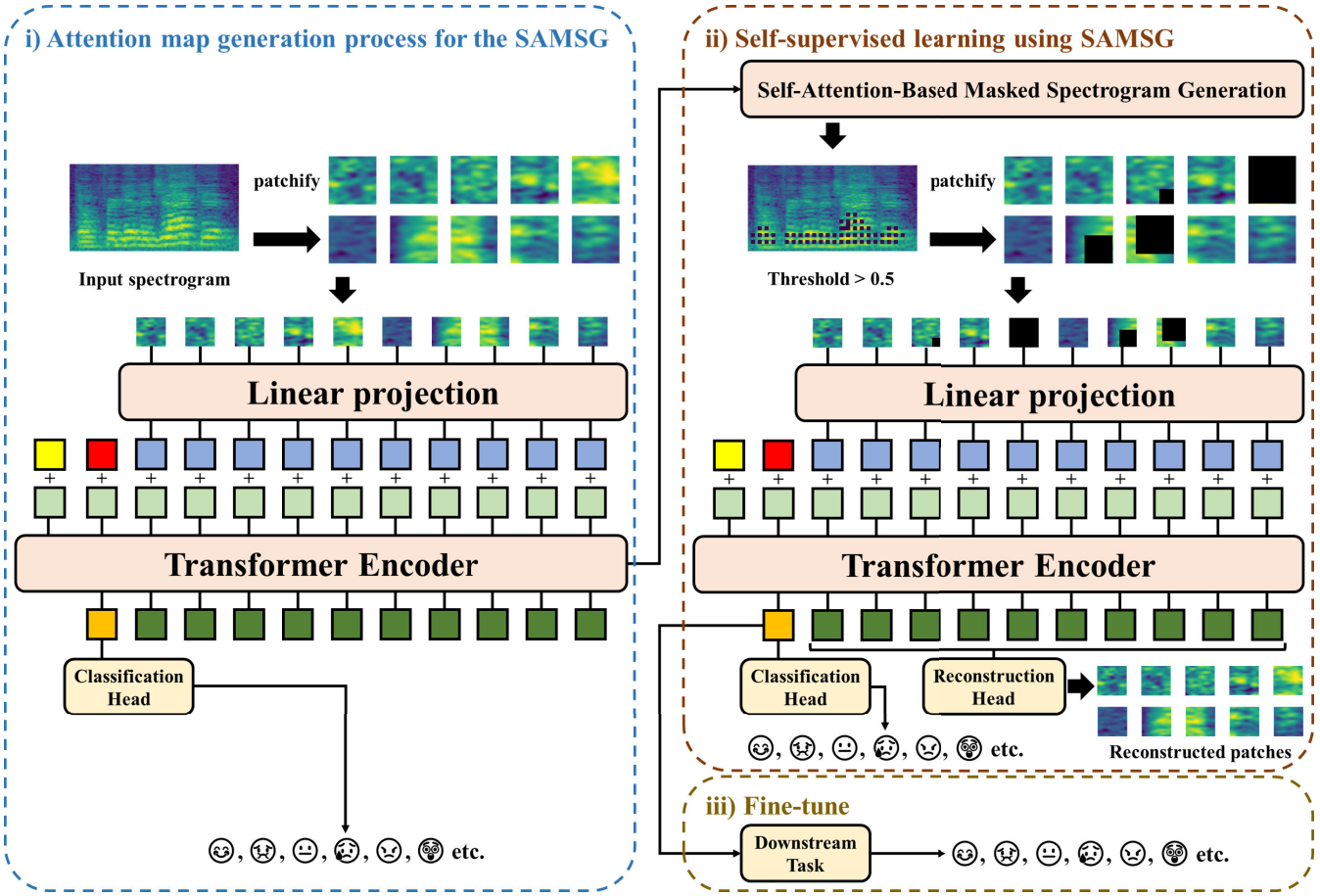

FIGURE 1. The entire process of the SAMSG method proposed in this paper. Yellow represents the CLS token, red represents the distillation token, blue represents the token after linear projection, light green represents the positional embedding, orange represents the CLS token and distillation token added and divided in half, and dark green represents the output token. i) Attention map generation process for the SAMSG: Training self-attention to determine which regions of the spectrogram are important for speech emotion recognition. ii) Self-supervised learning using the SAMSG: By applying attention rollout to the attention map generated in stage i, unimportant areas are removed, values are inverted to generate a mask, and after filtering once more with a threshold value, this is applied to the original image to perform masking. iii) After self-supervised learning, the final speech emotion recognition is performed through fine-tuning, which performs the downstream task with a single linear layer of the same structure as the classification head.

techniques, our method encourages the model to rely less on dominant cues and instead learn richer, more generalized emotional representations.

III. METHOD

In this section, we describe in detail the converting to log-Mel spectrogram, multi-head self-attention, the architecture of the models and the proposed SAMSG method. Figure 1 represent the entire process of the SAMSG method proposed in this paper.

A. CONVERT TO LOG-MEL SPECTROGRAM

In speech emotion recognition, MFCC are widely used due to their effectiveness in capturing perceptually relevant frequency features of speech signals. The MFCC extraction process is grounded in the Mel scale, which reflects the nonlinear characteristics of human auditory perception—particularly, the tendency to perceive pitch logarithmically rather than linearly.

Given an input time-domain speech signal $x(t)$, the Short-Time Fourier Transform (STFT) is first applied to analyze the time–frequency characteristics and obtain a complex spectrogram. Then, a Mel filter bank is used to convert the frequency axis f of the power spectrum to the Mel scale $m(f)$, emphasizing perceptually meaningful frequency bands.

Equations 1 and 2 represent the STFT and the Mel scale transformation, respectively, which form the basis for MFCC computation in our method.

$$X(n,\omega) = \sum_{m=-\infty}^{\infty} x[m] \cdot w[n-m] \cdot e^{-j\omega m} \tag{1}$$

$$m(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

B. MULTI-HEAD SELF-ATTENTION

Self-attention is a mechanism that models pairwise dependencies between positions in a sequence by computing their

mutual relevance. Given a sequence represented by an input matrix $X \in \mathbb{R}^{T\times d}$, where T is the sequence length and d is the feature dimension, self-attention projects X into three matrices—queries Q, keys K, and values V—using learned weight matrices:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V \tag{3}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d\times d_k}$ are trainable parameters, and $Q, K, V \in \mathbb{R}^{T\times d_k}$. The attention output is computed as a weighted sum of the value vectors, where the weights are determined by the scaled dot-product of queries and keys:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{4}$$

This operation allows the model to attend to different parts of the input sequence with varying importance, capturing long-range dependencies across time and frequency when applied to spectrogram representations of speech.

To enhance expressiveness, multi-head self-attention computes multiple sets of attention outputs in parallel, each with its own parameterized projections:

$$\begin{aligned}&\text{MultiHead}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O\\ &\text{where} \quad \text{head}_i = \text{Attention}(XW_i^Q, X\,W_i^K, X\,W_i^V)\end{aligned} \tag{5}$$

Here, h is the number of attention heads, and each projection matrix $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times d_k}$ is learned independently. The output dimension of each head is typically set to $d_k = \frac{d}{h}$, so that the concatenated result maintains the original dimensionality d.

In practical speech emotion recognition applications, the input log-Mel spectrogram is first divided into overlapping patches and linearly projected to form embeddings of shape $B \times N \times D$, where B is the batch size, N is the number of patches, and D is the embedding dimension. These embeddings are then passed through multiple multi-head self-attention layers, followed by pooling and a classification head to perform emotion recognition.

C. ARCHITECTURE OF MODELS

In this paper, we adopt the Data-efficient Image Transformer (DeiT) model as the backbone of the SAMSG framework. DeiT is based on the structure of ViT, but it is designed to be trained effectively with a relatively small amount of data. In particular, DeiT is one of the first models to directly apply the knowledge distillation method to the transformer structure, and is designed to insert distillation tokens learned from the CNN-based teacher model into the transformer so that the student model can effectively absorb the teacher's knowledge through attention. This plays a key role in improving performance without the need for external large-scale data.

In this study, we use the DeiT-base architecture, with embedding token dimension=768, depth=12, and head=12. The input is a log-Mel spectrogram image of size 128×1001, which is divided into 16×16 patches by patch embedding, and then converted into a 768-dimensional embedding vector by linear projection overlapping by 6 pixels. During this patch embedding, positional encoding is added to preserve positional information. The DeiT structure is purely self-attention based, which is effective in selectively capturing important regions across an image (or spectrogram), and this feature combines well with the important region masking strategy proposed in this paper.

SAMSG works by using DeiT's attention map to identify the regions in the spectrogram that the model is most focused on, then masking those regions, and learning from the post-masking reconstruction task. DeiT's strong attention inference capabilities are a key foundation for this self-supervised learning strategy. After learning self-supervised learning with SAMSG, the downstream task is to infer emotion from a classification head consisting of a single linear layer.

D. SELF-ATTENTION-BASED MASKED SPECTROGRAM GENERATION

In this paper, we propose SAMSG that utilizes DeiT to analyze self-attention maps, which improves the generalization performance of the model by masking emotionally salient regions. The entire process is organized as follows: First, train the log-Mel spectrogram based on the DeiT-base model to perform speech emotion recognition. Second, the trained DeiT model is used to collect self-attention maps generated by each transformer layer. Each attention map has the form (h,T,T) for the number of heads h and the number of tokens T, and represents the interaction relationship between input tokens. Third, the collected attention maps are subjected to an attention rollout process based on the methodology of Abnar et al. [13]. Attention rollout accumulates attention information per layer in a self-attention-based model such as ViT, and ultimately estimates the overall attention distribution for the input token. The attention matrix $A_{\text{raw}}^{(l)} \in \mathbb{R}^{N\times N}$ of each layer is normalized through softmax, and the identity matrix I is weighted and averaged to reflect the residual connection, as in Equation 6.

$$A^{(l)} = \alpha A_{\text{raw}}^{(l)} + (1-\alpha)I \tag{6}$$

where, α is a scaling parameter of the attention weight, which usually has a value such as 1 or 0.9 (we use 0.9). The final attention flow $\hat{A}$ from the input token to the output can be calculated by sequentially multiplying the attention $A^{(l)}$ containing these residuals by the matrix, and the entire attention rollout is defined as in Equation 7.

$$\hat{A} = A^{(1)}A^{(2)}\cdots A^{(L)} \tag{7}$$

where, L represents the total number of self-attention layers. This method is particularly useful for visualizing or interpreting the attention flow based on [CLS] tokens, and provides an intuitive understanding of the overall attentional focus area for the model's input.

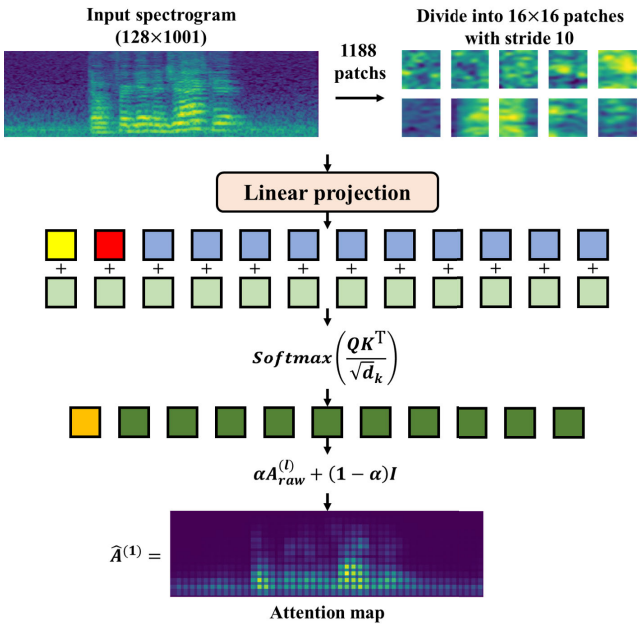

FIGURE 2. Diagram of the attention map generation process in this paper. The input spectrogram (128×1001) is split into patches of 16×16 size, and a stride of 10 is applied to generate a total of 1188 patches. The attention score is calculated as the softmax result based on the inner product between the query and the key. The attention matrix obtained in this way is accumulated through attention rollout, and the final attention map for the time-frequency domain that the model paid attention to in the spectrogram is generated.

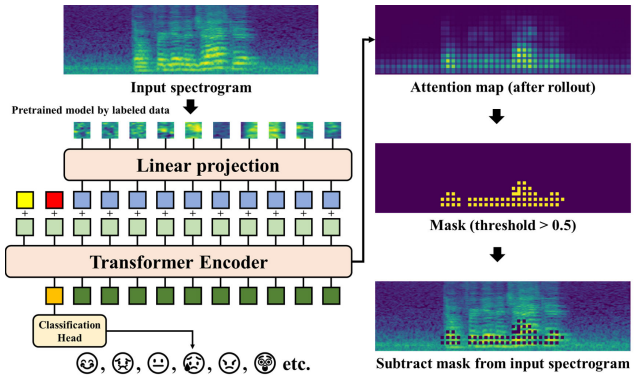

FIGURE 3. The masked spectrogram generation part of the proposed SAMSG method. The input spectrogram goes through the patch segmentation and linear projection processes and is input to the Transformer encoder. After passing through the encoder, an accumulated attention map is generated for the region that the model focused on in the time-frequency domain through attention rollout, and a mask is generated by extracting token locations that are greater than a certain threshold based on the attention value. After that, this mask is applied to the original spectrogram to perform masking that selectively removes only the regions that the model determines to be important.

Figure 2 is a diagram of the attention map generation process in this paper.

Fourth, we apply a threshold of 0.5 (empirically chosen based on rough visual assessment of the masked regions) to the final rollout attention map to identify areas that the model recognizes as important. We extract only tokens (patches) with attention values above 0.5, and consider these regions to be the parts of the input spectrogram that the model relies on. Fifth, we perform masking on the extracted important regions. This is a way to remove areas that the model focuses too much on, encouraging it to learn useful emotional information from a larger area. Masking is done by replacing the important patches with black (0), which produces a partially masked spectrogram. Sixth, the masked spectrogram is used as input to perform self-supervised learning using SAMSG. The output consists of two heads: a classification head for emotion classification and a reconstruction head for restoring masked regions. The classification head and reconstruction head are composed of a single linear layer. This allows the model to simultaneously learn emotion recognition and the secondary task of spectrogram reconstruction. Finally, in the downstream task, we freeze the weights of the self-supervised learning model using SAMSG and train the CLS token and single linear layer of the model to fine-tune it for the speech emotion recognition task. Figure 3 shows the masked spectrogram generation part of the proposed SAMSG method.

IV. EXPERIMENTS

In this section, we describes the speech emotion recognition datasets used in the experiments, the evaluation setup and the ablation study.

A. DATASETS

The SAVEE [40] consists of 480 english spoken sentences performed by four male actors and categorized into seven emotion categories: anger, fear, disgust, happiness, neutral, sadness, and surprise.

The EmoDB [41] contains a total of 535 speech data recorded by 10 German voice actors (5 female and 5 male), with seven emotion labels: anger, anxiety, boredom, disgust, happiness, neutral, and sadness.

The CREMA-D [1] consists of 7,442 videos of 91 actors (48 male and 43 female) of diverse ethnic backgrounds expressing emotions through 12 specific sentences. The dataset includes six emotion categories: anger, fear, disgust, happiness, neutral, and sadness.

Each dataset consists of speech waveforms collected from all participants, and the entire data was partitioned into train set, validation set, and test set in the ratio of 70% : 15% : 15% for the experiment.

B. EVALUATION SETUP

All speech data was normalized to a sampling frequency of 16 kHz, with a length limit of 4 seconds and zero-padding applied to samples shorter than 4 seconds. Then, STFT was applied to convert the speech data into log-Mel spectrograms, with the following parameters: FFT size N = 1024, hop size H = 64, window size = 512, and hamming window. In addition, various data augmentation techniques were applied in this paper to improve the generalization performance of the model. Specifically, we applied noise

perturbation, time shifting, and speed perturbation. The final size of the generated log-Mel spectrogram is 128×1001. The models used in the proposed SAMSG use an Adam optimizer [42] with a learning rate initialized to 1e-5 to train. The learning rate was set to decrease by a factor of 0.5 every 5 epochs to ensure stable convergence. The model was trained for a total of 25 epochs with a batch size of 4. We used cross-entropy loss as the objective function to optimize classification performance across sentiment categories. Following the Gong et al. [43], we used mean squared error loss as the objective function for reconstructing masked regions. All experiments were conducted using the same training configuration to ensure a fair comparison between different attentional mechanisms and architectural variants. The gradual learning rate reduction helps to avoid overfitting and stabilize learning over time, especially when working with limited data samples.

C. ABLATION STUDY

In this subsection, we quantitatively analyze the impact of the proposed SAMSG method on speech emotion recognition performance. By comparing experiments with and without SAMSG, we evaluate whether masking critical regions using self-attention contributes to model generalization and performance improvement. ViT trains the DeiT-base model directly on log-Mel spectrogram input without any masking. The proposed SAMSG method is to pretrain DeiT-base first, then mask important regions through attention rollout, perform masked spectrogram pretraining, and fine-tune the speech emotion recognition task through downstream tasks. We compare our proposed method with state-of-the-art methods. Table 1 shows the accuracy comparison results of the state-of-the-art methods on three datasets: SAVEE, EmoDB and CREMA-D.

TABLE 1. Accuracy comparison results of the state-of-the-art methods on three datasets.

Method	SAVEE↑	EmoDB↑	CREMA-D↑
Zhang et al. [32]	-	88.80%	-
Ong et al. [31]	-	90.57%	-
Ong et al. [35]	-	95.28%	-
Kakuba et al. [29]	93.75%	95.93%	-
Mishra et al. [34]	83.40%	90.51%	-
Dal Rì et al. [4]	-	-	68.22%
Ristea et al. [3]	-	-	70.47%
Jothimani et al. [33]	77.60%	-	66.80%
Radoi et al. [36]	-	-	78.50%
Radoi et al. [37]	-	-	74.20%
Goncalves et al. [38]	-	-	82.60%
DeiT (Gong et al.) [2]	79.17%	88.75%	67.20%
The SAMSG (Ours)	**94.44%**	**96.30%**	**85.94%**

In addition, we compute the macro-averaged F1-score to evaluate the performance of our proposed SAMSG applied model regardless of class imbalance. For each class i, the precision and recall are computed as:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i}, \qquad \text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \tag{8}$$

The F1-score for class i is then defined as:

$$\text{F1}_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{9}$$

Finally, the macro-averaged F1-score is calculated as the arithmetic mean of the F1-scores across all C classes:

$$\text{F1}_{\text{macro}} = \frac{1}{C} \sum_{i=1}^{C} \text{F1}_i \tag{10}$$

Here, TP_i, TN_i, FP_i, and FN_i refer to the number of:

- True Positives (TP): instances correctly predicted as class i
- False Positives (FP): instances incorrectly predicted as class i but belonging to another class
- False Negatives (FN): instances belonging to class i but incorrectly predicted as another class
- True Negatives (TN): instances that neither belong to class i nor are predicted as class i

Macro-averaged F1-score ensures that the metric is not biased toward majority classes and provides a balanced view of the model's performance across all emotion categories.

Table 2 shows the F1-score comparison results of the state-of-the-art methods on three datasets: SAVEE, EmoDB and CREMA-D.

TABLE 2. F1-score comparison results of the state-of-the-art methods on three datasets.

Method	SAVEE↑	EmoDB↑	CREMA-D↑
Zhang et al. [32]	-	0.8900	-
Kakuba et al. [29]	0.9250	0.9583	-
Goncalves et al. [38]	-	-	0.7790
Goncalves et al. [39]	-	-	0.7720
DeiT (Gong et al.) [2]	0.7660	0.8875	0.6730
The SAMSG (Ours)	**0.9401**	**0.9692**	**0.8595**

While recent deep learning approaches have achieved strong performance in speech emotion recognition, most of them focus on maximizing classification accuracy through feature engineering or network design, without directly addressing overfitting and shortcut learning issues that emerge in limited and repetitive datasets.

For example, Kakuba et al. [29] proposed a multi-branch attention-based architecture that combines dilated CNN and self-attention mechanisms to model long-range emotional dependencies in speech. Their model utilizes handcrafted acoustic features such as spectral and voice quality indicators and is optimized for parallel processing and efficiency. However, while effective at capturing temporal context, it does not address potential biases arising from repetitive sentence structures or dominant spectrotemporal patterns in emotional corpora.

Similarly, Jothimani and Premalatha [33] introduced MFF-SAug, a lightweight 1D CNN-based architecture that fuses multiple engineered features (MFCC, ZCR, RMS) and applies augmentation techniques including silence removal, white-noise injection, and pitch tuning. Although this

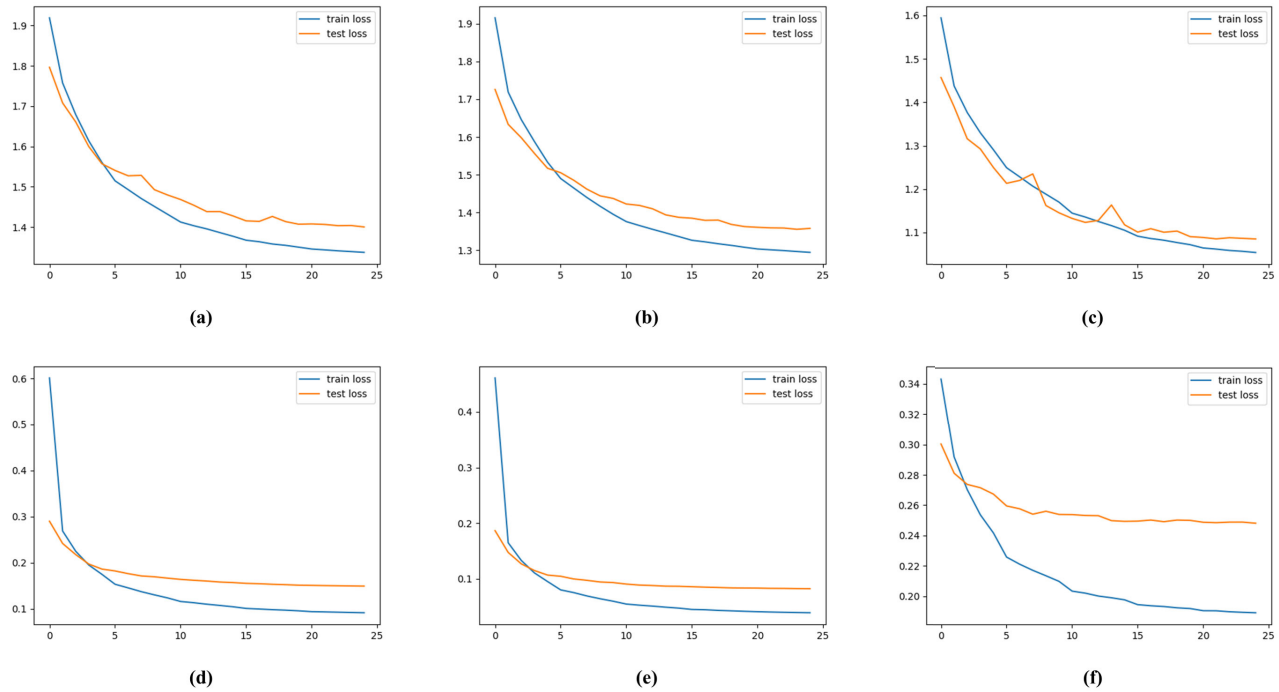

FIGURE 4. The training and testing loss curves of the DeiT model trained on (a) SAVEE, (b) EmoDB, and (c) CREMA-D, as well as the SAMSG applied DeiT model trained on (d) SAVEE, (e) EmoDB, and (f) CREMA-D. In the baseline DeiT (a–c), the models converge gradually, with both training and test loss steadily decreasing. SAVEE and EmoDB (a, b) show visible gaps between training and test loss, suggesting mild overfitting. CREMA-D (c) shows slightly better alignment between the two curves, though fluctuations in the test loss around epoch 7–10 indicate instability due to the dataset's higher complexity. After applying SAMSG (d–f), the loss curves demonstrate consistently faster convergence across all datasets. On SAVEE (d) and EmoDB (e), both training and test losses decrease rapidly and remain low, with minimal generalization gap—reflecting that these datasets are relatively easier for emotion classification. On CREMA-D (f), SAMSG also improves convergence speed and lowers training loss, but the test loss plateaus at a higher value, showing a moderate gap between training and test performance. This suggests that while SAMSG enhances generalization, challenges in CREMA-D—such as greater inter-speaker variability and sentence diversity—still pose difficulties.

improves data diversity and model robustness, the architecture remains dependent on low-dimensional handcrafted inputs and lacks mechanisms for analyzing or mitigating attention bias during training.

On the multimodal side, Goncalves et al. [38] proposed VAVL, a transformer-based framework that unifies audio-only, visual-only, and audio–visual emotion recognition in a single architecture. It integrates modality-specific conformer encoders with shared transformer layers and leverages an auxiliary reconstruction task to preserve unimodal characteristics. Later, Goncalves et al. [39] extended this idea with a two-stage training scheme, using pre-trained acoustic and visual encoders followed by frozen fusion through shared conformer layers trained only on multimodal labels. These transformer-based approaches improve generalization by leveraging multi-source supervision and cross-modal representation learning, but they do not explicitly investigate how internal attention distributions might guide or bias emotional reasoning.

In contrast, our proposed SAMSG framework is built upon a ViT backbone, and introduces a novel self-attention-guided masking mechanism. During training, SAMSG analyzes attention maps to identify spectrogram regions receiving high attention—often dominated by sentence-specific acoustic patterns—and masks them to encourage the model to learn from less salient but potentially more generalizable emotional cues. Unlike pretraining-based masking methods that require large unlabeled corpora, SAMSG operates fully in a supervised setting, using only labeled emotional data. Moreover, this approach reframes attention not only as a computational mechanism, but as a diagnostic and corrective signal to mitigate overfitting and dataset-specific bias.

Therefore, while previous works utilize CNN, LSTM, and Transformers primarily to improve representation capacity or cross-modal fusion, SAMSG contributes a complementary perspective by promoting representation accountability—specifically, encouraging models to learn emotional patterns that are robust across varying contexts, rather than overly relying on dataset artifacts. This makes it particularly suitable for real-world scenarios involving small, imbalanced, or structurally biased emotional speech datasets.

Figure 4 presents the training and testing loss curves of the DeiT model trained on (a) SAVEE, (b) EmoDB, and (c) CREMA-D, as well as the SAMSG applied DeiT model trained on (d) SAVEE, (e) EmoDB, and (f) CREMA-D. The curves highlight that integrating SAMSG leads to more stable

convergence and reduced overfitting, particularly in datasets prone to shortcut learning due to repeated sentence structures.

While the proposed SAMSG framework has demonstrated notable improvements in mitigating overfitting and enhancing generalization in speech emotion recognition tasks, it also presents several limitations that should be considered.

First, the SAMSG method requires an initial supervised training phase to obtain attention maps from a labeled dataset before applying masking, which inevitably increases the overall training time compared to standard models. Additionally, the need for both attention-guided masking and subsequent retraining can effectively double computational and memory requirements, posing practical constraints for deployment in resource-limited environments such as embedded or mobile devices. Second, the effectiveness of SAMSG is sensitive to the threshold used for generating masks from attention rollout results. Determining an optimal threshold often requires empirical tuning, which may not generalize well across different datasets or tasks. Third, SAMSG's performance is fundamentally dependent on the quality of the attention maps produced by the initial model. If the base model fails to generate meaningful or reliable attention distributions, the masking process may inadvertently remove informative regions, negatively impacting the learning of robust emotional features.

Figure 5 illustrates the confusion matrices of the proposed model trained and evaluated on (a) SAVEE, (b) EmoDB, and (c) CREMA-D, showing the effectiveness of SAMSG in enhancing emotion classification performance even in dataset challenging scenarios involving repeated sentence structures.

In Figure 5, the confusion matrix (a) of the model trained with the SAVEE dataset shows that the prediction accuracy is very high in most emotion classes, and in particular, neutral and sadness recorded 18 and 9 correct predictions, respectively. Overall, there is almost no confusion between classes, and the prediction errors are very limited, showing that the SAVEE dataset was able to distinguish emotions well due to its clear emotional expression and balanced distribution. The confusion matrix (b) of the model trained based on the EmoDB dataset shows high accuracy in the anger, boredom, happiness, and neutral classes, achieving 18, 12, 10, and 12 correct predictions, respectively. However, some confusion occurs in the disgust and fear classes, and in particular, the boundary between fear and happiness is unclear. This suggests that EmoDB is based on German speech, and the subtle differences between emotions are more subtly reflected in pronunciation, which may have made it somewhat difficult for the model to distinguish them. The confusion matrix (c) of the model trained based on CREMA-D shows the model's prediction results for a total of six emotion classes (anger, disgust, fear, happiness, neutral, sadness). The model shows very high accuracy in classes such as fear, happiness, and sadness, recording more than 167 correct predictions each. On the other hand, the anger class tends to be somewhat confused with disgust,

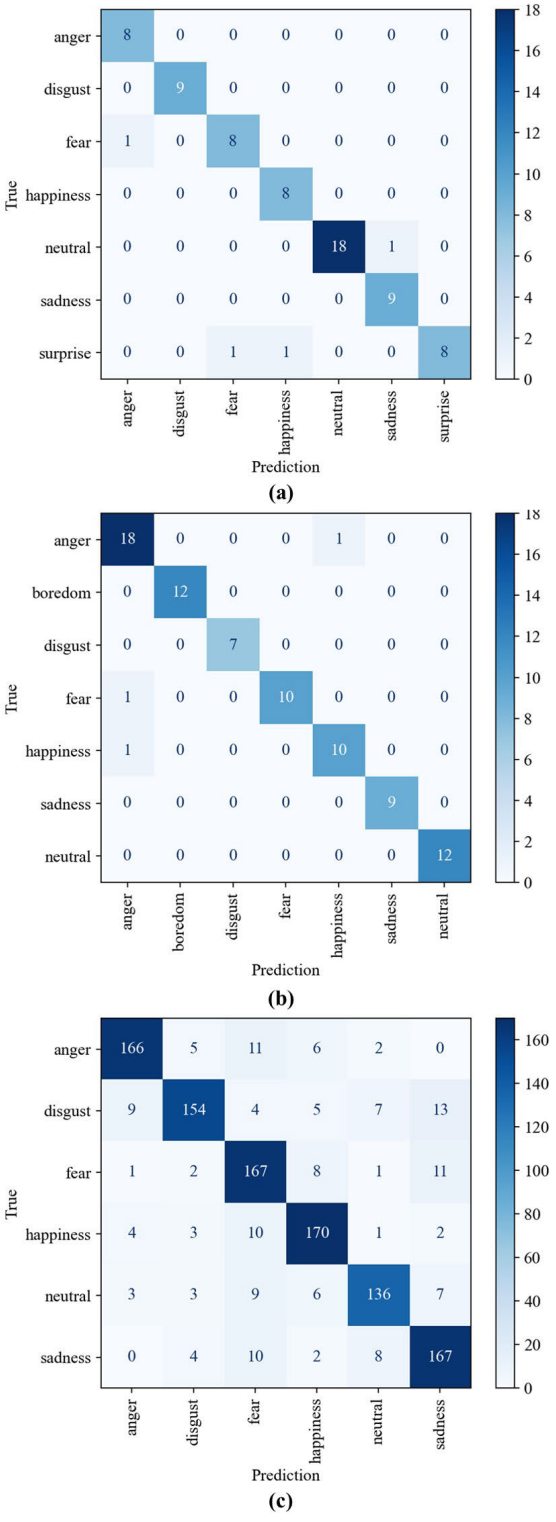

FIGURE 5. The confusion matrices of the proposed model trained and evaluated on (a) SAVEE, (b) EmoDB, and (c) CREMA-D. The model trained on SAVEE shows minimal confusion and high accuracy across all classes. The EmoDB-based model performs well overall but shows confusion between fear and happiness. The model trained on CREMA-D achieves high accuracy for most emotions but exhibits greater confusion, particularly in the anger and neutral classes, likely due to the dataset's larger size and the use of fixed sentences across emotions.

fear, and happiness, and the neutral class shows a pattern of being evenly confused with various emotions. These results suggest that the CREMA-D dataset is larger than the other two datasets and has high similarity because the same sentences are uttered regardless of emotion, making classification relatively more difficult.

V. CONCLUSION

In this paper, we proposed the SAMSG (Self-Attention-based Masked Spectrogram Generation), a novel training strategy designed to improve the generalization ability of self-attention-based models for speech emotion recognition. SAMSG selectively masks highly attended regions in spectrogram representations during training, encouraging the model to explore alternative emotional cues that are less dependent on repetitive or sentence-specific acoustic patterns.

Extensive experiments conducted on three widely used emotional speech datasets—SAVEE, EmoDB, and CREMA-D—demonstrate that SAMSG consistently improves model performance. Specifically, it achieves higher accuracy and macro-averaged F1-scores than baseline DeiT models, especially on datasets where sentence repetition and limited speaker diversity often lead to shortcut learning. Loss curve analyses further confirm that SAMSG promotes more stable and faster convergence, with smaller generalization gaps between training and test loss.

While SAMSG is particularly effective on relatively simpler datasets such as SAVEE and EmoDB, it also contributes to performance improvement on more challenging datasets like CREMA-D, indicating its robustness across varying levels of dataset complexity.

Future work may explore extending SAMSG to other modalities (e.g., multimodal emotion recognition) or integrating it with adaptive masking strategies that dynamically adjust to dataset characteristics.

REFERENCES

[1] 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.

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

[3] N. C. Ristea, R. T. Ionescu, and F. S. Khan, "SepTr: Separable transformer for audio spectrogram processing," in *Proc. Interspeech*, Sep. 2022, pp. 4103–4107.

[4] F. A. D. Rí, F. C. Ciardi, and N. Conci, "Speech emotion recognition and deep learning: An extensive validation using convolutional neural networks," *IEEE Access*, vol. 11, pp. 116638–116649, 2023.

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

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[7] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," Tech. Rep., 2019.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[10] J. J. Bird and A. Lotfi, "CIFAKE: Image classification and explainable identification of AI-generated synthetic images," *IEEE Access*, vol. 12, pp. 15642–15650, 2024.

[11] S. Li, P. Kou, M. Ma, H. Yang, S. Huang, and Z. Yang, "Application of semi-supervised learning in image classification: Research on fusion of labeled and unlabeled data," *IEEE Access*, vol. 12, pp. 27331–27343, 2024.

[12] Q. Pan, K. Liu, S. Zheng, and G. Wang, "A fine-grained image classification method based on ConvNeXt heatmap localization and contrastive learning," *IEEE Access*, vol. 13, pp. 80123–80132, 2025.

[13] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," 2020, *arXiv:2005.00928*.

[14] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[15] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 12449–12460.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 28492–28518.

[17] G. Woo Lee, H. Kook Kim, and D.-J. Kong, "Knowledge distillation-based training of speech enhancement for noise-robust automatic speech recognition," *IEEE Access*, vol. 12, pp. 72707–72720, 2024.

[18] M. A. Jalal, R. Milner, and T. Hain, "Empirical interpretation of speech emotion perception with attention based model for speech emotion recognition," in *Proc. Interspeech*, Oct. 2020, pp. 4113–4117.

[19] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7342–7346.

[20] S. Sadok, S. Leglaive, and R. Séguier, "A vector quantized masked autoencoder for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. Workshops (ICASSPW)*, Jun. 2023, pp. 1–5.

[21] T. Rajapakshe, R. Rana, S. Khalifa, B. Sisman, B. W. Schuller, and C. Busso, "EmoDARTS: Joint optimization of CNN and sequential neural network architectures for superior speech emotion recognition," *IEEE Access*, vol. 12, pp. 110492–110503, 2024.

[22] S. Hong, G. Lee, W. Jang, and S. Kim, "Improving sample quality of diffusion models using self-attention guidance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7428–7437.

[23] 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. Int. Conf. Learn. Represent.*, 2020.

[24] E. Akleman, "Deep learning," *Computer*, vol. 53, no. 9, pp. 1–17, Sep. 2020.

[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] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 843–852.

[27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jegou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[28] S. Kanwal and S. Asghar, "Speech emotion recognition using clustering based GA-optimized feature set," *IEEE Access*, vol. 9, pp. 125830–125842, 2021.

[29] 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.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[31] K. L. Ong, C. P. Lee, H. S. Lim, K. M. Lim, and A. Alqahtani, "MelMViTv2: Enhanced speech emotion recognition with mel spectrogram and improved multiscale vision transformers," *IEEE Access*, vol. 11, pp. 108571–108579, 2023.

[32] 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.
[33] S. Jothimani and K. Premalatha, "MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," *Chaos, Solitons Fractals*, vol. 162, Sep. 2022, Art. no. 112512. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960077922007147
[34] 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.
[35] K. L. Ong, C. P. Lee, H. S. Lim, K. M. Lim, and A. Alqahtani, "MaxMViT-MLP: Multiaxis and multiscale vision transformers fusion network for speech emotion recognition," *IEEE Access*, vol. 12, pp. 18237–18250, 2024.
[36] A. Radoi, A. Birhala, N.-C. Ristea, and L.-C. Dutu, "An end-to-end emotion recognition framework based on temporal aggregation of multimodal information," *IEEE Access*, vol. 9, pp. 135559–135570, 2021.
[37] A. Radoi and G. Cioroiu, "Uncertainty-based learning of a lightweight model for multimodal emotion recognition," *IEEE Access*, vol. 12, pp. 120362–120374, 2024.
[38] L. Goncalves, S.-G. Leem, W.-C. Lin, B. Sisman, and C. Busso, "Versatile audio-visual learning for emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 16, no. 1, pp. 306–318, Jan. 2025.
[39] L. Goncalves, H.-C. Chou, A. N. Salman, C.-C. Lee, and C. Busso, "Jointly learning from unimodal and multimodal-rated labels in audio-visual emotion recognition," *IEEE Open J. Signal Process.*, vol. 6, pp. 165–174, 2025.
[40] S. Haq and P. Jackson, "Multimodal emotion recognition," in *Machine Audition: Principles, Algorithms and Systems*. Hershey, PA, USA: IGI Global, Aug. 2010, pp. 398–423.
[41] 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.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[43] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-supervised audio spectrogram transformer," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 10, pp. 10699–10709.

JEONG-YOON KIM received the Associate's 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 2022. 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. Since November 2022, he has been the Vice President of Hanbat National University, where he is currently a Professor with the Department of Electronic Engineering. 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, 2021, and 2024, 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. He received 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.