IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 7, 2020, accepted December 17, 2020, date of publication January 1, 2021, date of current version January 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3048675

Polyphonic Sound Event Detection Based on Residual Convolutional Recurrent Neural Network With Semi-Supervised Loss Function

NAM KYUN KIM, (Student Member, IEEE), AND HONG KOOK KIM, (Senior Member, IEEE)
School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology, Gwangju 61005, South Korea

Corresponding author: Hong Kook Kim (hongkook@gist.ac.kr)

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government [Ministry of Science and ICT (MSIT)] (Development of AI Technology for Early Screening of Infant/Child Autism Spectrum Disorders Based on Cognition of the Psychological Behavior and Response) under Grant 2019-0-00330.

ABSTRACT Polyphonic sound event detection (SED) is an emerging area with many applications for smart disaster safety, security, life logging, etc. This paper proposes a two-stage polyphonic SED model when strongly labeled data are limited but weakly labeled and unlabeled data are available. The first stage of the proposed SED model is constructed by a residual convolutional recurrent neural network (RCRNN)-based mean teacher model with convolutional block attention module (CBAM)-based attention. Then, the second stage fine-tunes the student model from the first stage by applying the proposed semi-supervised loss function to accommodate the noisy targets of weakly labeled and unlabeled data. The proposed SED model is applied to both Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Challenge Task 4 and DCASE 2020 Challenge Task 4, and its performance is compared with those of the baseline and top-ranked models from both challenges by measuring the F1-score and polyphonic sound detection score (PSDS). The experiments show that the RCRNN-based first-stage model with CBAM-based attention achieves a higher F1-score and PSDS than the baseline and top-ranked models for both challenges. Furthermore, the proposed two-stage SED model with the semi-supervised loss function improves the F1-score by 6.1% and 4.6% compared to the top-ranked models from DCASE 2019 and 2020, respectively.

INDEX TERMS Convolutional block attention module, polyphonic sound event detection, residual convolutional recurrent neural network, semi-supervised loss function, unlabeled data, weakly labeled data.

I. INTRODUCTION

Sounds contain important information in our daily lives and help us to perceive auditory scenes according to individual sound events occurring around us. Sound event detection (SED) aims to classify specific sound events in diverse sound environments and to detect the onset and offset times of each sound event. SED could affect a wide range of applications associated with sound sensing [1]. For example, acoustic monitoring could detect physical events, such as glass breaking, a gun firing, tires skidding, or a car crashing. SED can also be incorporated into audio captioning for understanding social media content in more detail [2], audio monitoring in smart cities [3], life assistance and healthcare [4], etc.

Many approaches for constructing a SED model rely upon strongly labeled data. Over the last decade, machine learning-based SED methods have been proposed, such as the support vector machine (SVM) [5] and Gaussian mixture model (GMM)-hidden Markov model (HMM) [6], [7], in which audio signals are parameterized using mel-frequency cepstral coefficients. Recently, deep neural network-based approaches for SED have become mainstream due to the advances of computing power and significant improvement of deep learning compared to machine learning or statistical approaches. Different neural network architectures, such as fully connected neural networks [8], convolutional neural networks (CNNs) [9], [10], recurrent neural networks (RNNs) [11], and convolutional recurrent neural networks

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

(CRNNs) [12], [13], have been applied to SED. These SED models predict a strong label for each analysis frame and then detect the onset and offset times of a sound event, referred to as a timestamp, according to the predicted labels.

In general, the supervised learning described above requires a sufficient amount of sound data with strong labels to train an SED model. Moreover, such training data should be collected in a real environment in which a target application using SED could be deployed [1]. However, it is time consuming to give strong labels to all training data because strong labels should include the event type and the timestamp of each sound event, which results in a limited amount of strongly labeled data. An alternative is the combination of limited strongly labeled data with an ample amount of weakly labeled data whose labels only include the sound event types without any information on the timestamps of the events [15], [16]. Therefore, collecting weakly labeled data is much easier and cheaper than collecting strongly labeled data.

To handle weakly labeled data for SED, a popular approach is based on the multiple instance learning (MIL) framework [17]. In the MIL framework, weakly labeled data are processed by a CNN or an RNN to obtain a time-dependent feature map, which is then used to construct an instance-level prediction layer that is further pooled into a bag-level prediction layer. The loss function is the binary cross-entropy (BCE) between the output of the bag-level prediction layer and the target event label. In the inference stage, single or multiple event types included in the input audio clip are predicted by the bag-level prediction layer output while the instance-level predictions provide the timestamp of each predicted event.

Instead of only using strongly labeled data, weakly labeled data can be additionally used to construct an SED model based on supervised and weakly semi-supervised learning (SWSL) [18]. In one study, SWSL was good as the unified framework for handling strongly and weakly labeled data, but the graph search required considerable computations to realize the loss function [19].

In addition to strongly and weakly labeled data, unlabeled data can be used to improve the prediction accuracy of SED. One of the approaches using strongly labeled, weakly labeled and unlabeled data is mean teacher learning [20]. The mean teacher learning-based SED method has two parallel models in the MIL framework, where the student model is trained with the target labels that are predicted by the teacher model. Then, the teacher model is also updated according to the weight changes of the student model for each epoch. In particular, the loss function of the mean teacher model is a composition of different functions according to the data types. In other words, the loss function for strongly labeled data is the cross-entropy between the output of the instance-level prediction layer and the timestamp with the target event label for strongly labeled data. In addition, the loss function for weakly labeled data is the cross-entropy between the output of the instance-level prediction layer and the target label for weakly labeled data. To treat unlabeled data, the loss function includes the mean squared error (MSE) between the output of the bag-level prediction layer and that of the teacher model and the MSE between the output of the instance-level prediction layer and that of the teacher model. Note here that the MSEs are computed for all training data, including strongly labeled, weakly labeled, and unlabeled data. For the inference, only the student model is used, and the prediction is made as in the conventional MIL framework. This mean teacher learning-based model was introduced as a baseline in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2019 Challenge Task 4 [21], where a CRNN was used to represent a feature map in the MIL framework [17]. As described above, the outputs of the mean teacher model could provide noisy labels and timestamps for weakly labeled and unlabeled data because the teacher model is not perfect. To handle noisy targets, a soft bootstrapping technique that discounts the noisy target contribution in the loss function was introduced [22].

Even though these approaches described heretofore improved the SED performance by adding weakly labeled and unlabeled data to strongly labeled data, several avenues can improve the performance of SED. First, the feature map for a mean teacher model can be improved by replacing the CRNN with another architecture. Each sound event is characterized by its loudness, pitch, perceived duration, timbre and spaciousness [23]. Moreover, the sound events belonging to the same sound class might have different characteristics. Thus, to obtain a better feature map for all the sound events, the number of convolutional blocks should be increased compared to the baseline mean teacher model [21]. In this case, a residual block prior to each convolutional block enables us to train the neural network by overcoming the vanishing gradient issue [24]. Consequently, instead of a CRNN, a residual convolutional recurrent neural network (RCRNN) is newly proposed to improve the feature map for the mean teacher model. In addition, the feature map can be further improved by incorporating an attention mechanism into the convolutional layers by suppressing the unimportant features for the feature map [25]. In particular, the proposed RCRNN-based mean teacher model is designed by incorporating the convolutional block attention module (CBAM) in [25].

Second, instead of directly utilizing the RCRNN-based mean teacher model for SED, this paper proposes a two-stage approach for SED. The RCRNN-based mean teacher model provides a lower detection accuracy for weakly labeled and unlabeled data than for strongly labeled data because the predicted labels for weakly labeled and unlabeled data are error-prone, as indicated by [22]. Therefore, the first stage of the proposed two-stage SED model, which is the RCRNN-based mean teacher model, acts as a pretrained model to generate labels for weakly labeled and unlabeled data. Then, the second stage is constructed using the same network architecture of the student model from the first-stage mean teacher model by following the knowledge distillation technique [26]. Consequently, the second stage provides a fine-tuning model to perform SED.

Third, the first stage predicts noisy labels and timestamps for weakly labeled and unlabeled data. In the second stage, the target labels for the weakly labeled and unlabeled data are noisy or erroneous; thus, a semi-supervised loss function is also proposed here to accommodate such noisy target labels in the second stage.

The effectiveness of the proposed two-stage SED model with the semi-supervised loss function is evaluated on the DCASE 2019 and 2020 Challenge Task 4 datasets by measuring the event-based F1-score and polyphonic sound detection score (PSDS) [27]. In addition, the performance of the proposed SED model is compared with those of conventional ones, including the CRNN-based SED method [21], [28], which is the baseline model for DCASE 2019 and 2020 Challenge Task 4, and the SED models that ranked first in both challenges [29], [30].

The remainder of the paper is organized as follows. Following this introduction, Section II briefly reviews the mean teacher model that is the baseline for DCASE 2019 and 2020 Challenge Task 4. Then, Section III proposes the two-stage SED model, focusing on the RCRNN-based mean teacher model in the first stage of the proposed SED model and the semi-supervised loss function for the fine-tuning network in the second stage of the proposed SED model. Next, Section IV evaluates the performance of the proposed SED model and compares it with those of the baseline and top-ranked SED models for DCASE 2019 and 2020 Challenge Task 4. Finally, Section V concludes this paper.

II. CRNN-BASED MEAN TEACHER MODEL FOR SED

Fig. 1 shows the schematic diagram of the baseline SED model for DCASE 2019 and 2020 Challenge Task 4 [21], [28], where the network architectures for the teacher and student models of the mean teacher model are based on CRNN, and they are identical. As shown in the figure, the CRNN of both the teacher model and the student model stacks three different neural networks: a CNN, an RNN, and a fully connected (FC) layer. In this CRNN structure, the CNN, which is composed of a series of convolutional modules with a convolutional layer and a pooling layer, plays the role of a feature extractor. The numbers of convolutional modules are set to three and seven for DCASE 2019 and 2020, respectively. In addition, a gated linear unit (GLU) is used for the activation function in each convolutional layer for both DCASE 2019 and 2020. The pooled feature map from the last CNN module is used as the input of the bidirectional gated recurrent unit (BiGRU), which is one of the RNNs. Here, two BiGRUs are used to form a time-dependent feature map, and the output of the second BiGRU is connected with an FC layer, resulting in the 2-dimensional (2-D) output whose dimension is the number of frames by the number of sound events to be detected in SED. This 2-D output becomes a form of a strong label because it has timestamp information according to each sound event. Finally, the 2-D output is further processed by an aggregation layer that is implemented

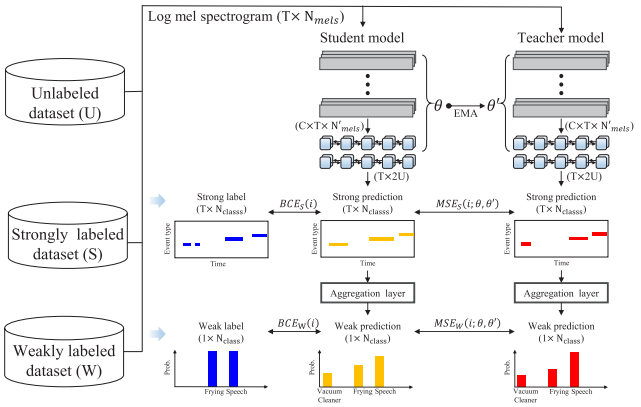

FIGURE 1. Schematic diagram of the CRNN-based mean teacher model.

by attention pooling, which results in a weak label (i.e., target label without a timestamp) for the input audio clip.

To train the student model in the mean teacher model, a loss function is defined by the sum of four different functions:

$$\begin{aligned} &L_{MeanTeacher}\,(\theta) \\ &=\sum_{i\in S} BCE_S\,(i;\theta) + \sum_{i\in W} BCE_W\,(i;\theta) \\ &\quad +\lambda \sum_{i\in\{S,W,U\}} MSE_S\left(i;\theta,\theta'\right) + \lambda \sum_{i\in\{S,W,U\}} MSE_W\left(i;\theta,\theta'\right) \end{aligned} \tag{1}$$

where i denotes the i-th training audio clip and θ and θ' are the student and teacher models, respectively. In addition, S, W, and U indicate the sets of strongly labeled, weakly labeled, and unlabeled data, respectively. The first two equations on the right-hand side in (1) correspond to two BCE functions: one is the BCE between the target label and the output from the FC layer for the strongly labeled training data, and the other is the BCE between the target label and the output of the aggregation layer for the weakly labeled training data. Here, the BCE is defined as

$$BCE\,(i;\theta) = -\left(y_i\,log\,\hat{y}_{i,\theta} + (1-y_i)\,log\left(1-\hat{y}_{i,\theta}\right)\right) \tag{2}$$

where y_i and $\hat{y}_{i,\theta}$ are the target label and the output of the student model, θ, for the i-th audio clip, respectively. Therefore, $BCE_S\,(i;\theta)$ and $BCE_W\,(i;\theta)$ in (1) are defined as in (2) for strongly labeled data and weakly labeled data, respectively.

To compute the MSEs in the other two functions in (1), the outputs of the FC layer from the teacher model, θ', are assumed to be strong labels composed of the target event type and timestamps for a given audio clip while the outputs of the aggregation layer from θ' correspond to weak labels predicting only target event types. Then, the MSE between the two outputs from θ and θ' is computed for strong labels and weak labels, respectively. The MSE for the i-th audio clip is defined as

$$MSE\left(i;\theta,\theta'\right) = \left\|\hat{y}_{i,\theta} - \hat{y}_{i,\theta'}\right\|_2^2. \tag{3}$$

In other words, the MSE loss functions distill knowledge from the teacher to the student [26]. Finally, the contribution of the MSE over the BCE is controlled by a hyperparameter, λ.

After computing the loss function, the parameters of the student model, θ, are updated according to the error back-propagation. In addition, the parameters of the teacher model, θ', are updated by the exponential moving average (EMA) of the pre-epoch parameters of the teacher model with the newly updated parameters of the student model, which means that the teacher model parameters are updated slightly toward the student model parameters. After terminating the training, the student model is finally used to predict the event type and timestamp for each test audio clip in the baseline model.

The mean teacher model described so far provides a good solution for unlabeled data training because there is no target label regarding the event type and timestamp for each unlabeled audio clip [20]. However, the outputs from the teacher model, which act as the target labels for the MSE loss functions, are noisy, so the performance of the student model is highly dependent on the degree of noisiness of the teacher model outputs. Therefore, in this paper, a semi-supervised loss function will be defined in Section III-B to discount the noisiness for the weakly labeled and unlabeled data. In addition, the network architecture of the mean teacher model will be changed from the CRNN to the RCRNN with attention to provide a better feature map, which will be described in Section III-A.

III. PROPOSED TWO-STAGE POLYPHONIC SED MODEL

This section proposes a two-stage SED model with strongly labeled, weakly labeled, and unlabeled data. As shown in Fig. 2(a), the first stage of the proposed SED model trains a mean teacher model that is composed of the proposed RCRNN architecture with CBAM-based attention, where the training procedure follows the approach described in Section II. After finishing the RCRNN-based mean teacher model, the student model of the mean teacher model is taken as a pre-trained model for the second stage of the proposed SED model. Compared to the first-stage training, the second-stage training is done with the instance-level predictions in the MIL framework for all the strongly labeled, weakly labeled, and unlabeled data. To obtain instance-level predictions, the student model of the first stage is used to decode instance-level predictions for the weakly labeled and unlabeled data, as shown in Fig. 2(b). Note here that the strongly labeled data have their corresponding strong target labels with a form of instance-level predictions. The target labels for the weakly labeled and unlabeled data are noisy or erroneous; thus, a semi-supervised loss function is also proposed here to accommodate such noisy target labels in the second stage. A more detailed explanation of the RCRNN with attention and the semi-supervised loss function will be given in the following subsections.

TABLE 1. Network architecture of a residual convolutional neural network in the proposed RCRNN.

Name	Layers	Output shape
Input layer	Input: log-mel spectrogram	1×628×128
Stem block	$(7 \times 7, \text{Conv2D}, @16, \text{GLU}, \text{BN})$ 2×2 pooling layer	16×314×64
	$(7 \times 7, \text{Conv2D}, @32, \text{GLU}, \text{BN})$ 2×2 pooling layer	32×157×32
Residual convolutional block	$\begin{pmatrix} 3 \times 3, \text{Conv2D}, @64, \\ \text{ReLU, BN} \end{pmatrix} \times 2$ Self-attention module (CBAM) 1×2 pooling layer	64×157×16
	$\begin{pmatrix} 3 \times 3, \text{Conv2D}, @128, \\ \text{ReLU, BN} \end{pmatrix} \times 2$ Self-attention module (CBAM) 1×2 pooling layer	128×157×8
	$\begin{pmatrix} 3 \times 3, \text{Conv2D}, @128, \\ \text{ReLU, BN} \end{pmatrix} \times 2$ Self-attention module (CBAM) 1×2 pooling layer	128×157×4
	$\begin{pmatrix} 3 \times 3, \text{Conv2D}, @128, \\ \text{ReLU, BN} \end{pmatrix} \times 2$ Self-attention module (CBAM) 1×2 pooling layer	128×157×2
	$\begin{pmatrix} 3 \times 3, \text{Conv2D}, @128, \\ \text{ReLU, BN} \end{pmatrix} \times 2$ Self-attention module (CBAM) 1×2 pooling layer	128×157×1

A. FIRST STAGE: PROPOSED RCRNN-BASED MEAN TEACHER MODEL

Fig. 3 shows a network architecture of the proposed RCRNN with CBAM-based attention. First, as described in [28], the signal of a given audio clip is downsampled from 44.1 kHz to 16 kHz and segmented into consecutive frames of 2,048 samples with a 255-sample hop size. Then, a 2,048-point fast Fourier transform (FFT) is applied to each frame, and the magnitude spectrum is converted into 128-dimensional log mel-filterbanks. Next, 628 frames are grouped to make a (628×128) spectral image, which is then used as the input feature to the proposed residual CNN (RCNN) with attention. As shown in the figure, the RCNN is composed of one stem block and five residual convolutional blocks, where the stem block consists of two convolutional blocks with 16 and 32 kernels for the first and second convolutional blocks, respectively. Each convolutional block has (7×7) kernels with a stride of (1×1), and it is followed by batch normalization, GLU activation, and a (2×2) average pooling layer.

Fig. 4 shows a residual convolutional block with CBAM-based attention [25]. As shown in the figure, each residual convolutional block is composed of two convolutional layers followed by batch normalization and rectified linear unit (ReLU) activation; it also has the shortcut connection for the residual learning [24]. The detailed hyperparameters for the residual convolutional block are listed in Table 1. Next, CBAM-based attention [25] is applied to the output of each residual convolutional block, F, such as

$$F' = M_c(F) \otimes F, \tag{4}$$

$$F'' = M_s(F') \otimes F' \tag{5}$$

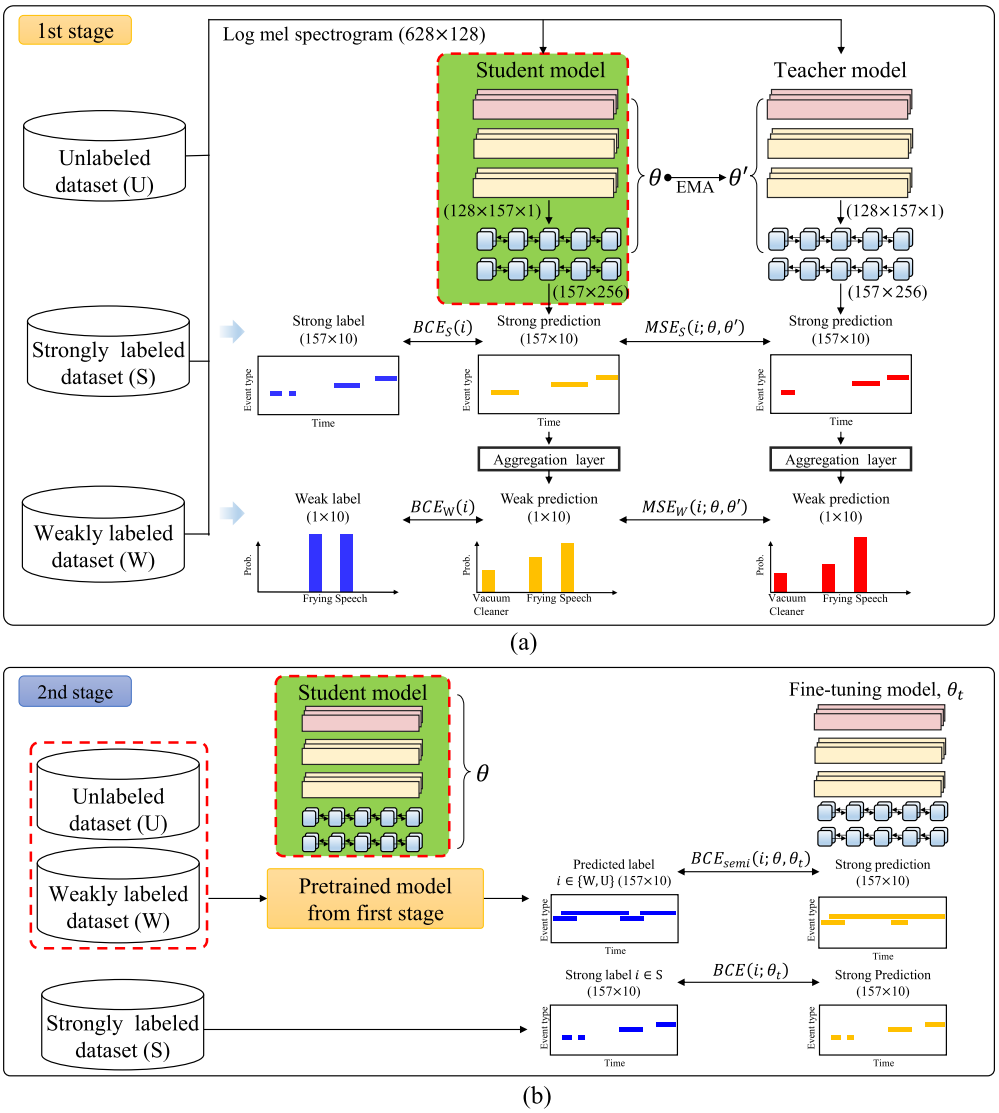

FIGURE 2. Block diagram of the proposed two-stage SED model: (a) the first stage based on the RCRNN-based mean teacher model and (b) the second stage using the proposed semi-supervised loss function.

where ⊗ denotes elementwise multiplication. In other words, the channel attention, $M_c(\cdot)$, is first applied to F, and then the spatial attention, $M_s(\cdot)$, is applied to the output of the channel attention, F'. The channel attention in (4) is calculated by the equation

$$M_c(F) = \sigma\left(MLP\left(AvgPool_c(F)\right) + MLP\left(MaxPool_c(F)\right)\right) \tag{6}$$

where σ denotes a sigmoid function and MLP is a multilayer perceptron with one hidden layer. In addition, $AvgPool_c(\cdot)$ and $MaxPool_c(\cdot)$ are the average pooling and max pooling functions over the spatial dimension, respectively. Similarly, the spatial attention is applied as

$$M_s(F) = \sigma\left(f\left(\left[AvgPool_s(F); MaxPool_s(F)\right]\right)\right) \tag{7}$$

where f represents a convolution operation with (7×7) convolution filters. Moreover, $AvgPool_s(\cdot)$ and $MaxPool_s(\cdot)$ represent the average pooling and max pooling operations over the channel dimension, respectively. Finally, the channel-spatially refined feature map, F'', is further processed by a (1×2) pooling layer.

After finishing all the residual convolutional blocks, the $(128 \times 157 \times 1)$ feature map is applied to a recurrent block, as shown in Fig. 4. The recurrent block consists of two BiGRUs to learn the temporal context information, where ReLU is used as an activation function for each GRU. The (157×256) output of the recurrent block is processed by an FC layer and then by a sigmoid function, resulting in a (157×10) output, where 10 denotes the number of sound events to be detected. Consequently, the input dimension of 628 is reduced to the output dimension of 157; thus, the

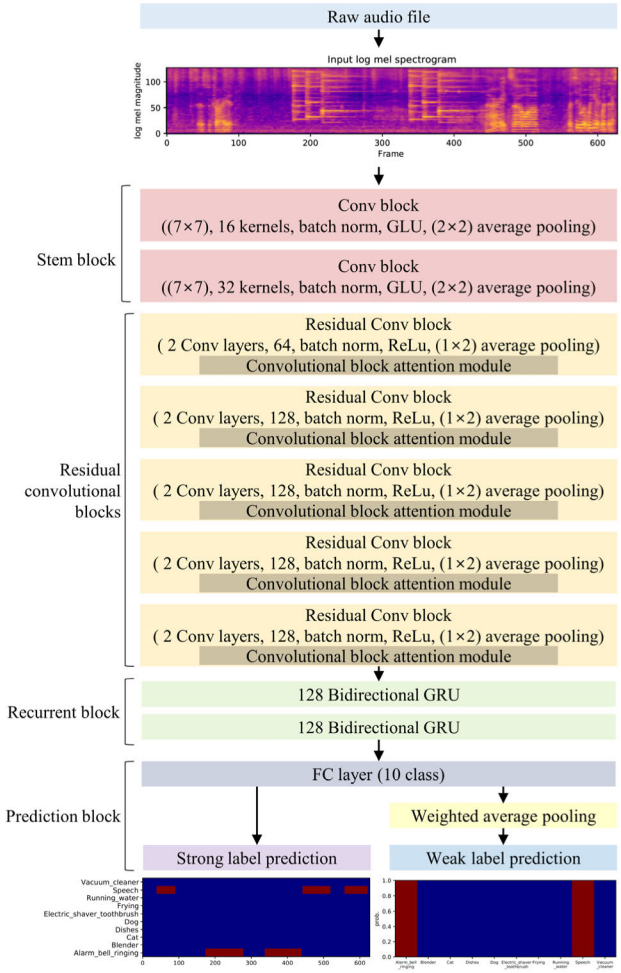

FIGURE 3. Network architecture of the proposed RCRNN-based mean teacher model for the first stage of the proposed SED model.

time resolution of the output corresponds to 1,020 samples. Note that an (157 × 10)-dimensional output is related to a strong label including the sound event type and timestamp. Moreover, a weighted pooling layer is applied to the (157 × 10)-dimensional output to obtain an (1 × 10)-dimensional output that predicts a weak label for the given audio clip.

This RCRNN-based first stage of the proposed SED model is trained according to the loss function defined in (1) for all the training data composed of strongly labeled, weakly labeled, and unlabeled data. After finishing the model training, the student model of the RCRNN-based mean teacher model is brought to the second stage of the proposed SED model.

B. SECOND STAGE: FINE-TUNING WITH SEMI-SUPERVISED LOSS FUNCTION

As mentioned earlier, the student model of the first stage provides the predicted target labels for each audio clip from weakly labeled or unlabeled data, where the prediction is performed to give only strong labels, as shown in Fig. 3. In other words, while the student model can predict weak

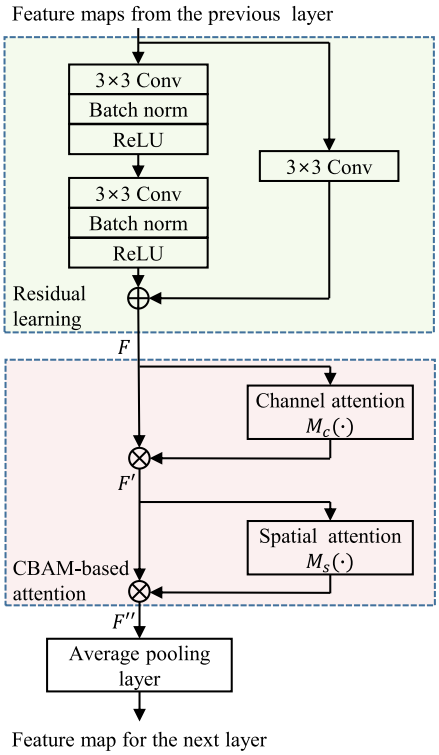

FIGURE 4. Block diagram of a residual convolutional block with CBAM-based attention.

labels and strong labels as its output, only strong labels are taken as the target labels for the fine-tuning model, as shown in the right part of Fig. 2(b). Notice that the correct target labels for strongly labeled data are already given.

The network architecture of the fine-tuning model in this paper is identical to that of the student model, as shown in Fig. 4, except that there is no network path for the weak label prediction because predicted strong labels for weakly labeled or unlabeled data are available from the pretrained student model. To train the fine-tuning model, a loss function should be defined. First, since strong labels are available regardless of strongly labeled, weakly labeled, and unlabeled data, a loss function can be given by modifying (1) as

$$L(\theta) = \sum_{i \in \{S,W,U\}} BCE_S(i;\theta). \tag{8}$$

However, the target labels for computing (8) are noisy or erroneous for the weakly labeled or unlabeled data while they are given as references for strongly labeled data. It is known that a soft bootstrapping approach can improve the prediction performance when the target is noisy [22]. Inspired by this approach, a semi-supervised loss function is proposed here.

For the semi-supervised loss function, the strong labels predicted from the student model for weakly labeled and unlabeled data are first binarized as

$$\bar{y}_{i,\theta} = \begin{cases} 0 & \hat{y}_{i,\theta} < th_\delta \\ 1 & otherwise \end{cases} \tag{9}$$

where θ denotes the student model and th_δ is a threshold for the binarization that is set to 0.5 as in [8]. Next, a semi-supervised loss function is defined by using both the BCE for strongly labeled data and soft bootstrapping for weakly labeled and unlabeled data, such as

$$L_{semi} = \sum_{i \in S} BCE(i; \theta_t) + \sum_{i \in \{W,U\}} BCE_{soft}(i; \theta_t) \quad (10)$$

where θ_t denotes the fine-tuning model of the second stage of the proposed SED model, as shown in the right part of Fig. 2(b). In addition, $BCE\ (i;\theta_t)$ is defined as in (2), and $BCE_{soft}(i;\theta_t)$ is the BCE between the binarized target from (9) and the predicted output from θ_t. In other words, $BCE_{soft}(i;\theta_t)$ is defined as

$$BCE_{soft}\ (i;\theta_t) = -(\bar{y}_i\ log\ \hat{y}_{i,\theta_t} + (1-\bar{y}_i)\ log\ \left(1 - \hat{y}_{i,\theta_t}\right) \quad (11)$$

where $\hat{y}_{i,\theta_t}\hat{y}_{i,\theta_t}$ is the output of the fine-tuning model, θ_t, for the i-th audio clip. In (11), $\bar{y}_i$ is an interpolated target between the binarized strong labels in (9), and it is computed as

$$\bar{y}_i = \beta \hat{y}_{i,\theta} + (1 - \beta)\, \hat{y}_{i,\theta_t} \quad (12)$$

where β $(0 < \beta \leq 1)$ is an interpolation parameter to control the influence of weakly labeled and unlabeled data on the predicted target label. As β in (12) is set to be close to one, it implies that the predicted target label from the student model is reliable. However, β is set to be smaller than one if the binarized predicted target label is noisy. In the following section, the performance evaluation depending on different settings of β is presented.

IV. PERFORMANCE EVALUATION

A. DATASET

The performance of the proposed two-stage RCRNN-based SED model was evaluated on two different tasks: DCASE 2019 Challenge Task 4 and DCASE 2020 Challenge Task 4. The two tasks focused on large-scale sound detection in domestic environments, and there were 10 sound event types including speech, dog, cat, alarm/bell/ringing, dishes, frying, blender, running motor, vacuum cleaner, and electric shaver/toothbrush sound events [21].

Table 2 describes the data distributions of the training set, development set, and evaluation set for DCASE 2019 Challenge Task 4 and DCASE 2020 Challenge Task 4. As shown in the table, DCASE 2019 Challenge Task 4 included three datasets: 1) a weakly labeled dataset without timestamps, 2) an unlabeled in-domain dataset without any labels, and 3) a strongly labeled synthetic dataset. The weakly labeled and the unlabeled in-domain datasets were taken from the AudioSet dataset [31] while the strongly labeled dataset was generated using the Scaper soundscape synthesis and augmentation library [32]. The numbers of audio clips were 14,412, 1,578, and 2,045 in the unlabeled, weakly labeled, and strongly labeled datasets, respectively. Each audio clip was stored as both mono- and stereo-channel signals that were sampled at 44.1 kHz with a maximum duration of 10 seconds. DCASE 2020 Challenge Task 4 also had three datasets as in DCASE 2019 Challenge Task 4. The unlabeled and weakly labeled datasets of DCASE 2020 Challenge Task 4 were identical to those of DCASE 2019 Challenge Task 4 while the number of audio clips for the strongly labeled dataset was increased from 2,045 to 2,584. In this paper, 20% of the strongly labeled data were used to validate the neural network models during training while 80% of the strongly labeled data with all the weakly labeled and unlabeled data were used to train the models. Note here that the SED models applied to DCASE 2019 Challenge Task 4 and DCASE 2020 Challenge Task 4 were trained using only the training sets denoted in the second and third rows of Table 2, respectively.

TABLE 2. Comparison of the datasets between DCASE 2019 Challenge Task 4 and DCASE 2020 Challenge Task 4.

	Dataset	DCASE 2019 Task 4	DCASE 2020 Task 4
Training set	Strongly labeled dataset	2,045	2,584
	Weakly labeled dataset	1,578	1,578
	Unlabeled dataset	14,412	14,412
Dev set	Validation test dataset	1,168	1,168
Eval set	Evaluation test dataset	692	-

For the performance evaluation of the SED models, two datasets were prepared: a validation test dataset, referred to as the Dev set; and an evaluation test dataset, referred to as the Eval set. The Dev set was composed of 1,168 strongly labeled audio clips that contained 4,093 sound events for both DCASE 2019 and 2020 Challenge Task 4. The Eval set was only available for the evaluation of DCASE 2019 Challenge Task 4, and it had 692 strongly labeled audio clips with 2,765 sound events [21].

B. EXPERIMENTAL SETUP

To train the proposed SED model, the neural network weights of the mean teacher model in the first stage were initialized by using Xavier initialization [33], but the biases were all initialized to zero. Next, the mini-batchwise adaptive moment estimation (ADAM) optimization algorithm [34] was applied, where dropout was also applied at a rate of 0.5 [35]. In addition, the learning rate was set according to the ramp-up strategy [36], where the maximum learning rate reached 0.001 after 50 epochs.

For the second stage of the proposed SED model, the student model from the mean teacher model of the first stage was fine-tuned using 5-fold cross-validation that divided all the data in the training set into 5 folds, where 4 out 5 folds were used for training and the remaining fold was used for validation. Here, the learning rate was initially set to 0.001 and it was reduced by a simple learning rate schedule when the semi-supervised loss function defined in (10) plateaued on the validation set (commonly known as ReduceLROnPlateau in PyTorch) [37]. After finishing the 5-fold cross-validation training, the 5-fold models were linearly combined to form an ensemble classifier. In this paper, the hyperparameter, β, in (12) was set from 0.3 to 0.9 at a step size of 0.2 in order to examine the effect of different degrees of influence of weakly labeled and unlabeled data on the predicted target label.

The performances of the proposed SED model applied to DCASE 2019 and 2020 Challenge Task 4 were first compared with those of the baseline models released by DCASE 2019 Challenge Task 4 [21] and DCASE 2020 Challenge Task 4 [28]. In addition to the baselines, the performances of the top-ranked models throughout the challenges were compared. The top-ranked models in the DCASE 2019 and DCASE 2020 Challenge Task 4 were based on a teacher-student model with guided learning [29] and a mean teacher model with a transformer [30], respectively.

All the neural network models in this paper were implemented using a deep learning package in Python 3.6.9 with PyTorch 1.6.0, and the training and evaluation of the models were conducted on an Intel Core i7-7700 workstation with an NVidia GTX 1080ti GPU. However, instead of running the baselines and top-ranked models in DCASE 2019 and 2020, their performances were taken from the reports corresponding to each model [29], [30].

C. EVALUATION METRICS

The performance of SED models was compared in objective measures, such as F1-score, error rate (ER), and PSDS. The F1-score was defined as [38]

$$\text{F1} - \text{score} = \frac{2 \cdot \text{Precison} \cdot \text{Recall}}{\text{Precison} + \text{Recall}} \tag{13}$$

where $\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$ and $\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$. Here, TP, FP, and FN are the numbers of true positive, false positive, and false negative samples, respectively. A higher F1-score implies better detection performance for SED. In fact, an event-based F1-score was first calculated following [21]. Then, the macro average F1-score was computed by averaging the F1-scores across all sound event types.

The ER measured the number of errors in terms of insertions (I), deletions (D), and substitutions (S), and it was defined as [38]

$$\text{ER} = \frac{\sum_i S(i) + \sum_i D(i) + \sum_i I(i)}{\sum_i N(i)} \tag{14}$$

where $I(i)$, $D(i)$, $S(i)$, and $N(i)$ are the numbers of inserted, deleted, substituted, and ground truth sound events in the i-th audio clip, respectively. Note that a lower ER indicates better SED performance.

Last, PSDS was defined as the normalized area under the polyphonic sound event detection–receiver operating characteristic (PSD-ROC) curve, $r(e)$, as [27]

$$PSDS = \frac{1}{e_{max}} \int_0^{e_{max}} r(e)\, de. \tag{15}$$

To compute $r(e)$, the effective FP rate (eFPR) and effective TP rate (eTPR) were computed as [27]

$$\text{eFPR} : e_c^* \triangleq R_{\text{FP},c} + \alpha_{\text{CT}} \frac{1}{|C| - 1} \sum_{\substack{\hat{c} \in C \\ \hat{c} \neq c}} R_{\text{CT},c,\hat{c}}, \tag{16}$$

$$\text{eTPR} : r(e) \triangleq \mu_{\text{TP}}(e) - \alpha_{\text{ST}} \sigma_{\text{TP}}(e) \tag{17}$$

TABLE 3. Comparison of the F1-scores and ERs between the baseline of DCASE 2019 Challenge Task 4 and the first stage of the proposed RCRNN-based SED model with or without CBAM-based attention.

Model	Dev set		Eval set	
	F1-score	ER	F1-score	ER
Baseline of DCASE 2019 [21]	23.7	-	29.0	-
First stage of the proposed model (RCRNN w/o CBAM)	42.6	1.19	48.8	0.95
First stage of the proposed model (RCRNN with CBAM)	46.8	1.12	51.4	0.92

where $R_{\text{FP},c}$ is FP for the c-th sound type and $R_{\text{CT},c,\hat{c}}$ is the cross-trigger (CT) rate. In particular, $R_{\text{FP},c}$ and $R_{\text{CT},c,\hat{c}}$ correspond to the false alarm rate in the classification and the substitution ER of an event type, c, as $\hat{c}$, respectively. By differently setting α_{CT} in (16), the eFPR could reflect the effect of substitution over the overall false alarms. In (17), $\mu_{\text{TP}}(e)$ and $\sigma_{\text{TP}}(e)$ were the mean and standard deviation of TP across all sound types, respectively. α_{ST} controlled the degree of the confidence intervals. Moreover, e_{max} in (15) was set to the maximum value of e_c^*. Actually, the computation of PSDS was done by using a public open-source [27] where the parameters of $(\rho_{\text{DTC}}, \rho_{\text{GTC}}, \rho_{\text{CTTC}}, \alpha_{\text{CT}}, \alpha_{\text{ST}})$ were set to (0.5, 0.5, 0.3, 0, 0) for PSDS and (0.5, 0.5, 0.3, 1.0, 0) for PSDS CT.

D. EXPERIMENTAL RESULTS ON DCASE 2019 TASK 4

First, the performance of the first stage of the proposed two-stage SED model was compared with that of the DCASE 2019 baseline. Compared to the baseline, the first stage of the proposed SED model was characterized by residual convolutional layers and CBAM-based attention. Table 3 compares the eventwise F1-scores of both models. Note here that the ER for the baseline was not available. As shown in the table, the F1-scores of the RCRNN-based mean teacher model without CBAM-based attention of the proposed SED model were improved by 18.9% and 19.8% for the Dev set and Eval set, respectively, compared to the CRNN as the baseline. Moreover, applying CBAM-based attention to the RCRNN achieved further improvements of 23.1% and 22.4% for the Dev set and Eval set, respectively, compared to the baseline. From then on, the performance of the proposed SED model was used when the CBAM-based attention was applied to the first stage.

Next, the proposed two-stage SED model including the first and second stages was compared with the top-ranked model in DCASE 2019 Challenge Task 4. To examine the effect of the semi-supervised loss function on the performance of the proposed SED model, the interpolation parameter, β, in (12) was differently set from 0.3 to 0.9 at a step size of 0.2. Note that $\beta = 1.0$ made the semi-supervised loss function, L_{semi}, in (10) be the same as $L(\theta)$ in (8) with binarized target values. Table 4 compares the F1-scores and ERs of the proposed SED models according to different βs and the top-ranked model. Here, the performance of each of the proposed SED models with RCRNN was obtained by averaging the F1-scores and ERs over 5-fold cross-validation

TABLE 4. Comparison of the F1-scores and ERs between the top-ranked SED model of DCASE 2019 Challenge Task 4 and the proposed RCRNN-based SED model with different semi-supervised loss functions.

Model	Dev set		Eval set	
	F1-score	ER	F1-score	ER
Top-ranked model of DCASE 2019 Task 4 [29]	44.5	-	45.5	-
RCRNN, β=0.3	50.1	0.99	55.5	0.81
RCRNN, β=0.5	**50.6**	0.99	**56.1**	0.81
RCRNN, β=0.7	49.8	1.00	55.5	0.81
RCRNN, β=0.9	49.9	0.99	54.4	0.83
RCRNN, β=1.0	48.9	1.00	52.7	0.80

TABLE 5. Comparison of the F1-scores and ERs between the top-ranked ensemble SED model of DCASE 2019 Challenge Task 4 and the proposed RCRNN-based ensemble SED model with different semi-supervised loss functions.

Model	Dev set		Eval set	
	F1-score	ER	F1-score	ER
Top-ranked model of DCASE 2019 Task 4 (6 model ensemble) [29]	45.3	-	47.7	-
RCRNN, β=0.3 (5 model ensemble)	**52.3**	0.93	55.6	0.8
RCRNN, β=0.5 (5 model ensemble)	52.2	0.94	**56.2**	0.8
RCRNN, β=0.7 (5 model ensemble)	50.6	0.98	55.3	0.8
RCRNN, β=0.9 (5 model ensemble)	51.7	0.94	55.0	0.8
RCRNN, β=1.0 (5 model ensemble)	51.1	0.94	54.0	0.81

while the performance of the top-ranked model was set as the top-1 in the literature [29]. As shown in the table, among the proposed SED models with different βs, the proposed SED model with $\beta = 0.5$ achieved the highest F1-score and the lowest ER. Moreover, compared to the top-ranked model of DCASE 2019, the proposed SED model increased the F1-scores by 6.1% and 10.6% for the Dev set and Eval set, respectively. In addition, the semi-supervised loss function with $0 < \beta < 1$ provided better performance than the supervised loss function with $\beta = 1$. Finally, the proposed SED model was constructed by ensembling single models from 5-fold cross-validation, and its performance was compared with that of the ensemble version of the top-ranked model. As shown in Table 5, all the proposed SED models with different βs always had higher F1-scores than the top-ranked model. In particular, the best performance was achieved for the proposed SED model when $\beta = 0.3$ and $\beta = 0.5$ for the Dev set and Eval set, respectively. Therefore, it is concluded that the proposed SED model, which was characterized by the RCRNN structure, CBAM-based attention, and the semi-supervised loss function, achieved better performance than the state-of-the-art model for DCASE 2019 Challenge Task 4.

E. EXPERIMENTAL RESULTS ON DCASE 2020 TASK 4

In this subsection, the proposed SED model was applied to DCASE 2020 Challenge Task 4, and its performance was compared with those of the baseline and the top-ranked model throughout the challenge. Table 6 compares the event-wise F1-scores of the different SED models, such as the

TABLE 6. Comparison of the F1-scores and ERs between the baseline, top-ranked SED model of DCASE 2020 Challenge Task 4 and the proposed RCRNN-based SED model with different semi-supervised loss functions.

Model	Dev set	
	F1-score	ER
Baseline of DCASE 2020 [28]	34.8	-
Top-ranked model of DCASE 2020 Task 4 [30]	46.0	-
CRNN, β=0.5 [39] (single model)	42.5	1.14
First stage of the proposed model (RCRNN with CBAM)	46.8	1.13
RCRNN, β=0.3	50.2	1.00
RCRNN, β=0.5	50.0	1.01
RCRNN, β=0.7	**50.6**	**0.98**
RCRNN, β=0.9	49.1	1.01
RCRNN, β=1.0	49.9	1.00

TABLE 7. Comparison of the F1-scores and ERs between the top-ranked ensemble SED model of DCASE 2020 Challenge Task 4 and the proposed RCRNN-based ensemble SED model with different semi-supervised loss functions.

Model	Dev set	
	F1-score	ER
Top-ranked model of DCASE 2020 Task 4 (6 model ensemble) [30]	50.6	-
CRNN, β=0.5 [39] (5 model ensemble)	45.2	1.05
RCRNN, β=0.3 (5 model ensemble)	51.4	0.97
RCRNN, β=0.5 (5 model ensemble)	50.8	0.98
RCRNN, β=0.7 (5 model ensemble)	**51.6**	**0.95**
RCRNN, β=0.9 (5 model ensemble)	50.3	0.97
RCRNN, β=1.0 (5 model ensemble)	50.6	0.97

DCASE 2020 Challenge Task 4 baseline, the top-ranked model in DCASE 2020 Challenge Task 4, the first stage of the proposed SED model, and the proposed SED models with different βs. In addition, the performance of our SED model when we participated in DCASE 2020 Challenge Task 4 [39] was compared. Compared to the proposed SED model in this paper, our SED model, which was the version that participated in DCASE 2020 Challenge Task 4, used the CRNN for the mean teacher model and applied the semi-supervised loss function without any binarization to the noisy targets. Note here that the ERs for the baseline and top-ranked SED models were not available. As shown in the table, the first stage of the proposed SED model provided a higher F1-score than the baseline and top-ranked models. Moreover, it significantly improved our previous version of the SED model. In particular, the proposed two-stage SED models improved the F1-score by more than 3.8% regardless of the setting of β, and the best performance was achieved when $\beta = 0.7$.

Second, the ensemble versions of the proposed SED and top-ranked model in DCASE 2020 were compared, as shown in Table 7. Consequently, it was shown that the proposed SED model with $\beta = 0.7$ had the best F1-score and ER, and it increased the F1-score by 1.0% compared to that of the top-ranked model.

Third, the F1-score was decomposed depending on the event class in order to investigate the effectiveness of the proposed SED model on each sound event class. Table 8 compares the F1-scores for each sound event class, where the performances of the first-stage-only and two-stage SED models

TABLE 8. Comparison of the F1-scores according to 10 different audio event types between the baseline and proposed SED models on DCASE 2020 Challenge Task 4.

Event Type \ Model	Baseline	Proposed	
		First-stage only	Two-stage
Alarm/bell/ringing	35.5	48.5	50.4
Blender	30.6	46.2	60.2
Cat	35.0	45.0	49.5
Dishes	24.8	33.1	38.2
Dog	24.2	32.8	37.3
Electric shaver/toothbrush	33.0	53.1	59.2
Frying	33.7	50.0	54.0
Running water	29.2	42.5	42.5
Speech	52.7	61.5	65.3
Vacuum cleaner	49.8	55.4	59.2

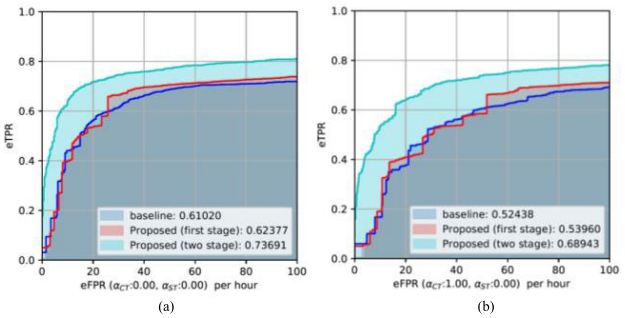

FIGURE 5. Comparison of the PSD-ROCs and areas, where (a) (α_{CT}, α_{ST}, e_{max}) = (0, 0, 100) and (b) (α_{CT}, α_{ST}, e_{max}) = (1, 0, 100).

were taken when $\beta = 0.7$ in Tables 6 and 7, respectively. As shown in the table, the first-stage-only and the two-stage proposed SED models improved the F1-scores for all classes compared to the baseline. As expected, the proposed two-stage SED model outperformed the first-stage-only model for all the classes except the running water class.

Fourth, the PSD-ROCs were depicted using the toolkit [27] to examine the performance of the proposed SED model across all possible operating points. Fig. 5 shows the

TABLE 9. Comparison of the PSDSs between the baseline and proposed SED models on DCASE 2020 Challenge Task 4.

Model	PSDS	PSDS-CT
DCASE 2020 baseline [28]	0.610	0.524
First stage of proposed model (RCRNN with CBAM)	0.624	0.542
RCRNN, β=0.7	0.737	0.689

PSD-ROCs averaged over all sound event classes for two different cases: 1) average TP vs. average FP by setting (α_{CT}, α_{ST}) = (0, 0) in (16) and (17) and 2) average TP vs. FP with a penalty for misclassification by setting (α_{CT}, α_{ST}) = (1, 0), where the x-axis was discretized into 100 intervals by setting $e_{max} = 100$. In addition, Table 9 compares the PSDSs by integrating the PSD-ROCs for the baseline, the first stage of the proposed SED model, and the proposed two-stage SED model, respectively. As shown in the table, the proposed two-stage SED model achieved the highest PSDS for both cases. In particular, all eTPRs of the proposed two-stage SED model were higher than those of the baseline, as shown in Fig. 5. This implies that the proposed two-stage SED model could provide better detection performance than the baseline by setting an arbitrary decision threshold.

Finally, to examine how effectively the feature map is constructed when using the proposed RCRNN, visualization using gradient-weighted class activation mapping (Grad-CAM) [40] was performed. Fig. 6 compares the feature map constructed by the CRNN in the baseline and that constructed by the RCRNN without/with CBAM-based attention in the proposed SED model. The figure shows that the RCRNN, regardless of the implementation of attention, provided a clearer map than the CRNN. By investigating the areas indicated by dotted boxes, the feature map constructed by applying CBAM-based attention had a higher correlation to the strong labels compared to that without attention.

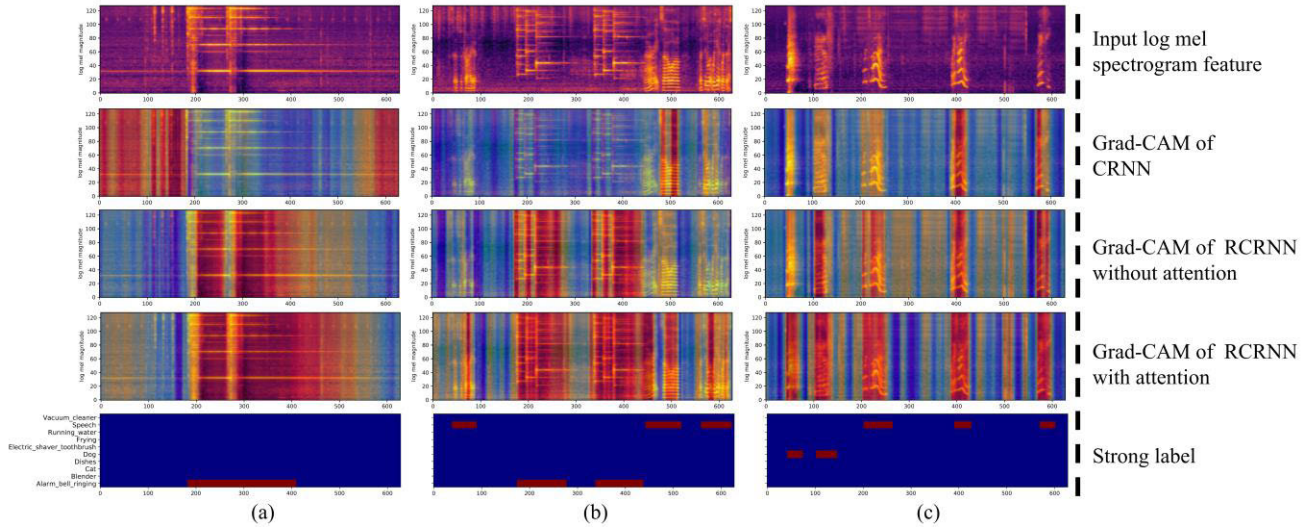

FIGURE 6. Grad-CAM illustrations of feature maps constructed by CRNN, RCRNN without CBAM-based attention, and RCRNN with CBAM-based attention for three different audio clips from Dev set: (a) Y5UMWvLV5DGU_40.000_50.000.wav, (b) Y0oudYrPGNN8_30.000_40.000.wav, and (c) Y5lw-BHt_rXY_30.000_40.000.wav.

V. CONCLUSION

This paper proposed a two-stage polyphonic SED model using a mixture of strongly labeled, weakly labeled, and unlabeled data. Compared to the baseline of DCASE 2019 and DCASE 2020 Challenge Task 4, the main contributions were as follows:

1) The RCRNN-based mean teacher model was formed by combining a residual convolutional network and a recurrent neural network to improve the feature map of sound events compared to CRNN-based feature map in the challenge baselines. In addition, CBAM-based attention was applied to the convolutional layers in the proposed RCRNN-based mean teacher model to further improve the feature map.
2) By using the RCRNN-based mean teacher model as a predefined model for labeling weakly labeled and unlabeled data, a two-stage SED model was formed. The mean teacher model was directly used as the SED model in the baselines.
3) A semi-supervised loss function was used to train the second stage model of the proposed SED model to accommodate the noisy target labels from the first stage of the proposed two-stage SED model.

In other words, the first stage of the proposed SED model consisted of an RCRNN-based mean teacher model with CBAM-based attention, and the second stage was a fine-tuning model from the student model trained in the first stage, where a semi-supervised loss function was used to effectively train the weakly labeled and unlabeled data.

The effectiveness of the proposed two-stage SED model was evaluated by applying it to both DCASE 2019 and DCASE 2020 Challenge Task 4, and its performance was compared with those of the baseline and top-ranked models from both challenges by measuring the F1-scores, ERs, and PSDSs. First, the contribution of the proposed RCRNN-based mean teacher model to SED was confirmed by comparing the performance of the CRNN-based mean teacher model of the baseline for each of the challenges. Consequently, it was shown that the F1-scores of the proposed RCRNN-based mean teacher model were improved by 23.1% and 12.0% for the Dev set of DCASE 2019 Challenge Task 4 and DCASE 2020 Challenge Task 4, respectively, compared to those of the CRNN-based baseline. Second, the performance of the proposed two-stage SED model was compared to that of the baseline that corresponded to the first stage of the proposed two-stage model. When the semi-supervised loss function was not employed, i.e., $\beta = 1.0$ in (12), the F1-scores of the proposed two-stage SED model were improved by 2.1% and 3.1% for the Dev set of DCASE 2019 and 2020 Challenge Task 4, respectively, compared to those of the first-stage model only. Third, after applying the semi-supervised loss function, the F1-scores of the proposed two-stage SED model were further increased by 1.7% and 0.7% for the Dev set of DCASE 2019 and 2020 Challenge Task 4, respectively, compared to those of the proposed two-stage SED model without the semi-supervised loss function.

Furthermore, the proposed two-stage model with the semi-supervised loss function was the best in terms of the F1-score and PSDS. Interestingly, the semi-supervised loss function worked well with any value of the control parameter.

REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Cham, Switzerland: Springer, 2018, ch. 1, pp. 3–12.
[2] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 374–378.
[3] J. P. Bello, C. Silva, O. Nov, R. L. DuBois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: A system for the monitoring, analysis and mitigation of urban noise pollution," *Commun. ACM*, vol. 62, no. 2, pp. 68–77, Feb. 2019.
[4] Y. Zigel, D. Litvak, and I. Gannot, "A method for automatic fall detection of elderly people using floor vibrations and sound—Proof of concept on human mimicking doll falls," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 12, pp. 2858–2867, Dec. 2009.
[5] A. Temko and C. Nadeu, "Acoustic event detection in meeting-room environments," *Pattern Recognit. Lett.*, vol. 30, no. 14, pp. 1281–1288, Oct. 2009.
[6] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *Proc. 18th Eur. Signal Process. Conf. (EUSIPCO)*, Aalborg, Denmark, Aug. 2010, pp. 1267–1271.
[7] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP J. Audio, Speech, Music Process.*, vol. 2013, no. 1, pp. 1–13, Jan. 2013.
[8] E. Cakir, T. Heittola, H. Huttunen, and T. Virtanen, "Polyphonic sound event detection using multi label deep neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Killarney, Ireland, Jul. 2015, pp. 1–7.
[9] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brisbane, QLD, Australia, Apr. 2015, pp. 559–563.
[10] H. Phan, L. Hertel, M. Maass, and A. Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 3653–3657.
[11] G. Parascandolo, H. Huttunen, and T. Virtanen, "Recurrent neural networks for polyphonic sound event detection in real life recordings," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 6440–6444.
[12] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.
[13] S. Adavanne, P. Pertila, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 771–775.
[14] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 121–125.
[15] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 11, pp. 2180–2193, Nov. 2018.
[16] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Brighton, U.K., May 2019, pp. 31–35.
[17] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.
[18] A. Kumar and B. Raj, "Audio event and scene recognition: A unified approach using strongly and weakly labeled data," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Anchorage, AL, USA, May 2017, pp. 3475–3482.
[19] A. Shah, A. Kumar, A. G. Hauptmann, and B. Raj, "A closer look at weak label learning for audio events," 2018, *arXiv:1804.09288*. [Online]. Available: http://arxiv.org/abs/1804.09288

[20] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 1195–1204.
[21] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. Detection Classification Acoust. Scenes Events Workshop (DCASE)*, New York, NY, USA, 2019, pp. 253–257.
[22] J. Goldberger and E. Ben-Reuven, "Training deep neural networks using a noise adaptation layer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017, pp. 1–9.
[23] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, no. 10, pp. 1733–1746, Oct. 2015.
[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.
[25] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Jun. 2018, pp. 3–19.
[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531
[27] C. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulovic, "A framework for the robust evaluation of sound event detection," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 61–65.
[28] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," in *Proc. Workshop Detection Classification Acoust. Scenes Events (DCASE)*, Tokyo, Japan, Nov. 2020, pp. 200–204.
[29] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning convolution system for DCASE 2019 task 4," in *Proc. Detection Classification Acoust. Scenes Events Workshop (DCASE)*, New York, NY, USA, Oct. 2019, pp. 134–138.
[30] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Convolution augmented transformer for semi-supervised sound event detection," in *Proc. Workshop Detection Classification Acoust. Scenes Events (DCASE)*, Tokyo, Japan, Nov. 2020, pp. 100–104.
[31] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 776–780.
[32] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, USA, Oct. 2017, pp. 344–348.
[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res.*, vol. 9, pp. 249–256, May 2010.
[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[36] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," 2016, *arXiv:1610.02242*. [Online]. Available: http://arxiv.org/abs/1610.02242
[37] M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar, "Adaptive methods for nonconvex optimization," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Montréal, QC, Canada, Dec. 2018, pp. 9793–9803.
[38] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for poly-phonic sound event detection," *Appl. Sci.*, vol. 6, no. 6, pp. 162–178, May 2016.
[39] N. K. Kim and H. K. Kim, "Polyphonic sound event detection based on convolutional recurrent neural networks with semi-supervised loss function for DCASE Challenge 2020 task 4," in *Proc. DCASE Challenge*, Jun. 2020, pp. 1–4.
[40] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626.

NAM KYUN KIM (Student Member, IEEE) received the B.S. degree in electrical engineering from Cheonnam National University, South Korea, in 2008, and the M.S. degree in information and communications engineering from the Gwangju Institute of Science and Technology (GIST), South Korea, in 2015, where he is currently pursuing the Ph.D. degree. His current research interests include sound event detection, speech recognition, audio signal processing, and machine learning.

HONG KOOK KIM (Senior Member, IEEE) received the B.S. degree in control and instrumentation engineering from Seoul National University, South Korea, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 1990 and 1994, respectively. From 1990 to 1998, he was a Senior Researcher with the Samsung Advanced Institute of Technology (SAIT), South Korea. From 1998 to 2003, he was a Senior Technical Staff Member with the Voice-Enabled Services Research Laboratory, AT&T Labs-Research, Florham Park, NJ, USA. Since August 2003, he has been a Professor with the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea. He is also jointly affiliated with the AI Graduate School, GIST. From 2014 to 2015, he was a Visiting Professor with the City University of New York, New York, USA. His current research interests include statistical and deep learning approaches on large vocabulary speech recognition, sound event detection, unsupervised anomaly detection, and speech/audio enhancement and source separation. He serves a member of the APSIPA Speech, Language, and Audio Technical Committee. He has served as an Editorial Committee Member. He has served as an Area Editor of *Digital Signal Processing*.