IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 November 2022, accepted 20 December 2022, date of publication 28 December 2022,
date of current version 5 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3232807

RESEARCH ARTICLE

Environmental Sound Classification With Low-Complexity Convolutional Neural Network Empowered by Sparse Salient Region Pooling

HAMED RIAZATI SERESHT AND KARIM MOHAMMADI
School of Electrical Engineering, Iran University of Science and Technology, Tehran 16846-13114, Iran
Corresponding author: Hamed Riazati Seresht (hamed_riazati@elec.iust.ac.ir)

ABSTRACT Environmental Sound Classification (ESC) is an important field in a broad range of applications, such as smart cities, audio surveillance, and health care. Recently, Convolutional Neural Networks (CNNs) have taken the lead from traditional approaches and have produced promising results. However, the achieved improvements are often accompanied by increasing depth, complexity, and size of the network, which prevents their usage in many practical applications. In this work, our goal is to empower a small-size low-complexity CNN model to achieve superior performance. To this end, we concentrate on the importance of global pooling technique, which is less investigated in ESC. In most previous works, models utilize global average pooling layer which does not consider regional saliency, and thus weakens the salient time-frequency regions contributions to the classification, and also to the training of convolutional kernels. We propose a novel global pooling method, called Sparse Salient Region Pooling (SSRP), which computes the channel descriptors using a sparse subset of features, and guides the model to effectively learn from the more salient time-frequency regions. Experimental results demonstrate that the proposed model with only 700K parameters yields accuracies of 86.7% on ESC-50 and 94.8% on ESC-10, which are comparable to that of the state-of-the-art methods. Compared to the baseline model, our model achieves absolute improvement of 21.8% in accuracy on ESC-50, with 98% smaller model size. Our visual analyses show that SSRP intensifies the responses of low-energy regions such that they contribute even more than high-energy regions to the classification of specific sound classes.

INDEX TERMS Convolutional neural networks, environmental sound classification, global feature pooling, low complexity, regional saliency.

I. INTRODUCTION

In recent years, development of Environment Sound Classification (ESC) methods has been one of the hottest topics in audio classification domain due to its potential use in various application areas such as smart cities [1], audio surveillance systems [2], health care [3], security control systems [4], and Internet of Things (IoT) [5]. For example, ESC can be used to automatically identify different environmental sound events, such as gun shots [6], siren [7], and bird sounds [8].

The quite limited knowledge of temporal and frequency characteristics, the lower signal to noise ratio, and less static patterns make ESC more challenging than other audio-related tasks such as music genre classification and speech recognition.

With the remarkable achievements of Deep Learning (DL) methods in different applications, such as image classification [9], [10], speech recognition [11], [12], music processing tasks [13], [14], different attempts have been made to employ DL framework for ESC. Various methods have been proposed based on using one-dimensional Convolutional Neural Network (CNN) to directly process raw audios [15], [16]. Despite the advantage of eliminating the need for manual feature

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Shariq Imran.

extraction and tuning several hyper-parameters, they cannot benefit from the important frequency clues and they suffer from high computational complexity due to use of numerous convolutional layers (i.e. up to 34 layers in [15]). Latest methods follow the DL framework proposed by Piczak et al. [17] where a Time-Frequency (T-F) representation is utilized as input image to a two-dimensional (2-D) CNN. To benefit from complementary features, a multi-stream network was proposed in [18] where each stream receives a different input feature. However, the multi-stream CNN is not only too complex, but also requires a large amount of memory. Several methods focused on increasing the depth of the network to extract more abstract and higher-level features [19], [20] and employed well-known image recognition architectures such as ResNet and DenseNet [20], [21]. However, these methods require large amount of training data, and limited available samples for several sound classes has hindered their further success. Also, No specific domain knowledge is incorporated in their design which is necessary to achieve superior performance. In [22], in order to distinguish between different frequency bands, a model consisting of an ensemble of several CNNs was proposed, which processes each frequency band separately. Recently, several works have attempted to combine CNN with recurrent neural networks which has improved the CNN performance at the cost of higher model parameters and complexity.

As mentioned above, most of these methods are mainly accompanied by a large increase in model size and complexity. The heavier computational load and the need for large memory greatly limit their usage for many low-power and low-resource applications such as hearing aids or IoT devices. Rather than increasing network depth or architectural complexity, we, instead, pose this question: how can a simple, small-size and not-so-deep CNN with a single input audio representation achieve superior performance for ESC task under limited data condition? To find a solution to this question, we believe that the following characteristics of environmental sounds are necessary to consider:

- *Temporal characteristics*: Environmental sound signals exhibit complex temporal structures with different levels of local relationship and varied durations. As can be seen in Figure 1, environmental sounds can be transient (e.g. crying baby), continuous (e.g. rain), or intermittent (e.g. clock tick). Moreover, local T-F patterns are highly shift-invariance across time axis so that temporal translation has little effect on the classification of sound events.
- *Spectral characteristics:* Compared to other audio signals, environmental sounds have a broader range of frequency information with diverse spectral profiles which are either scattered across frequency bands, concentrated at low, middle or higher frequency bands, or spread across all frequency bands [22], [23]. Also, unlike the time dimension, translation across the frequency dimension can significantly affect the performance of the sound classification [24].

As a result of these characteristics, T-F representations of environmental sounds exhibit diverse energy modulation patterns with considerable intra-class variations. Moreover, the input data may contain many noisy or silent frames with only a few semantically relevant frames (e.g. *clock tick*). The variation imposed by noises and sound sources unrelated to the target sound event adds to the variety of T-F patterns.

Revisiting a regular CNN shows that convolutional layers do not discriminate between different local T-F regions from the whole input representation and treat them equally. Also, the Global Average Pooling (GAP) layer, which is applied to CNN output channels in popular architectures such as ResNet [20], mixes all local features extracted from the whole input representation into a one-value descriptor via similar weighting. Thus, GAP ignores the details and leads to channel descriptors that have similar values. Given the mentioned spectral and temporal characteristics and the degree of variety of T-F patterns, this equal treatment of different regions reduces the model ability to learn a discriminative representation, especially when small training data is available. As a solution, attention mechanisms have been proposed to increase emphasis on temporal frames or channels that are more relevant to the target sound [19], [25]. But these methods rarely exploit the joint T-F information [23], [24], and bring about limited improvements at the cost of adding lots of learnable parameters.

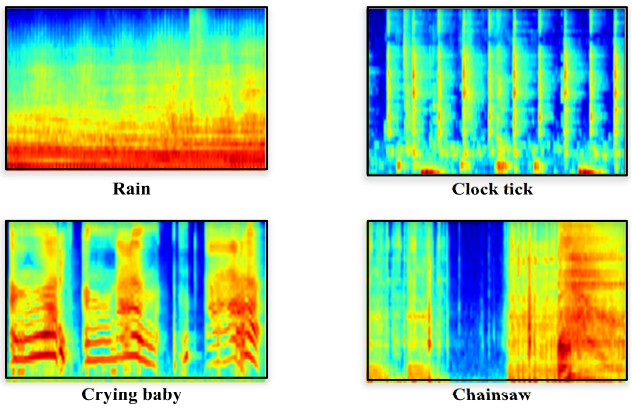

FIGURE 1. Examples of log mel spectrogram of different environmental sounds in ESC-10 dataset.

In this paper, we propose a novel and efficient approach which empowers a small-size and not-so-deep CNN to learn discriminative representation. To this end, we introduce a new global pooling scheme which is called Sparse Salient Region Pooling (SSRP) and does not include any extra learnable parameters. Figure 2 depicts the overall structure of the low-complexity CNN model equipped with SSRP. Among the mentioned diverse input patterns, we believe that there are few patterns that contain the most salient and discriminative information of each sound class, and other less informative patterns could be beneficial only if large amount of training data is available. Hence, the main idea behind SSRP is restricting the features that contribute to the channel

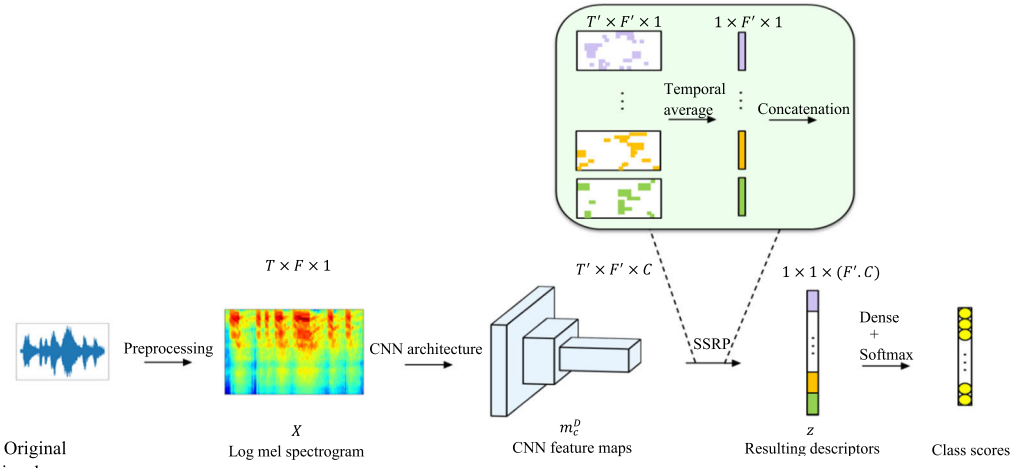

FIGURE 2. The overall structure of the proposed model.

descriptors to a subset of very sparse features which play a key role in the classification of a sound class, and ignoring all other features. This is similar to the mechanism that the human auditory system has developed to tune in a single voice in a crowded room full of interfering signals or noisy surroundings (also known as the cocktail party effect). Strikingly, studies conducted in [26] found that the auditory cortical representation only reflects the neural responses of the target speaker and all information coming from superfluous speakers was completely discarded. During training, SSRP largely affects the back-propagation process which forces the model to increase its focus on those frequent patterns that can contribute the most to the classification (i.e. class-dependent salient patterns). To fully utilize the information within different salient patterns, SSRP observes the aforementioned spectral and temporal characteristics by discriminating between local spectro-temporal patterns at different frequency bands and capturing the long-term temporal structures of different sound classes. To the best of our knowledge, in contrast to this work, none of the existing methods for ESC has investigated the impact of the global pooling method. Results of experiments conducted on ESC-50 and ESC-10 datasets show that our not-so-deep CNN model equipped with SSRP can yield comparable performance to the state-of-the-art methods under much lower computational complexity, and much smaller parameter space.

In summary, the main contributions of this paper are as follows:

- This paper proposes a novel global pooling method, named SSRP, which combines local and global information in a way that guides CNN kernels to efficiently learn from more informative and salient T-F regions of input representation.
- We show that the most salient information can reside in a small percentage of total frames, and thus extracting features only from these informative frames leads to considerable improvements.
- Our proposed model achieves 21.8% and 14.3% absolute improvements over the baselines of ESC-50 and ESC-10 datasets respectively, a notable performance for a lightweight network with total parameters of 0.7 M.

The rest of this paper is organized as follows. We briefly review the existing methods proposed for ESC and provide a description about different pooling methods developed for visual domain in Section II. Section III presents details of the proposed global pooling method and its impact on the back-propagation process. Section IV provides the experimental settings and results on ESC-50 and ESC-10 datasets. Finally, Section V concludes the paper.

II. RELATED WORK

A. SOUND CLASSIFICATION NETWORKS

During the past few years, many attempts have been made to employ CNN models for ESC. As the first work, Piczak [17] fed a CNN with log mel spectrogram, as a low-level 2-D representation, and achieved significant progress over traditional support vector machine, random forest and K-nearest neighbor classifiers. In [21], using the same 2-D input features, two well-known CNN architectures (AlexNet and GoogLeNet) in image classification were utilized for ESC. In [15], very deep 1-D CNNs were proposed to learn directly from raw audio waveforms. In [16], Gammatone filterbanks were utilized to initialize the first layer of an end-to-end 1-D CNN which could deal with inputs of different lengths. Another end-to-end system was proposed in [27] where two 1-D convolutional layers first work on the raw audio and learn to extract a 2-D representation which is then processed by 2-D CNNs. In [28], a multi-scale time-domain convolutional layer was used as the first layer to provide an improved frequency resolution and build more discriminating 2-D representation. The 2-D representation is then fused with log mel spectrogram using a two-phase method. In [19], a concatenation of multiple features consisting of MFCC, Gammatone Frequency Cepstral Coefficients

(GFCC), the Constant Q-transform (CQT), and Chromagram were used as input of a deep 2-D CNN which discriminates between time and frequency domains via spatially separable convolution kernels with different sizes. In [25], a Convolutional Recurrent Neural Network (CRNN) is used where 2-D convolutional layers first process log mel spectrogram and then recurrent layers model temporal dynamics. Several works [18], [29], [30] have attempted to utilize temporal attention mechanism to focus on the more informative frames. In [30], an architecture was designed which takes advantage of CNNs to model spectral information, LSTMs to model long-term temporal information, and deep neural networks to classify the obtained features. Also, an attention mechanism is used to determine the importance of different time outputs of the Long Short-Term Memory (LSTM) layer. In [18], a multi-stream CNN equipped with temporal attention was proposed which relies on three different input streams consisting of raw audio and spectral features. Increasing the depth and complexity of the deep learning models intensifies the need for huge amounts of training data. To deal with the scarcity of labeled training data for ESC, data augmentation techniques have been employed in the literature. For example in [17], [21], and [31], deformations such as time stretching, pitch shifting and dynamic range compression were applied to the available data to generate new training samples. Several works [19], [25], [32] employed mixup technique; another widely-used data augmentation which simply mixes up different features and their labels with one another [33]. While these approaches improve the CNN-based classifier performance, they typically suffer from a large parameter space and high-complexity.

B. FEATURE POOLING METHODS

To the best our knowledge, global pooling layers are not investigated in the architectures developed for ESC, and simply average pooling is used in most networks [20], [34], [35]. To gain insights into other pooling strategies, here, we review some of the pooling methods proposed in the visual domain. Stochastic pooling [36] randomly picks the activation within each pooling region according to a multinomial distribution. S3Pool [37] takes the maximum of activations of a sub-region which is randomly selected within the pooling region. Detail-preserving pooling [38] computes weighted average pooling where the weights are proportional to the differences of activations. Rank-based pooling methods [39], [40], [41] sort the activations in descending order and compute the average of top-K values. In Rank-based weighted pooling [41], prior to taking the sum of activations, each activation is weighted by a coefficient computed based on its rank. To reflect advantages of max and average pooling methods, mixed pooling method chooses between the two pooling values where the selection is done based on a random coefficient [42]. In [43], the stochastic coefficient is replaced with a learnable coefficient in order to benefit from the two pooling methods at the same time. These pooling methods have been mostly utilized to be used between convolutional layers, and GAP is still used as the global feature pooling at the end of most popular architectures. Some researchers have attempted to improve the averaging mechanism of GAP. To increase focus on more semantically important regions, entropy pooling [44] assigns entropy-based weights to different features prior to computing the average value. In [45] AlphaMEX was proposed which uses log mean exponential function to aggregate the feature maps. Stochastic region pooling [46] randomly selects some regions within each feature map and takes the average value of the selected features. This is done to encourage the responses of detail features to be enhanced during training, and make the network learn more representative channel descriptors. Global learnable pooling [47], as another weighted average pooling method, assigns lots of learnable weights to all features within output feature maps, where the weights learn to highlight the contribution of more distinctive features.

III. PROPOSED GLOBAL POOLING METHOD

A. SPARSE SALIENT REGION POOLING (SSRP)

In this section, we present the proposed method which aims at increasing the focus of a lightweight CNN on the most salient input T-F patterns. These spectro-temporal patterns are formed by energy modulations of frequency content across the temporal dimension with horizontal or diagonal directions [48]. Identifying these class-dependent salient patterns in the input representation has many difficulties and requires enough acoustic knowledge of different sound classes to develop low-level saliency clues. Developing a separate sub-network which learns to identify these patterns, and then utilizing its decisions for the main network, similar to [49], can also be considered as an option, but it would greatly increase the model complexity and the total number of model parameters, and intensify the need for training data. We, instead, introduce a new global pooling scheme which affects the training of the model so that it is gradually biased towards learning the most salient input patterns.

The most commonly used global pooling method, GAP, treats all features of each CNN output channel equally, and mixes them to a single-value channel descriptor via calculating the mean value $z_c = 1/(T' \times F') \times \sum_{T'} \sum_{F'} m_c^D(t,f)$, where m_c^D is the c^{th} channel of the D^{th} convolutional layer (last layer) containing T' temporal frames of F' frequency bands. The input Receptive Fields (RFs) of many of these features can correspond to different non-overlapping T-F regions of input representation, depending on the total pooling size in the previous layers. Thus, the saliency of information embedded in different components of m_c^D may vary a lot because of the varied spectral and temporal characteristics of environmental sounds, as mentioned before. However, GAP does not consider regional saliency and lets all these features make the same contribution to the classification and then to the training of the corresponding convolutional kernels. Thus, those regions of low relevancy and importance may lessen the effect of regions containing discriminative patterns.

In the proposed SSRP, we restrict the computation of z_c to a subset Ω_c that contains indices of very sparse features of m_c^D and drop all other features outside Ω_c. This is equivalent to applying a sparse binary mask with a few non-zero components to each output channel z_c prior to computing the channel descriptor. By imposing such regional bottleneck on the flow of information in the forward path, only a small percentage of input T-F regions can contribute to the classification and then to the training of the kernels of the last convolutional layer.[1] At the early steps of training, the model may focus on capturing information within the less salient or even irrelevant patterns and thus the subset would contain no discriminating features, which would not lead to large reduction of the total loss of the network. This keeps the gradient values large which would force the model to search for the most informative input T-F patterns. Note that the desired T-F regions contain the most informative and frequent patterns that occur in most of the samples of each target sound class. Thus, we expect further training steps to encourage the responses of salient patterns in the output feature maps (z_c) to gradually intensify such that the selected sparse subset Ω_c be enriched with more discriminative features during training. To achieve this, we consider the selection routine of members of Ω_c as a decisive part of the proposed scheme. For instance, if sparse subset Ω_c is formed based on a stochastic selection routine (as in [37] and [46]), the variation of features could be very large between successive training steps and also no discriminative features may be involved in many training steps; attributes that will cause the idea to fail (as will be shown in Figure 7). Hence, we observe temporal and spectral characteristics in the design of SSRP. First, we note that each particular frequency band of m_c^D is of different importance among various classes. Thus, SSRP pools each frequency band separately to prevent discriminative features with lower activation levels from weakening by more dominant features at other frequency bands. Second, /the most informative and salient patterns have strong local correlations which should be preserved under the imposed restriction. Discarding parts of these features or including features of non-salient patterns would result in undesirable increased variations in the channel descriptors. Due to the mentioned diverse temporal structures, these patterns may last within a single local temporal interval of varied length, or may reside within multiple distant intervals, depending on the sound class. Hence, SSRP should also capture temporal dependencies among different temporal frames. Taking these aspects into consideration, we design a two-stage aggregation scheme where the first stage captures local information within different temporal intervals, and the second stage aggregates the first-stage results. Here, we propose three simple methods to implement SSRP. In the basic form, which we refer to as SSRP-B, we first slide a rectangular window of length W with unit stride over temporal features of each frequency band of m_c^D and compute the mean of the activations within each window

$$s_c(t,f) = \frac{1}{W}\sum_{i=1}^{W} m_c^D(t+i-W,f) \tag{1}$$

Then, in the second stage, $s_c(t,f)$ is aggregated into $z_c(f)$ by picking the highest mean of activations

$$z_c(f) = \max_t (s_c(t,f)) \quad t = 1,\ldots,T' \tag{2}$$

This way, members of Ω_c corresponds to the local features within the temporal windows selected at different frequency bands. Using the long-term max pooling across the time dimension in (2) is in accordance with the fact that patterns are invariant to high amounts of translation in the time domain, and makes SSRP non-sensitive to the position of salient patterns. Also, the temporal interval of the selected windows can be different among frequency bands, as illustrated in Figure 2. Thus, Ω_c is not limited to similar temporal intervals for all frequency bands and can contain discriminative features of different frequency bands irrespective of their temporal positions.

Note that increasing the size of W in (1) would decrease the level of sparsity (i.e. $(T' - W)/T'$), and thus increase the diversity of features that contribute to the descriptor. For the special case of $W = T'$, no restriction is imposed and SSRP-B is degenerated to a frequency-dependent temporal average pooling. Also, the sparsity level would be maximum when $W = 1$, for which SSRP-B would select only the most dominant feature at each frequency band and thus may lose some discriminative information. Thus, the size of W should be selected so that the descriptor can represent the feature map well enough for different sound classes.

Figure 3 shows validation accuracy curves of the proposed model during training using two different sparsity levels. Comparing the accuracies achieved for sparsity levels of 92% and 81% shows that at early epochs of training, the model with smaller sparsity level learns more quickly and achieves accuracies up to 20% higher than that of the model with larger sparsity level. This indicates that the more severe the regional bottleneck (i.e. the larger the sparsity), the harder it is for the training process to bias convolutional kernels towards detecting and capturing information of salient patterns. However, the performance of the model with higher sparsity level is improved by further training so that it finally achieves accuracy of 83.50%, while smaller sparsity level leads to accuracy of 81.75%.

The basic form of SSRP has one-degree of freedom to aggregate temporal local features and is able to capture one single interval of fixed size. Considering the complex temporal structures of environmental sound classes, this would cause some discriminative information to be lost or some features extracted from non-salient patterns to be a member of the sparse subset. Also, it would lose the global temporal dependencies between distant salient patterns. To address these issues, we develop two other forms of SSRP. In the

[1] As will be discussed in section III.B, SSRP also disrupts the flow of gradients in the backward path and thus affects the training of convolutional kernels of the previous layers.

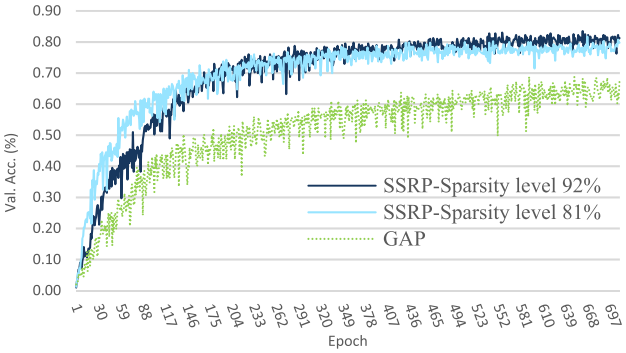

FIGURE 3. Validation accuracy curves of the proposed model (described in section IV.A) obtained during training using SSRP-B with W sizes of 8 and 20 frames ($T' = 107$), which are equivalent to the sparsity level of 92% (shown in dark blue) and 81% (shown in light blue), respectively. The model with higher sparsity level manages to surpass the model with smaller sparsity level after 182 epochs of training. The results obtained using GAP are also shown (in green round dotted), which demonstrates that training of model equipped with SSRP settles more quickly than the model equipped with GAP, indicating increased speed of training gained by SSRP.

second form, we compute the moving mean of activations using L levels of increasingly coarser windows

$$s_c^l(t,f) = \frac{1}{W_l}\sum_{i=1}^{W_l} m_c^D(t+i-W_l,f) \tag{3}$$

$$W_l = l \times W_0, \quad l = 1,..,L \tag{4}$$

where W_l is the length of the window at level l, and W_0 is the size of the finest window. Then, the temporal channel descriptor for each frequency band $z_c(f)$ is computed by taking average of L highest means of activation (one max at each level)

$$z_c(f) = \frac{1}{L}\sum_{l=1}^{L} \max_t \left(s_c^l(t,f)\right), \quad t = 1,\ldots,T' \tag{5}$$

We name this form SSRP-Multi-Scale (MS), as it combines multiple scales of local features; the finest level captures more local features while the coarsest level can preserve patterns of longer temporal correlation.

Moreover, as illustrated in Figure 4, if the selected windows are overlapping, then the total weights used to combine the selected features can be different from the uniform weighting of SSRP-B, depending on the amount of overlaps.

The third form we propose, referred to as SSRP-Top (T), also uses multiple intervals of local features but using fixed-size windows. In the first stage of SSRP-T, the moving mean of activations $s_c(t,f)$ is computed via (1), the same as SSRP-B. After sorting these locally aggregated features across the time dimension $s_c(f) : Q \rightarrow \{(q_1,\ldots q_n) : q_1 > \cdots > q_n$, the temporal descriptor $z_c(f)$ is calculated by taking average of top K values of Q

$$z_c(f) = \frac{1}{K}\sum_{i\in[1,K)} q_i \tag{6}$$

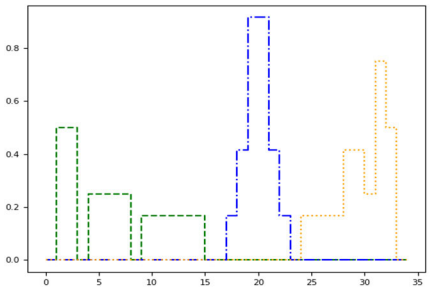

FIGURE 4. Illustration of non-uniform weighting performed by SSRP-MS with $L = 3$ and $W_0 = 2$, when the windows are non-overlapping (green dash line), overlapping and concentric (blue dash-dot line), and partially overlapping (orange dot-dot line).

Therefore, SSRP-T selects K different intervals that have the largest mean of activations at each frequency band. This way, the K intervals can be over-lapping and close to each other to capture temporal structures of different lengths, or far from each other to capture intermittent temporal structures. Thus, SSRP-T can also assign non-uniform weights to features selected by Ω_c, when the windows are over-lapping.

After computing the frequency-wise descriptors $z_c(f) \in R^{F'\times C}$, the final representation vector is formed by concatenation of $z_c(f)$ along both frequency and channel dimensions, as illustrated in Figure 2. Thus, SSRP not only pools each frequency band separately, but also it lets the classifier learn to weight them appropriately. This frequency-awareness enables the classifier to distinguish between features extracted from similar patterns but at different frequency bands. In order to investigate the impact of the mentioned frequency-awareness, we derive another variant of the basic form, referred to as SSRP-No Frequency Awareness (SSRP-NFA), which mixes the information of different frequency bands by taking average of m_c^D along the frequency dimension before applying (1)-(2). Hence, SSRP-NFA leads to the final representation vector of size $1 \times C$ rather than $F' \times C$. In addition, we consider another variant of SSRP, named as SSRP-Random (SRRP-R), where the temporal windows forming the sparse subset Ω_c are selected randomly.

B. THE IMPACT OF SSRP ON THE BACK-PROPAGATION PROCESS

The proposed global pooling scheme has a great impact on the training of the kernels of the last convolutional layer, and of the previous layers although to a lesser extent. To illustrate this, let z_c, E and $\delta_c = \partial E / \partial z_c$ be the c^{th} channel descriptor calculated by global pooling layer, the final loss function, and the back-propagated errors from the dense layers, respectively. Based on convolution operation, the c^{th} output channel is computed by

$$m_c^D(t,f) = \sum_{a=0}^{K_t}\sum_{b=0}^{K_f}\sum_{v=0}^{C^{D-1}} \left(m_v^{D-1}(t+a,f+b)\times w_{a,b,v,c}^D\right) \tag{7}$$

where $w^{D}_{a,b,v,c} \in R^{K_t \times K_f \times C^{D-1} \times C^{D}}$ is the weight of the convolutional kernel of size $K_t \times K_f$ which are convolved with the v^{th} channel of feature maps of $(D-1)^{th}$ layer m_v^{D-1} to produce the c^{th} channel of the D^{th} convolutional layer. Hence, the gradient of loss w.r.t. $w^{D}_{a,b,v,c}$ when GAP is used as the global pooling layer would be

$$\begin{aligned}
&\frac{\partial E}{\partial w^{D}_{a,b,v,c}} \\
&= \frac{\partial E}{\partial z_c} \times \frac{\partial z_c}{w^{D}_{a,b,v,c}} \\
&\xrightarrow{\delta_c = \frac{\partial E}{\partial z_c}} = \delta_c \times \sum_{t=0}^{T'} \sum_{f=0}^{F'} \left(\frac{\partial z_c}{\partial m_c^D(t,f)} \times \frac{\partial m_c^D(t,f)}{\partial w^{D}_{a,b,v,c}} \right) \\
&= \delta_c \times \sum_{t=0}^{T'} \sum_{f=0}^{F'} \left(\frac{1}{T' \times F'} \times \frac{\partial m_c^D(t,f)}{\partial w^{D}_{a,b,v,c}} \right) \\
&= \delta_c \times \left(\frac{1}{T' \times F'} \right) \times \sum_{t=0}^{T'} \sum_{f=0}^{F'} m_v^{D-1}(t+a, f+b) \qquad (8)
\end{aligned}$$

As can be seen, the updating value for each weight of the last convolutional layer ($\partial E / \partial w^{D}_{a,b,v,c}$) depends on the average of all components of the corresponding feature map (v^{th}) of its input. The same computation can be performed for the proposed SSRP methods via z_c computed in (2), (5), and (6)

$$\begin{aligned}
&\frac{\partial E}{\partial w^{D}_{a,b,v,c}} \\
&= \sum_{f=0}^{F'} \left(\frac{\partial E}{\partial z_c(f)} \times \frac{\partial z_c(f)}{w^{D}_{a,b,v,c}} \right) \\
&\xrightarrow{\delta_c(f) = \frac{\partial E}{\partial z_c(f)}} = \sum_{f=0}^{F'} \left(\delta_c(f) \times \sum_{t=0}^{T'} \left(\frac{\partial z_c(f)}{\partial m_c^D(t,f)} \times \frac{\partial m_c^D(t,f)}{\partial w^{D}_{a,b,v,c}} \right) \right) \\
&\xrightarrow{SSRP-B} = \sum_{f=0}^{F'} \left(\delta_c(f) \times m_v^{D-1}(t_{max}+a, f+b) \right) \qquad (9) \\
&\xrightarrow{SSRP-MS/T} = \sum_{f=0}^{F'} \left(\delta_c(f) \times \frac{1}{|\Omega_c(f)|} \times \sum_{t \in \Omega_c(f)} m_v^{D-1}(t+a, f+b) \right) \qquad (10)
\end{aligned}$$

In above operations, zero gradients are assigned for non-maximum values when using the max operator in the forward path. Thus, $\partial z_c(f) / \partial m_c^D(t,f)$ has non-zero value(s) (unit value) only for temporal frame(s) within Ω_c. According to equations (9) and (10), it is seen that few T-F locations of the feature maps of the previous layer (m_v^{D-1}) contribute to the updating values of the final convolutional layer ($\partial E / \partial w^{D}_{a,b,v,c}$). These T-F locations are the corresponding locations of m_v^D that are members of Ω_c.

With similar computations and back-propagating the errors to the $(D-1)^{th}$ layer, $\partial E / \partial w^{D-1}_{a,b,v,c}$ for different global pooling methods would be

$$\begin{aligned}
&\frac{\partial E}{\partial w^{D-1}_{a,b,v,c}} | \boldsymbol{GAP} \\
&= \sum_{c=0}^{C} \sum_{t=0}^{T'} \sum_{f=0}^{F'} \left(\delta_c \times \left(\frac{1}{T' \times F'} \right) \times w^{D}_{a,b,v,c} \times \frac{\partial m_v^{D-1}(t,f)}{\partial w^{D-1}_{a,b,v,c}} \right) \qquad (11) \\
&\frac{\partial E}{\partial w^{D-1}_{a,b,v,c}} | \boldsymbol{SSRP-B/T} \\
&= \sum_{c=0}^{C} \sum_{f=0}^{F'} \left(\delta_c(f) \times w^{D}_{a,b,v,c} \times \frac{\partial m_v^{D-1}(t_{max}, f)}{\partial w^{D-1}_{a,b,v,c}} \right) \qquad (12) \\
&\frac{\partial E}{\partial w^{D-1}_{a,b,v,c}} | \boldsymbol{SSRP-MS/T} \\
&= \sum_{c=0}^{C} \sum_{f=0}^{F'} \left(\delta_c(f) \times w^{D}_{a,b,v,c} \times \frac{1}{|\Omega_c(f)|} \times \sum_{t \in \Omega_c(f)} \frac{\partial m_v^{D-1}(t,f)}{\partial w^{D-1}_{a,b,v,c}} \right) \qquad (13)
\end{aligned}$$

Comparing equations (11)-(13) shows that the large restriction imposed on the feature maps of the final convolutional layer has also affected the updating values of the layer before this layer. However, as illustrated in Figure 5, the T-F regions contributing to $\partial E / \partial w^{D-1}_{a,b,v,c}$ are larger than that of the final convolutional layer due to the summation over channels (c), which aggregates the locations of non-zero gradients from different channels of the D^{th} layer. Therefore, we infer that the closer to SSRP layer, the more restriction is imposed on the training of the convolutional kernels. This effect is reasonable as the bottom layers, which learn to extract more general features, are trained using larger T-F regions of input and learn from more diverse patterns. But the training of the top layers, which learn to extract more abstract features, are more selectively done using less varied patterns.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENT SETUP

We evaluate the proposed global pooling method on two publically available datasets: ESC-50 and ESC-10. ESC-50 consists of 2000 5-second recordings, which is of total duration of 168 minutes, from 5 major categories of animals (e.g. dog barking), natural soundscapes and water sounds (e.g. chirping birds and pouring water), human non-speech sounds (e.g. clapping), domestic sounds (e.g. keyboard typing) and exterior sounds (e.g. car horn), each containing 10 equally-balanced classes of sound events [50]. ESC-10 consists of 10 classes, with a total duration of 33 minutes, selected from ESC-50, with 40 samples for each class.

We use a simple and small-size CNN architecture with three convolutional layers with kernel sizes of 3 × 3. Each convolutional layer is followed by batch normalization [51] and Rectified Linear Unit (ReLU) activation. The number of kernels in each convolutional layer is twice the number in its

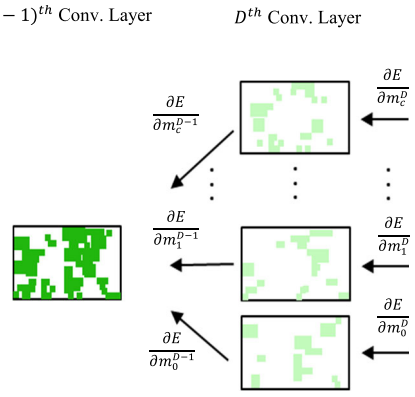

FIGURE 5. Illustration of T-F regions that contribute to compute gradient of E w.r.t. the weights of $(D-1)^{th}$ layer ($\frac{\partial E}{\partial w_{a,b,v,c}^{D-1}}$).

preceding layer, i.e. 32 → 64 → 128. Similar to [20], average pooling layers with kernel sizes of 2 × 2 are applied to the first two convolutional layers, as average pooling has been resulted in higher performances than max pooling [52]. The input of the network is a one-channel log-mel spectrogram of size 431 × 40×1 ($T \times F \times 1$), which is extracted by using 40 mel filters, with window and hop sizes of 1024 (46.4 ms) and 256 (11.6 ms), respectively. The convolutional section is followed by global pooling layer and one dense layer containing 128 neurons with ReLU activations. Dropout is applied after the global pooling and the dense layer with the rate of 0.5. Finally, the output probabilities are produced using a softmax layer.

The model is trained to optimize the categorical cross-entropy loss using stochastic gradient descent with batch size of 64, and learning rate and momentum of 0.1 and 0.9, respectively. We use the five-fold cross-validation setup provided in ESC-10 and ESC-50 datasets, and the mean accuracy of the five folds is reported as the final accuracy. To boost our model's performance, we employ mixup data augmentation [33], which extends the distribution of training data by mixing two random training samples and their one-hot labels via a random mix ratio drawn from beta distribution with $\alpha = 0.2$. We employ Keras library with Tensorflow as backend to implement the proposed model, and Librosa [53] for audio processing and feature extraction. All models are run on a system with a 16 GB RAM and a NVIDIA GTX 1060 GPU.

B. PERFORMANCE COMPARISON WITH STATE-OF-THE-ART

Table 1 compares the performance of our five-layer CNN model equipped with the proposed global pooling method to other models reported in previous works. Moreover, we observe the effect of doubling the number of convolutional kernels and neurons of the dense layer on the model performance. This model, which is referred to as CNN-D, has still much lower parameters than state-of-the-art models.

From Table 1, we see that the proposed SSRP layers show remarkable improvements over GAP. SSRP does not include any learnable parameters in its descriptor computations and the slight increase in the size of model equipped with SSRP compared to that of with GAP is due to computing multiple frequency-dependent descriptors from each channel which increases the number of weights of the dense layer. It is seen that both SSRP-MS and SSRP-T improve SSRP-B which indicates the effectiveness of use of multiple windows. Compared to the pioneer work of Piczak et al., as the baseline model, we see that CNN+SSRP-T with model size lower than 0.01 of the size of PiczakCNN, shows 19.8% and 14.3% higher accuracies than PiczakCNN on ESC-50 and ESC-10 datasets, respectively. Also, our model with only 0.2 M parameters yields 3.4% more accuracy than EnvNet-v2 (with 101 M parameters) on ESC-10.

Results show that the model performance is further enhanced on ESC-50 (up to 2.1% absolute improvements) by doubling kernels/neurons of the model. As can be seen, the best performance of the proposed CNN-D (with 0.7 M parameters) on ESC-50 is obtained by using SSRP-T which achieves 86.7%. In addition, CNN-D produces accuracy of 94.8% on ESC-10. When comparing with other models of Table 1, it should be noted that [19], [25], [54] benefit from temporal attention mechanisms, and the models proposed in [25] and [54] utilize Recurrent Neural Networks (RNNs) to model temporal structure of environmental sound signals. Also, [18] and [19] employ multiple input feature channels, and [55] fine-tunes an ensemble of well-known pre-trained image classification networks, such as Inception, ResNet50 and ResNet101. However, the proposed CNN-D model equipped with SSRP-MS/T achieves higher and comparable accuracies without use of multiple input feature channels, attention mechanisms, or recurrent networks, and too many learnable parameters.

Figure 6 shows the normalized confusion matrix generated by the proposed CNN-D+SSRP-T model for ESC-50 dataset. It is seen that most classes achieve classification accuracy higher than 80%. Particularly, all signals belonging to *pouring waters* are classified correctly (100% accuracy). Also, 10% of *cat* samples and 12% of *footsteps* samples are misclassified into *crying baby* and *door knock*, respectively. Moreover, only 38% of *helicopter* signals are classified correctly, and *airplane*, *washing machine* and *engine* classes have attracted 14%, 11% and 10% of *helicopter* samples, respectively. These misclassifications could be due to similar characteristics between these environmental sounds. But the overall classification results are highly satisfactory.

C. ANALYSIS OF COMPUTATIONAL COMPLEXITY

In order to further demonstrate the efficiency of the proposed method, we also evaluated the computational complexity of the proposed model for ESC-50 dataset. Besides the number of model parameters, studied in Table 1, computational complexity is another aspect of significant

TABLE 1. Comparison of accuracies obtained by different methods on ESC-50 and ESC-10 datasets. The best results of the previous works and our methods are presented in bold. For methods that use an ensemble of models, the maximum depth is considered.

Model	ESC-10	ESC-50	Depth	Feature	Number of Parameters (In Million (M))
PiczakCNN [17]	80.5%	64.9%	4	Log Mel-Delta	31.5 M
SoundNet [56]	92.1%	74.2%	8	Raw Data	3.2 M
EnvNet-v1 [27]	87.2%	70.8%	7	Raw Data	48.0 M
EnvNet-v2 [57]	91.4%	84.9%	13	Raw Data	101.2 M
DS-CNN [58]	92.6%	83.1%	9	Raw Data-Log Mel	2.3 M
Residual Network [35]	87.3%	-	19	Log Mel	11.7 M
Attention-based Residual Network [35]	92.0%	-	19	Log Mel	11.9 M
Multi-Stream CNN [18]	93.7%	83.5%	16	Raw Data-Log Mel	-
Attention-based CNN-GRU [54]	94.2%	86.5%	11	Log Mel-Delta	-
Attention-based CNN-GRU [25]	93.7%	86.1%	11	Log Mel-Delta	3.8 M
Attention-based DCNN [19]	94.7%	87.4%	16	MFCC-GFCC-CQT-Chromagram	1.3 M
T-FCNN [23]	-	84.4%	9	Log Mel	1.6 M
DCNN [59]	**94.9%**	**89.3%**	5	Log Mel	3.2 M
M-LM-C CNN [60]	-	85.6%	8	Log Mel	11.3 M
Ensemble of CNNs [55]	-	88.6%	101	Log Mel-Cochleagram-	250.0 M
Human [50]	95.7%	81.3%			
CNN + GAP	90.6%	70.3%	5	Log Mel	0.1 M
CNN + SSRP-B	94.2%	84.1%	5	Log Mel	0.2 M
CNN + SSRP-MS	94.6%	84.9%	5	Log Mel	0.2 M
CNN + SSRP-T	**94.8%**	84.7%	5	Log Mel	0.2 M
CNN-D + GAP	91.5%	75.4%	5	Log Mel	0.4 M
CNN-D + SSRP-B	94.6%	86.2%	5	Log Mel	0.7 M
CNN-D + SSRP-MS	**94.8%**	86.4%	5	Log Mel	0.7 M
CNN-D + SSRP-T	**94.8%**	**86.7%**	5	Log Mel	0.7 M

importance when implementing the model in low-resource and low-power applications. To evaluate the computational complexity, we computed the number of floating point operations (FLOPs)[2] and the inference time of the model when processing a sample audio recording file.[3] Table 2 shows values computed for the proposed model, baseline model of ESC-50 [17], and three other convolutional networks [25], [27], [35].[4] From Tables 1 and 2, we can see that EnvNet-V2 has improved the accuracy of the baseline model at the cost of considerable increase of FLOPs and inference time (i.e. 3.19 billion more FLOPs and 440 milliseconds more time). Also, the residual network and CRNN have greatly increased either FLOPs or inference time. However, it is seen that the inference time of the proposed model is very close to the baseline model and the increase in FLOPs (0.1 billion) is much less than that of the other models. Therefore, the comparison indicates that our 5-layer CNN equipped with the proposed SSRP layer efficiently utilizes the network capacity, and results in similar or higher accuracies than the compared models, but with much lower computational complexity.

D. IMPACT OF HYPER-PARAMETERS

Referring to Eq. (1-6), we conduct several experiments on ESC-50 to study the impact of three hyper-parameters: the windows size W, the maximum scale level L in SSRP-MS, and the number of windows K in SSRP-T.

First, we evaluate the impact of sparsity on the performance of the model equipped with SSRP-B via using different sizes of W. Figure 7 shows the obtained results. It is observed that when no sparsity is imposed (i.e. $W = 431$), SSRP-B achieves accuracy of 79.3%, making 11.6% absolute improvement over GAP. This large gain, achieved under no sparsity condition, indicates the considerable importance of the frequency-awareness derived by separate pooling of different frequency bands and prevention of mixing the salient information of them. We can see that decreasing the size of W from the maximum value (i.e. $T' = 431$), and thus imposing sparsity, significantly improves the accuracies obtained for both SSRP-NFA and SSRP-B. Strikingly, the best accuracies are obtained for very short temporal windows. For instance, SSRP-B reaches its highest accuracy (83.5%) using $W = 6$ which is equivalent to about 0.01 of the total frames or 140 ms of the input signal. This verifies that the most salient information may reside in very small interval of frames and that imposing sparsity could make model learn to extract the most salient

[2]A convolutional layer with N output channels of size $w \times h$, kernels of size $k_w \times k_h$, and M input channels, performs $N \times M \times w \times h \times k_w \times k_h$ multiplication operations, and almost the same number of addition operations.

[3]To estimate the inference time, we computed the average value of inference time needed to process all 400 samples within one validation fold of ESC-50.

[4]For the FLOPs presented in Table 2, we reproduced the network architectures according to their descriptions and used Ptflops library.

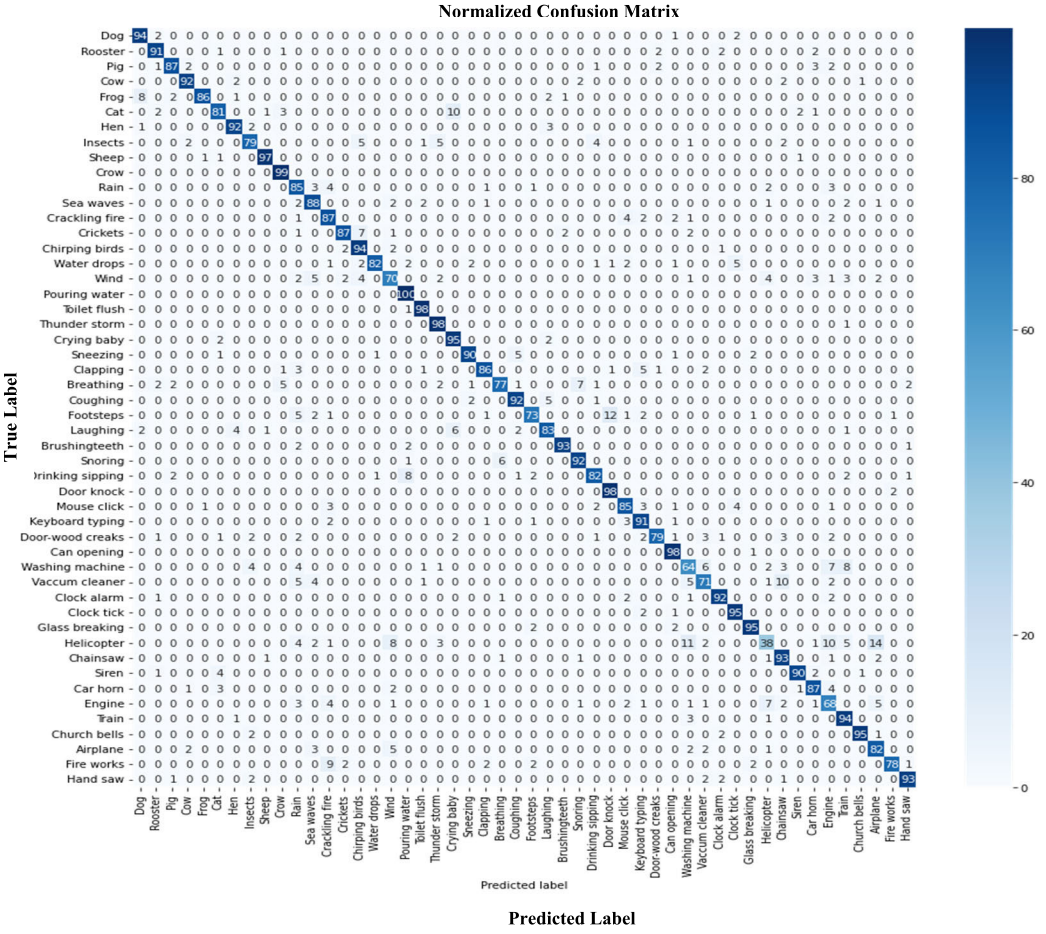

FIGURE 6. Normalized confusion matrix of CNN-SSRP-T ($K = 4$) with an average classification accuracy of 86.7% on ESC-50 dataset. The values are normalized to the total number of samples for each class, and shown in percentage (e.g. the accuracies gained for sheep, crow, and pouring water are 97%, 99% and 100%, respectively).

TABLE 2. Analysis of computational complexity of the proposed model, the baseline model, and three other networks composed of convolutional, fully-connected, and bi-directional gated recurrent layers denoted by Conv, FC, and BiGRU, respectively.

	Input Size	Architecture Details	FLOPs ($\times 10^9$)	Inference Time (ms)
Baseline [17]	41×60×2	2 Conv, 3 FC	0.069	9.1
EnvNet-v2 [27]	1×1×66650	10 Conv, 3 FC	3.260	449.2
Residual Network [35]	431×128×1	17 Conv, 2 FC	3.280	45.8
Convolutional RNN [25]	128×128×2	8 Conv, 2 BiGRU, 1 FC	0.915	108.2
Proposed CNN	431×40×1	3 Conv, 2 FC	0.166	10.7

information. Also, it is seen that a large range of window sizes ($6 \leq W \leq 35$) leads to accuracies as much as 1% lower than that of maximum accuracy of SSRP-B. This indicates that use of single window of fixed size and thus applying similar level of sparsity might not be the appropriate choice regarding the differences in the temporal structure of different sound classes. Moreover, accuracies obtained by SSRP-R confirm that random selection of the sparse subset Ω_c could notably degrade the model performance. These results were obtained by using no temporal pooling between convolutional layers. This minimizes the input temporal RF of the network down to 7 frames, as illustrated in Figure 8, which means that the CNN output features correspond to input T-F regions with smaller amounts of overlaps.[5] Hence, more temporal resolution is available to SSRP to distinguish salient T-F regions from other less relevant regions. However, lack of temporal pooling could decrease the CNN ability to learn T-F

[5]Using frame hops of 256 samples and sample rate of 22050 Hz, 7 frames is equivalent to $(7 \times 256)/22050 = 80ms$.

patterns and thus a trade-off must be made. We found that the model performance is maximized at temporal pool size of 2, and further increase in pooling size degrades the model performance.

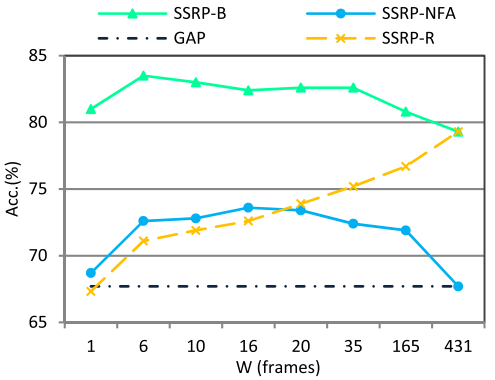

FIGURE 7. Accuracies obtained on ESC-50 datasets using SSRP-B, SSRP-NFA and SSRP-R for different window sizes when no temporal pooling is performed and $T' = T = 431$. The performance of model obtained using GAP is also shown as the reference.

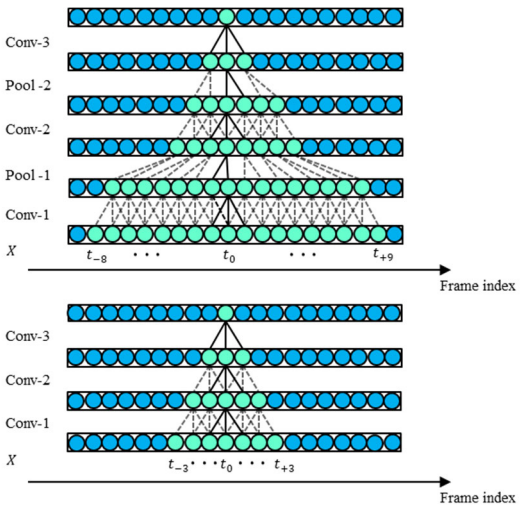

FIGURE 8. Comparison of CNN temporal RF size when no temporal pooling is performed (lower section), where each output feature corresponds to 7 frames of input representation, and when pooling with stride of 2 is performed (upper section), where each output feature corresponds to 18 frames of input representation.

Figure 9 shows the accuracies obtained for different sets of (W, L) and (W, K) used for SSRP-MS and SSRP-T, respectively. Clearly, the results indicate that use of multiple windows improves the model performance, especially when the window size is small which let the model capture both transient small T-F patterns, and those with longer temporal structures. It is observed that both SSRP-MS and SSRP-T outperform the basic form and the best accuracy (84.9%) is achieved by SSRP-MS using the pair of $(W = 2, L = 5)$, and the model performance declines for larger window sizes (e.g. 83.1% for $W = 10$ and $L = 4$).

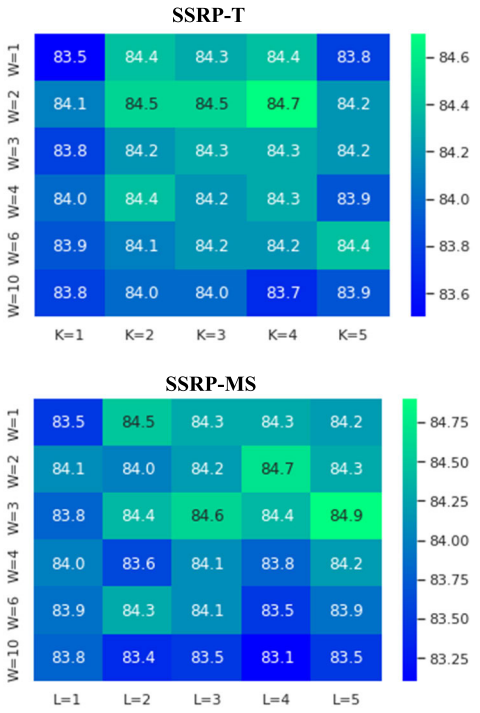

FIGURE 9. Accuracies obtained on ESC-50 using SSRP-T (upper section) and SSRP-MS (lower section) under different pairs of hyper parameters.

E. VISUAL ANALYSIS OF THE PROPOSED SSRP

To have a better understanding of the effect of the proposed global pooling method, we study the areas of input log mel spectrogram on which SSRP layer has focused for different sound classes. To this end, we form a mask for each output channel, where the nonzero values correspond to input RFs of the members of Ω. Also, we use GradCAM [61] to study the visualization results of network. Figure 10 shows the resulting masks and Grad-CAM heat-maps produced for three sound log mel-spectrogram samples and exemplary output channels. The produced masks (Figure 10 (b)) show that each channel focuses on specific time-frequency regions which are different from those of other channels. Also, they show that members of Ω for each channel correctly correspond to the essential temporal frames while the silent frames are ignored. Comparing the Grad-CAM visualizations of GAP and SSRP indicates that SSRP not only focuses on the high-energy T-F regions but it also intensifies the responses of low-energy T-F regions. For instance, the focus on low-energy regions at low-frequencies of *clock tick* sample (mel indices in range of 0-10) or at high-frequencies of *crying baby* (mel indices in range of 30-40) is increased by SSRP. This denotes that many low-energy T-F regions may contain salient patterns which can contribute to the classification of different sound events, more than high-energy T-F regions. This is more apparent in the visualization obtained for *chainsaw* sample where the highest-energy regions exhibit lower contributions to the classification than low-energy T-F patterns that reside at mel indices in the range of 30-40 of the frames close to the end of signal.

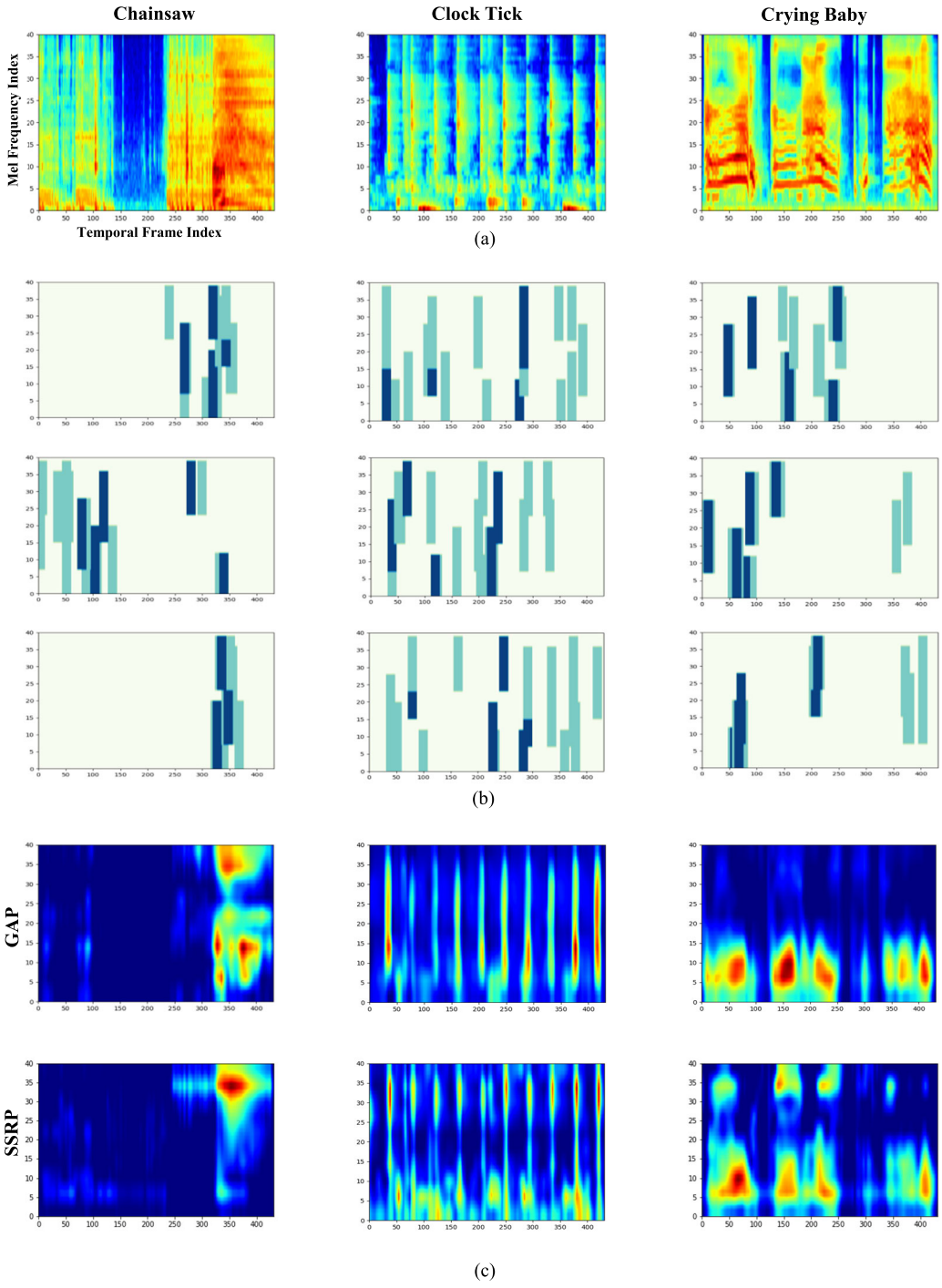

FIGURE 10. Illustration of the regions under focus of the network; (a) three sample input log mel spectrograms, (b) the produced masks for three sample channels (c) Grad-CAM visualizations [61] for GAP and SSRP-T with $K = 4$. The darker colors within each produced mask indicate the regions related to the highest mean of activations in each frequency band of $m_c^D(t, f)$.

V. CONCLUSION

In this paper, we proposed a global pooling method, called SSRP, which aims at enhancing the representation ability of a lightweight CNN via reducing the effect of less informative T-F regions on the training of convolutional kernels. By observing spectral and temporal characteristics of environmental sounds, SSRP restricts the features that contribute to channel descriptors to a subset of sparse features that makes the model learn from the more salient T-F regions. Experimental results on ESC-10 and ESC-50 demonstrated that our

CNN model equipped with SSRP achieves comparable performance to the state-of-the-art methods, with much smaller model size and much lower complexity. Our model makes absolute improvements of 14.3% and 21.8% in accuracy relative to the baseline models of ESC-10 and ESC-50 datasets, respectively. The obtained results denoted the importance of separately pooling each frequency band, which helps model distinguish between patterns that occur at different frequency bands. Moreover, our experiments revealed that the most salient information can reside in very short temporal intervals and thus imposing high degree of sparsity at each frequency band can significantly boost the model performance. We also performed visual analyses which indicated that the responses of low-energy salient T-F regions are also intensified by SSRP, and can contribute even more than high-energy T-F regions to the classification of specific sound classes. Considering the capability of the proposed model in ignoring silent and less informative frames, our plan for future work is to test the robustness of the model against different noise scenarios.

REFERENCES

[1] Y. Alsouda, S. Pllana, and A. Kurti, "A machine learning driven IoT solution for noise classification in smart cities," 2018, *arXiv:1809.00238*.

[2] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2005, pp. 158–161.

[3] Z. Yu, C. Licia, W. Ouri, and Y. Hai, "Urban computing: Concepts, methodologies, and applications," *ACM Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–55, Sep. 2014.

[4] T. Lv, H.-Y. Zhang, and C.-H. Yan, "Double mode surveillance system based on remote audio/video signals acquisition," *Appl. Acoust.*, vol. 129, pp. 316–321, Jan. 2018.

[5] L. Turchet, G. Fazekas, M. Lagrange, H. S. Ghadikolaei, and C. Fischione, "The Internet of Audio Things: State of the art, vision, and challenges," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 10233–10249, Oct. 2020.

[6] R. Nijhawan, S. A. Ansari, S. Kumar, F. Alassery, and S. M. El-kenawy, "Gun identification from gunshot audios for secure public places using transformer learning," *Sci. Rep.*, vol. 12, no. 1, pp. 1–10, Aug. 2022.

[7] V.-T. Tran and W.-H. Tsai, "Acoustic-based emergency vehicle detection using convolutional neural networks," *IEEE Access*, vol. 8, pp. 75702–75713, 2020.

[8] G. Gupta, M. Kshirsagar, M. Zhong, S. Gholami, and J. L. Ferres, "Comparing recurrent convolutional neural networks for large scale bird species classification," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, Aug. 2021.

[9] A. Krizhevsky and I. H. G. E. Sutskever, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural. Inf. Process. Syst.*, vol. 25, 2012, pp. 1–15.

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

[11] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2015.

[12] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4580–4584.

[13] S. Dieleman and B. Schrauwen, "End-to-end learning for music audio," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6964–6968.

[14] Y. M. G. Costa, L. S. Oliveira, and C. N. Silla Jr., "An evaluation of convolutional neural networks for music classification using spectrograms," *Appl. Soft Comput.*, vol. 52, pp. 28–38, Mar. 2017.

[15] W. Dai, C. Dai, S. Qu, J. Li, and S. Das, "Very deep convolutional neural networks for raw waveforms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 421–425.

[16] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," *Expert Syst. Appl.*, vol. 136, pp. 252–263, Dec. 2019.

[17] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *Proc. IEEE 25th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, Boston, MA, USA, Sep. 2015, pp. 1–6.

[18] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-stream network with temporal attention for environmental sound classification," 2019, *arXiv:1901.08608*.

[19] J. Sharma, O.-C. Granmo, and M. Goodwin, "Environment sound classification using multiple feature channels and attention based deep convolutional neural network," in *Proc. Interspeech*, 2020, pp. 1186–1190.

[20] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020.

[21] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Proc. Comput. Sci.*, vol. 112, pp. 2048–2056, Jan. 2017.

[22] T. Qiao, S. Zhang, S. Cao, and S. Xu, "High accurate environmental sound classification: Sub-spectrogram segmentation versus temporal-frequency attention mechanism," *Sensors*, vol. 21, no. 16, p. 5500, Aug. 2021.

[23] W. Mu, B. Yin, X. Huang, J. Xu, and Z. Du, "Environmental sound classification using temporal-frequency attention based convolutional neural network," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, Nov. 2021.

[24] H. Wang, Y. Zou, D. Chong, and W. Wang, "Environmental sound classification with parallel temporal–spectral attention," 2019, *arXiv:1912.06808*.

[25] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," *Neurocomputing*, vol. 453, pp. 896–903, 2021.

[26] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, May 2012.

[27] Y. Tokozume and T. Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2721–2725.

[28] B. Zhu, C. Wang, F. Liu, J. Lei, Z. Huang, Y. Peng, and F. Li, "Learning environmental sounds with multi-scale convolutional neural network," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2018, pp. 1–8.

[29] J. Wang and S. Li, "Self-attention mechanism based system for DCASE2018 challenge task1 and task4," in *Proc. DCASE Challenge*, 2018, pp. 1–5.

[30] J. Guo, N. Xu, L.-J. Li, and A. Alwan, "Attention based CLDNNs for short-duration acoustic scene classification," in *Proc. Interspeech*, 2017, pp. 469–473.

[31] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 171–175.

[32] Z. Zhang, S. Xu, S. Cao, and S. Zhang, "Deep convolutional neural network with mixup for environmental sound classification," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, 2018, pp. 356–367.

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

[34] X. Dong, B. Yin, Y. Cong, Z. Du, and X. Huang, "Environment sound event classification with a two-stream convolutional neural network," *IEEE Access*, vol. 8, pp. 125714–125721, 2020.

[35] A. M. Tripathi and A. Mishra, "Environment sound classification using an attention-based residual neural network," *Neurocomputing*, vol. 460, pp. 409–423, Oct. 2021.

[36] M. D. Zeiler and R. Fergus, "Stochastic pooling for regularization of deep convolutional neural networks," 2013, *arXiv:1301.3557*.

[37] S. Zhai, H. Wu, A. Kumar, Y. Cheng, Y. Lu, Z. Zhang, and R. Feris, "S3Pool: Pooling with stochastic spatial sampling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4970–4978.

[38] F. Saeedan, N. Weber, M. Goesele, and S. Roth, "Detail-preserving pooling in deep networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9108–9116.

[39] A. Kumar, "Ordinal pooling networks: For preserving information over shrinking feature maps," 2018, *arXiv:1804.02702*.

[40] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 695–711.

[41] Z. Shi, Y. Ye, and Y. Wu, "Rank-based pooling for deep convolutional neural networks," *Neural Netw.*, vol. 83, pp. 21–31, Nov. 2016.

[42] D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *Proc. Int. Conf. Rough Sets Knowl. Tech.*, 2014, pp. 364–375.
[43] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," in *Artificial Intelligence and Statistics*. Spain: PMLR, 2016.
[44] W. Wan, J. Chen, T. Li, Y. Huang, J. Tian, C. Yu, and Y. Xue, "Information entropy based feature pooling for convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3405–3414.
[45] B. Zhang, Q. Zhao, W. Feng, and S. Lyu, "AlphaMEX: A smarter global pooling method for convolutional neural networks," *Neurocomputing*, vol. 321, pp. 36–48, Dec. 2018.
[46] M. Luo, G. Wen, Y. Hu, D. Dai, and Y. Xu, "Stochastic region pooling: Make attention more expressive," *Neurocomputing*, vol. 409, pp. 119–130, Oct. 2020.
[47] X. Zhang and X. Zhang, "Global learnable pooling with enhancing distinctive feature for image classification," *IEEE Access*, vol. 8, pp. 98539–98547, 2020.
[48] H. R. Seresht, S. M. Ahadi, and S. Seyedin, "Spectro-temporal power spectrum features for noise robust ASR," *Circuits, Syst., Signal Process.*, vol. 36, no. 8, pp. 3222–3242, Aug. 2017.
[49] A. Almahairi, N. Ballas, T. Cooijmans, Y. Zheng, H. Larochelle, and A. Courville, "Dynamic capacity networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2549–2558.
[50] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 1015–1018.
[51] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
[52] 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.
[53] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.
[54] Z. Zhang, S. Xu, S. Zhang, T. Qiao, and S. Cao, "Learning attentive representations for environmental sound classification," *IEEE Access*, vol. 7, pp. 130327–130339, 2019.
[55] L. Nanni, G. Maguolo, S. Brahnam, and M. Paci, "An ensemble of convolutional neural networks for audio classification," *Appl. Sci.*, vol. 11, no. 13, p. 5796, Jun. 2021.
[56] Y. Aytar, C. Vondrick, and A. Torralba, "SoundNet: Learning sound representations from unlabeled video," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 1–15.
[57] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," 2017, *arXiv:1711.10282*.
[58] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, "An ensemble stacked convolutional neural network model for environmental event sound recognition," *Appl. Sci.*, vol. 8, no. 7, p. 1152, 2018.
[59] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," *Appl. Acoust.*, vol. 167, Oct. 2020, Art. no. 107389.
[60] Y. Su, K. Zhang, J. Wang, D. Zhou, and K. Madani, "Performance analysis of multiple aggregated acoustic features for environment sound classification," *Appl. Acoust.*, vol. 158, Jan. 2020, Art. no. 107050.
[61] 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)*, Oct. 2017, pp. 618–626.
[62] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 131–135.

HAMED RIAZATI SERESHT received the M.Sc. degree in electrical engineering from the Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2012, where he is currently pursuing the Ph.D. degree in electronic engineering. His current research interests include deep learning, artificial intelligence, environmental sound classification, and automatic speech recognition.

KARIM MOHAMMADI received the B.S. degree in electrical engineering from the Iran University of Science and Technology, Tehran, Iran, in 1972, the M.Sc. degree in electrical engineering from Wayne State University, Detroit, MI, USA, in 1978, and the Ph.D. degree in electrical engineering from Oakland University, Rochester, MI, USA, in 1981. He is currently a Professor in electrical engineering at the Iran University of Science and Technology. His current research interests include reconfigurable systems, fault-tolerant systems, reliable computing, digital systems, and microprocessors.