Article

Clustering by Errors: A Self-Organized Multitask Learning Method for Acoustic Scene Classification

Weiping Zheng [1,*], Zhenyao Mo [2] and Gansen Zhao [1]

[1] School of Computer Science, South China Normal University, Guangzhou 510631, China; gzhao@m.scnu.edu.cn
[2] School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China; csmozhenyaomyz@mail.scut.edu.cn
[*] Correspondence: zhengweiping@scnu.edu.cn

Abstract: Acoustic scene classification (ASC) tries to inference information about the environment using audio segments. The inter-class similarity is a significant issue in ASC as acoustic scenes with different labels may sound quite similar. In this paper, the similarity relations amongst scenes are correlated with the classification error. A class hierarchy construction method by using classification error is then proposed and integrated into a multitask learning framework. The experiments have shown that the proposed multitask learning method improves the performance of ASC. On the TUT Acoustic Scene 2017 dataset, we obtain the ensemble fine-grained accuracy of 81.4%, which is better than the state-of-the-art. By using multitask learning, the basic Convolutional Neural Network (CNN) model can be improved by about 2.0 to 3.5 percent according to different spectrograms. The coarse category accuracies (for two to six super-classes) range from 77.0% to 96.2% by single models. On the revised version of the LITIS Rouen dataset, we achieve the ensemble fine-grained accuracy of 83.9%. The multitask learning models obtain an improvement of 1.6% to 1.8% compared to their basic models. The coarse category accuracies range from 94.9% to 97.9% for two to six super-classes with single models.

Keywords: acoustic scene classification; convolutional neural network; acoustic scene clustering; multitask learning; late fusion

1. Introduction

Acoustic scene classification (ASC) refers to the task of associating a semantic label to an audio stream that identifies the environment in which it has been produced [1]. This task takes as input a relatively long sound clip and outputs predicted acoustic scene class, e.g., home, park, and bus. Classifying scenes by audio data has its unique advantages. The recording of audio data is not restricted by the camera angle and illumination condition, etc. As a result, the equipment for sound collection can be installed in a wider range where object occlusion is no more a problem. The collection can run indiscriminately in a dark environment. Moreover, the storage cost of audio data is relatively low compared to image or video data. Recently, ASC has shown huge potentials in many industrial and business applications [2,3], such as surveillance, life-logging, and advanced multimedia retrieval [4].

The inter-class similarity is a common challenge in machine learning research [5]. However, it is getting more prominent in ASC, as the labels of the scenes are commonly annotated according to the spatial functions related to the place where the audio segments are recorded. Consequently, there are audio segments that are quite similar in terms of acoustical characteristics while assigned with different labels, e.g., the segments of a library and those of an office. It is therefore a challenging task to distinguish these similar scenes even for humans and they are often misclassified by the machine learning algorithms.

Furthermore, in most cases, misclassification occurs among similar scenes. For example, in the TUT Acoustic Scenes 2017 dataset [6], the scene of the beach is misclassified in most cases as a residential area. Additionally, home is frequently misclassified as a library, and so on. These errors seem forgivable considering the similarities existing among the scenes. For example, pedestrians, laughter, blowing wind, and other sounds exist in both beach and residential areas. The scenes of home and library may have common aspects, for instance, the quietness, low-voice speaker, and phone ring. Hence, acoustic scenes tend to be misclassified as those having similar characteristics.

Based on the above, here we propose learning the similarity relations of acoustic scenes by taking advantage of the classification errors. In our method, we use the spectral clustering algorithm on the confusion matrix. The scene (class) set of a certain acoustic scene dataset is then divided into several subsets according to the similarity relation of the corresponding acoustic scenes. Each subset is assigned a super-class label. Using this approach, a two-level class hierarchy can be easily built in the label space of the acoustic scene dataset.

Ye et al. [7] proposed an acoustic event taxonomy construction approach based on between-dictionary distances. Li et al. [8] also proposed an acoustic scene clustering method using agglomerative hierarchical clustering on deep embedding extracted by Convolutional Neural Network (CNN). Their taxonomy heavily depends on the quality of acoustic feature learnt and the distance metric selected. Conversely, our construction approach is a simple solution that does not need any feature embedding.

In this paper, the two-level class hierarchy is further integrated into a multitask learning framework for ASC. To take advantage of the relevance between the super-class (coarse category) and original class (fine-grained category), a regularization method is adopted to optimize the training. Note that multitask learning is not a new idea for ASC. Tonami et al. [9] proposed a multitask learning-based solution for joint analysis of acoustic events and scenes where each sample was given both scene and event labels through manual annotation. Abrol et al. [10] also proposed a multitask model which was trained with hierarchical coarse and fine labels for ASC. They manually created a two-level class hierarchy by arranging the fine scene classes into coarse classes. In our proposed self-organized multitask learning method, the original label space is organized into a hierarchical structure by learning the similarity relationship from the confusion matrix. In our method, manual annotation is not required, and the class hierarchy is constructed automatically solely based on the original dataset. This is the reason that the proposed method is called "self-organized" multitask learning.

The proposed method is evaluated comprehensively on two publicly available datasets including the TUT Acoustic Scenes 2017, and the LITIS Rouen [11] datasets. As shown in Figure 1, there are 15 acoustic scenes and 3 super-classes in the TUT Acoustic Scenes 2017 dataset. It is arranged as a two-level class hierarchy by the original dataset publishers. We compare the constructed class hierarchy with the original one in the experiment. The LITIS Rouen dataset provides single-level classes. The experiment demonstrates that a single-level class dataset can also benefit from the proposed method.

The contributions of this paper are as the following:

(1) To the best of our knowledge, we are the first to automatically construct a taxonomy for acoustic scenes by learning similarity relationships from classification errors.

(2) By incorporating the constructed two-level class hierarchy, the proposed self-organized multitask learning method improves the performance of ASC.

The rest of the paper is organized as follows: Section 2 introduces the related works. Section 3 describes the proposed method. Experimental results and analyses are presented in Section 4. Discussion is provided in Section 5. Finally, we conclude in Section 6.
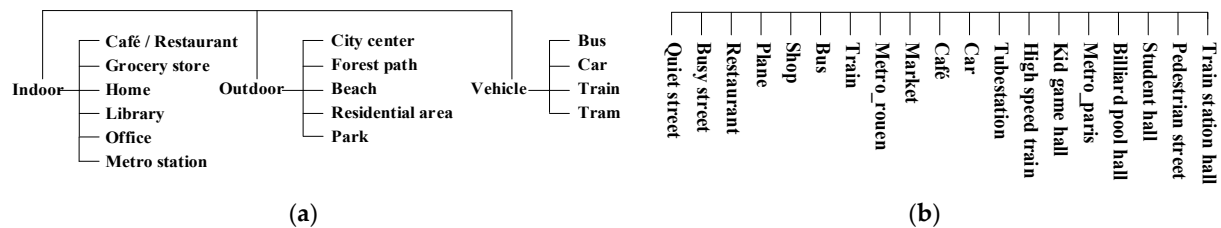
Indoor
- Café / Restaurant
- Grocery store
- Home
- Library
- Office
- Metro station

Outdoor
- City center
- Forest path
- Beach
- Residential area
- Park

Vehicle
- Bus
- Car
- Train
- Tram

(**a**)

Quiet street, Busy street, Restaurant, Plane, Shop, Bus, Train, Metro_rouen, Market, Café, Car, Tubestation, High speed train, Kid game hall, Metro_paris, Billiard pool hall, Student hall, Pedestrian street, Train station hall

(**b**)

Figure 1. Different organizations of the label spaces. The label space of TUT Acoustic Scenes 2017 dataset is originally organized as a two-level hierarchy, however, merely single-level labels are provided in the LITIS Rouen dataset. (**a**) Labels of the TUT Acoustic Scenes 2017 dataset. (**b**) Labels of the LITIS Rouen dataset.

2. Related Works

Audio classification has become a hot topic in the field of signal processing. As an essential part of the audio classification, ASC has been one of the main tasks in the IEEE DCASE Challenges (2013, 2016–2021). In the conventional ASC techniques, cepstrum coefficients, as well as other handcrafted audio features, are classified by the Gaussian mixture models (GMM), hidden Markov models (HMM) and support vector machine (SVM) methods [12]. For example, Ma et al. [13] used a hierarchical HMM-based model fed by MFCC features to classify the environmental sounds. Chakrabarty et al. [14] further proposed a hybrid GMM–HMM system, where average modulation statistics of the scene provided by the GMM and temporal trajectories of the modulations obtained by the HMM are fused to achieve better classification performance. Recently, deep learning techniques have also achieved impressive results on acoustic scene classification [15], e.g., CNN [16–19], RNN [20], LSTM [21], DNN [22,23], and their combinations [21]. Among these models, CNN is the most popular architecture which has shown promising results in most recent works. For instance, Eghbal-Zadeh et al. [16] proposed an approach using deep CNN and binaural i-vectors [24] for ASC. Bae et al. [21] also presented a structure composed of parallel CNN and LSTM [25] networks, which aimed to extract both spectro-temporal locality and sequential information.

As in the ASC, inter-class similarity has been a serious problem for fine-grained visual recognition. To distinguish the degree of similarity among classes, the common solution is to use triplet loss [26], quadruplet loss, or N-pair-mc loss [27]. In these methods, the hierarchical relation of classes should be carefully arranged. The selections of anchor points in these methods are challenging. Nevertheless, these approaches often require a larger number of training samples and result in more complex optimizations. Multitask learning is another solution that needs hierarchical labels. Xie et al. [28] studied the large intra-class and small inter-class variance in fine-grained image classification and proposed a multitask learning framework. Zhang et al. [26] also designed a multitask framework to learn fine-grained feature representations. In their framework, hierarchy and share attributes are embedded by optimizing both classification and similarity constraints. Wu et al. [29] further formulated a multitask loss on CNN architecture to utilize the semantic relationships among food categories.

In this paper, we focus on the inter-class similarities problem in ASC and use a similar multitask learning solution as in fine-grained visual recognition. Hierarchical labels were also utilized in [10,22] for multitask learning of ASC. Nevertheless, in those works, the problem of automatic construction of class hierarchy was not considered. Recently, multitask learning approaches have been proposed to perform a joint analysis of acoustic scenes and events [9,30]. However, in these methods, the datasets should be intentionally prepared using audio synthesis or additional manual annotation.

The size of datasets in ASC is relatively small thus it may lead to over-fitting. High-quality labeled data are expensive to acquire, especially for audio data. To increase the size of the datasets, data augmentation is widely used and proven as effective practice

[31,32]. Salamon et al. [32] augmented the data by deforming the audio signal directly before converting it into log-Mel spectrograms. The applied deformation included time stretching, pitch shifting, dynamic range compression, and background noise.

Nevertheless, not all augmentation techniques are helpful. Those samples augmented far from their original ones are harmful to the classification performance. To solve this problem, Lu et al. [33] proposed a metric learning-based framework to ensure appropriate augmentation for the appropriate training data. In [31], a GAN-based [34] method was used to generate additional samples for ASC. These samples were selected by an SVM hyperplane to ensure augmentation quality.

Zhong et al. [35] proposed a random erasing method for CNN data augmentation. In the method, a rectangle region is randomly selected within an image. The pixels in the region are erased with random values. This method is easy to implement, and random erasing keeps most of the information in the original image. As a result, the filtering operation to remove the harmful augmented samples performed in [31,33] is not necessary here. Gharib et al. [36] applied a similar random erasing method for ASC and achieved an improvement of 0.13 percent compared with their baseline system.

Mixup [37] is another interesting data augmentation method. This method constructs a new example using a linear interpolation of two random examples from the training set and their labels. Xu et al. [18] used a multi-channel CNN in ASC and applied mixup to improve prediction accuracy. In this paper, a class hierarchy construction method is proposed which appends super-class labels for the training examples. The method does not increase the size of the dataset. However, it extends the label space of the samples, hence providing more information. It is demonstrated in our experiment that using multitask learning with a two-level class hierarchy can effectively enhance the generalization of the CNN model. Although both mixup and class hierarchy construction bring changes into the label space, mixup modifies the labels from the one-hot into the ratio type, whereas the class hierarchy construction provides more one-hot labels by constructing super-class labels.

3. Proposed Method

3.1. Overview

In this paper, we propose a self-organized multitask learning method. The proposed solution includes two stages: a two-level class hierarchy is automatically constructed in the first stage using a basic model. The final classifier is then obtained by training a multitask learning model using the constructed super-class labels and the original fine-grained labels in the second stage. As shown in Figure 2, the proposed method for ASC includes the following four steps:

(1) Preparing spectrograms: transforming the raw audio segments into spectrograms that are suitable for CNN models.
(2) Getting a basic model: training a single-task CNN model as a basic model using the spectrograms and original fine-grained scene labels.
(3) Constructing a class hierarchy: testing the validation set on the basic model to obtain a confusion matrix. The spectral clustering is performed on the confusion matrix to generate super-classes.
(4) Getting the final model: training a multitask CNN model as the final classifier to predict both the original scene class and the constructed super-class using hierarchical labels.
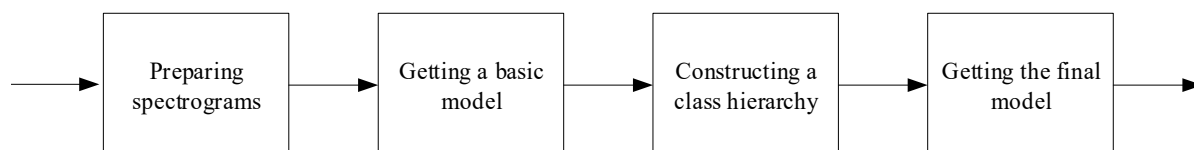
Figure 2. Flowchart of the proposed method.

3.2. Spectrograms Generation.

To apply CNN models, spectrograms are generated from the audio segments using certain signal processing methods, e.g., the Short-Time Fourier Transform (STFT) [38], Constant-Q-Transform (CQT) [39], and Mel Frequency Cepstral Coefficients (MFCC) [40]. They are split into multiple samples and fed into the CNN model. The spectrogram is considered as a time-frequency representation of the acoustic scene [41]. As CNN is effective in learning spatially local correlations from images, it can use the spatial and temporal information in the spectrograms. However, different spectrograms have different processing abilities for the corresponding frequency range, which can be used for characterizing different acoustic scenes. Therefore, CNN models are widely used as the deep feature extractor, and multiple spectrogram fusions are usually applied in ASC for performance enhancement [17,42,43].

In this paper, we generate three kinds of spectrograms, STFT, CQT, and log-Mel spectrograms, and then evaluate the proposed method on these different presentations, respectively. Details about the spectrogram generation are described in Section 4.1.

3.3. Basic Model

Using these spectrograms with fine-grained labels, a CNN model is trained to classify the acoustic scenes. The trained CNN model is referred to as the basic model. A VGG-like network [17] is considered here as the basic model and its structure is illustrated in Table 1. After training on different spectrograms, various basic models become available. Specifically, three basic CNN models are evaluated in the paper, namely the VGG-STFT, VGG-CQT, and VGG-Log-Mel models.

Table 1. Architecture of the proposed CNN network.

Layer	Conv1	Conv2	Pool1	Conv3	Conv4	Pool2	Conv5	Conv6	Conv7	Conv8	Pool3	Conv9	Conv10	Conv11	Full1
Kernel	5×5	3×3	Max, 2×2	3×3	3×3	Max, 2×2	3×3	3×3	3×3	3×3	Max, 2×2	3×3	1×1	1×1	–
Stride	2	1	2	1	1	2	1	1	1	1	2	1	1	1	–
Padding	2	1	0	1	1	0	1	1	1	1	0	0	0	0	–
Number of Channels	32	32	32	64	64	64	128	128	128	128	128	512	512	C	C
Dropout rate	–	–	0.3	–	–	0.3	–	–	–	–	0.3	0.5	0.5	–	–
Activation	ReLu	ReLu	–	ReLu	ReLu	–	ReLu	ReLu	ReLu	ReLu	–	ReLu	ReLu	ReLu	–
Batchnorm	Yes	Yes	–	Yes	Yes	–	Yes	Yes	Yes	Yes	–	Yes	Yes	Yes	–

Note: Symbol C in the last two columns represents the number of classes.

Without loss of generality, the architecture of CNN and the spectrogram are not specified below. Suppose we are given a set of n training samples $TS = \{(x_1, y_1^o), \dots, (x_n, y_n^o)\}$ with $y_i^o \in \{1, \dots, C\}$ indicating the fine-grained acoustic scene class label of image, $x_i, i \in [1, n]$ (namely a spectrogram patch); superscript o denotes original labels of the dataset.

The CNN network consists of multiple convolutional and pooling layers. At the end of the network, the output layer uses a softmax activation function to assign probabilities to each possible class, where there are C nodes in the output layer. Let $P(y_i^o|x_i)$ be the probability corresponding to the true-ground class of x_i. There are also L nodes in the next-to-last layer, which are mapped to C nodes using the fully connected layer. Let $W_{v,u}(v \in [1, C]; u \in [1, L])$ denote the weights of connections between these two layers. A negative log-likelihood loss is adopted in the basic model, i.e.,

$$Loss(W) = \frac{1}{n} \sum_{(x_i, y_i^o) \in TS} (-\log P(y_i^o|x_i, W)) + \alpha ||W||_2^2 \qquad (1)$$

3.4. Super-Class Labels Construction

Constructing the super-classes by merging similar acoustic scenes is a natural and straightforward method. However, it is difficult to find out scenes with similar acoustical properties. Generally, the audio segments are transformed into a certain kind of embedding, and distances defined on the embedding spaces are used to group the scenes into coarse categories. However, learning of embedding and distance definition is not easy, and the clustering results are hard to explain.

In our research, we use misclassification information to approximate the similarities among classes. Specifically, a certain set (e.g., the validation set) of samples is evaluated on a basic model. These predicted results are counted into a confusion matrix. Finally, a spectral clustering algorithm is applied to construct super-class labels and thus expand the label space. The pipeline is illustrated in Figure 3.
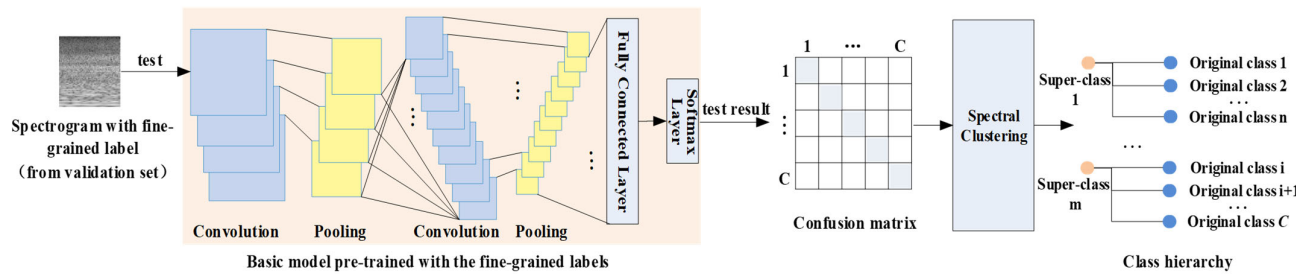
Figure 3. The pipeline of the construction of super-class labels. Each sample in the validation set is evaluated by the trained basic CNN model. The prediction results are collected to compute a confusion matrix. Then, a spectral clustering algorithm is performed upon this matrix. As a result, original acoustic scene classes are clustered into several subsets. The classes in the same subset are considered as similar scenes and the same super-class label is assigned to their corresponding samples.

For a certain basic CNN model, such as the VGG-STFT, a confusion matrix F can be calculated, with $F_{ci,cj}$ denoting the number of the samples of class ci that are classified as class cj by that model. After pre-processing, F is transformed to a matrix D to ensure symmetry:

$$D = (F + F^T)/2 \tag{2}$$

Using this matrix D, we apply spectral clustering [44] to divide the original C classes into N subsets $H_1, \dots, H_N$, $H_1 \cup \dots \cup H_N = \{1, \dots, C\}$; $H_{hi} \cap H_{hj} = \emptyset (hi \neq hj; hi, hj \in [1, N])$. The proposed clustering algorithm is provided in Algorithm 1.

Each subset is assigned a super-class label. The TS can be rewritten as $TS = \{(x_1, < y_1^o, y_1^e >), \dots, (x_n, < y_n^o, y_n^e >)\}$ with $y_i^e \in \{1, \dots, N\}$ indicating the super-class label of $x_i, i \in [1, n]$, where superscript e denotes the expanded label. Therefore,

$$\forall i \forall j \exists m ((y_i^o \in H_m \wedge y_j^o \in H_m) \rightarrow y_i^e = y_j^e) \tag{3}$$

The number of super-classes in the above construction (i.e., N) is a hyperparameter and is selected using experiments.

Algorithm 1. Clustering algorithm in super-class generation.

Input: the confusion matrix F, number of clusters N.

Output: super-class clusters $H_1, \dots, H_N$

1. Set the diagonal elements of F to zero:
    $F_{k,k} = 0, k \in \{1, \dots, C\}$
2. Normalize each row of F by the following equations:

$$SM_k = \sum_{d=1}^{C} F_{k,d}$$

$$F_{k,t} = F_{k,t}/SM_k \ , \ k,t \in \{1,...,C\}$$

3. Transform F into a symmetric matrix D:
    $$D = (F + F^T)/2$$
4. Assume that B is a diagonal matrix whose elements are set as:
    $$B_{e,e} = \sum_{q=1}^{C} D_{e,q}, \ e \in \{1,...,C\}$$
5. Construct the Laplacian matrix G by:
    $$G = B - D$$
6. Calculate the eigenvectors $a_1, a_2, ..., a_N$ corresponding to the N smallest eigenvalues of G.
7. Let $A \in \mathbf{R}^{C \times N}$ be the matrix containing $a_1, a_2, ..., a_N$ as columns; let $r_i \in \mathbf{R}^N$ be the i-th row of A.
8. Use K-means algorithm to cluster $\{r_1, r_2, \ ..., r_C\}$ into N clusters $RC_1, ..., RC_N$.
9. Output clusters $H_1, ..., H_N$ with $H_{hi} = \{z | r_z \in RC_{hi}\}, \ hi \in \{1, ..., N\}$.

3.5. Multitask Learning Model

In the second stage, the constructed two-level class hierarchy is incorporated into a multitask learning framework. The structure of the multitask learning model is illustrated in Figure 4. As it is seen the discrimination of super-class has become an additional task in the classification process. Consequently, the basic model is transformed into a multitask learning paradigm. The details are provided in the following.

Multitask learning model for final prediction
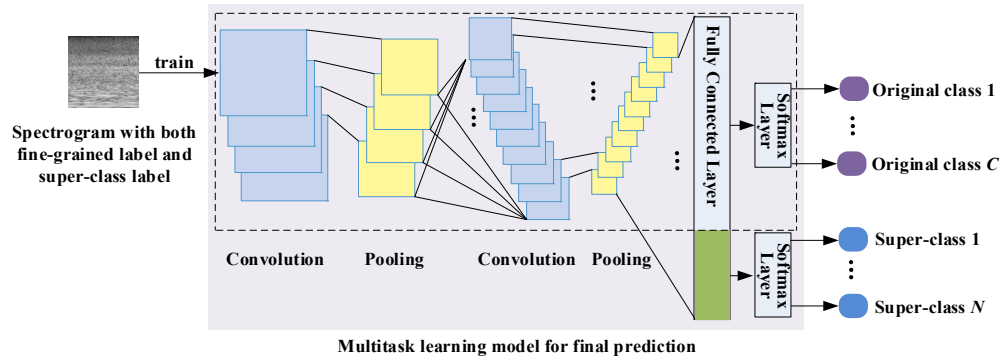
Figure 4. Structure of the multitask learning model. The structure within the dashed box is identical to the basic CNN model. An additional softmax layer has been added and fully connected to the next-to-last layer. The green area above represents the newly added connections in the fully connected layer. Both original acoustic scene labels and super-class labels are simultaneously predicted here.

To make the model aware of the inter-class similarities, we add another output layer onto the basic CNN model, leaving all other details of the model unchanged. The newly added layer has N output nodes and is fully connected onto the original next-to-last layer. The weights of the newly added connections are denoted as $U_{mj,mi} (mj \in [1,N]; mi \in [1,L])$. We then update the reconstruction error of the new model into a multitask learning form as the following:

$$E = \sum_{(x_i, <y_i^o, y_i^e>) \in TS} -(\gamma \log P(y_i^o | x_i, W; U) + (1-\gamma) \log P(y_i^e | x_i, W; U)) \tag{4}$$

where $\gamma \in [0,1]$ controls the proportion between the original task and the new task in the reconstruction error. The weight vector $W_t = (W_{t,1}, ..., W_{t,L})$ for original class t should capture similar high-level patterns [28,45] as the weight vector for the super-class $s(t)$ of class t, i.e., $U_{s(t)} = (U_{s(t),1}, ..., U_{s(t),L})$. Therefore, we introduce the following regularization into the loss function:

$$R = \sum_{t=1}^{C} ||W_t - U_{s(t)}||_2^2 \qquad (5)$$

Finally, the loss function of the new model can be defined as:

$$Loss_{ML}(\mathrm{W};\mathrm{U}) = E/n + \alpha \cdot R + \beta \cdot ||W;U||_2^2 \qquad (6)$$

where α and β are set to 0.0001.

After performing the self-organized multitask learning method, respectively, we can obtain boosted models, e.g., VGG-STFT-ML (from VGG-STFT), VGG-CQT-ML (from VGG-CQT), and VGG-Log-Mel-ML (from VGG-Log-Mel). As expected, our experiments confirm that the updated model outperforms the basic models.

Note that the CNN model is a building block in the proposed framework. It can be replaced by any other popular CNN architecture, such as ResNet [46] and GoogleNet [47]. The backbone of the multitask learning network is not necessary to be restricted by the basic model. We keep most of the layers unchanged in multitask learning model to facilitate performance comparison.

Furthermore, the class hierarchy construction and self-organized multitask learning approaches are not limited to CNN. Hence, similar ideas apply to other models such as RNN/LSTM, DNN, and DBN, and might be suitable for other applications.

4. Experiments and Results

4.1. Experiment Setup

The TUT Acoustic Scenes 2017 dataset [6] and LITIS Rouen dataset [11] (a revised version) are selected to evaluate the performance of our method. The TUT Acoustic Scenes 2017 dataset includes Development and Evaluation sets. We have trained the model on the Development set and evaluated it on the Evaluation set. We also follow the four-fold split provided by the dataset publishers.

The LITIS Rouen dataset is one of the commonly used publicly available datasets for ASC. However, it tends to provide over-optimistic results as some examples cut from the same long recordings are distributed into the training set and test set, respectively. To avoid the "album effect", Rakotomamonjy [48] had created a corrected version of the LITIS Rouen dataset, namely Rouen-15. However, it was not made public. To this end, we create a revised version by ourselves in this paper. We first merge the 3026 examples into 487 recording files according to the mapping relation presented by the dataset provider on their website. The dataset is then divided into the training set and test set by file. Similarly, four-fold cross-validation is performed. The training set is further split into four folds by file as well. At last, these files are restored into examples. Specifically, 684 examples are selected as a test set, while 2342 examples are used for training, which are split into 641, 617, 545, and 539 examples, respectively.

Three kinds of spectrograms are generated for the evaluation experiments including STFT, CQT, and log-Mel spectrograms. Spectrograms are generated for each channel (left and right) from the audio clips. To generate STFT spectrograms, the window size is set to 16 ms (706 points) and the hop length is 9.75 ms (430 points) at 44.1 KHz for the TUT Acoustic Scenes 2017 dataset. For the LITIS Rouen dataset, the window size is 32 ms (706 points) and the hop length is 19.5 ms (430 points) at 22.05 KHz.

Logarithmic power spectral densities ($10\log_{10}PSD$) are utilized to plot the spectrograms, which are generated in a one-sided fashion. The sizes of the spectrograms are 1024 × 354 pixels and 1537 × 354 pixels for the two datasets. The spectrograms are divided into patches with a width of 143 pixels and a shift step of 126 pixels. The size of each patch is 143 × 354 pixels. Therefore, we obtain 8 and 12 patches for each spectrogram on the two datasets, respectively.

CQT spectrograms are generated using a python library, Librosa 0.5.0. In the generation function, the sampling rate is set as 22.05 KHz, the filter scale is set to 2, and

the frequency bin is 110. Other parameters are set to their default values. The CQT spectrograms with the sizes of 862 × 110 pixels and 1292 × 110 pixels are generated for the TUT Acoustic Scenes 2017 and the LITIS Rouen, respectively. The spectrograms are split into patches with a 143-pixels width and an 80-pixels shift step. Consequently, we obtain 10 and 15 patches for each spectrogram per channel on the two datasets, respectively.

We then extract log-scaled Mel-spectrograms with 128 Mel-bands, using a window size of 92.8 ms (2048 points at 22.05 KHz) and a hop length of 46.4 ms. The sizes of the log-Mel spectrograms for the two datasets are 430 × 128 pixels and 646 × 128 pixels, respectively. The patch width is set as 143 pixels and the shift step is 71 pixels. We can generate 5 and 8 patches from each spectrogram on the two datasets. All patches are resized into 143 × 143 pixels before they are fed into the CNN networks. In addition, patches derived from both channels are separately treated as samples. Note that the above settings of hop lengths and shift steps are decided and justified in our previous work [17] and reused here for convenience.

The experiments are implemented using the TensorFlow [49] platform. A mini-batch size of 256 is used as well as an early stopping strategy with a patience parameter of 30 and a maximum epoch of 200. We use Adam [50] optimizer with a learning rate of 0.0001. In the following experiments, γ in Equation (4) is set to 0.6 for all the multitask models. An example-level majority voting accuracy is selected as a performance metric in the following experiments.

4.2. Selection of Super-Class Number

In the self-organized multitask learning models, the number of super-class is an important parameter that is closely related to the performance. For a given dataset with C classes, the maximum and the minimum super-class number is C-1 and 2, respectively. One way is to train and test the multitask learning models with all the possible super-class numbers and select the model with the highest accuracy. This, however, causes substantial waste of computing resources. To observe the differences, we have implemented the multitask learning model using CQT spectrograms on both datasets with all super-class numbers. Their corresponding accuracies are presented in Figures 5 and 6.
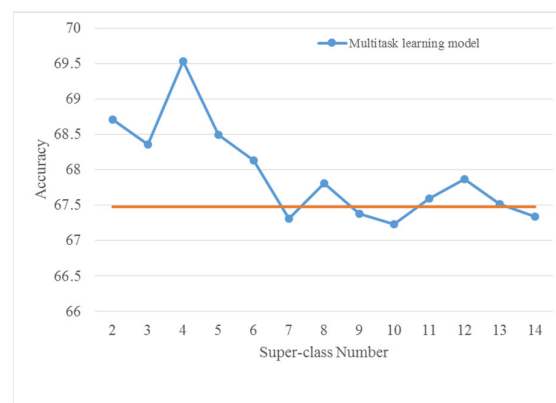
Figure 5. Accuracies of multitask learning models with different super-class numbers using CQT spectrograms on the TUT Acoustic Scenes 2017 dataset. The orange horizontal line means the accuracy of the basic model (VGG-CQT).
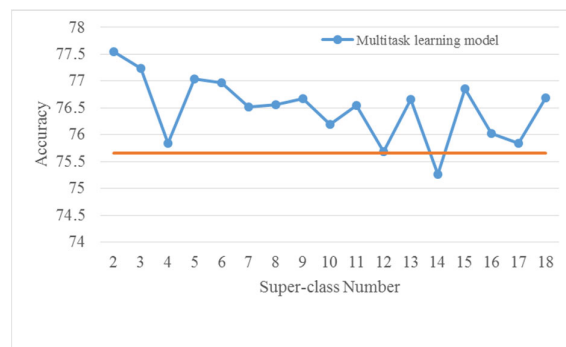
Figure 6. Accuracies of multitask learning models with different super-class numbers using CQT spectrograms on the LITIS Rouen dataset. The orange horizontal line means the accuracy of the basic model (VGG-CQT).

As it is seen in Figure 5, applied on the TUT Acoustic Scenes 2017 dataset the previous five models (with super-class numbers 2, 3, 4, 5, and 6) achieve higher accuracies than that of other models. In other words, the multitask learning model with a very large super-class number is not competitive in performance. Similarly, in Figure 6, the previous five models (except for the one with 4 super-classes) provide higher accuracies on the LITIS Rouen dataset. Although the decline of performances with the increase in super-class numbers is not apparent, the best results are still achieved with relatively small super-class numbers. Consequently, only the multitask learning models with two to six super-classes are explored in the following experiments.

Note that in Figures 5 and 6, it can be seen that the accuracies of the multitask learning models outperform the basic model in most cases.

4.3. Evaluations on the TUT Acoustic Scenes 2017 Dataset

Using the proposed method, the basic CNN models are trained on the STFT, CQT, and log-Mel spectrograms. Based on a specific basic model, a confusion matrix is generated for each validation set. To obtain more stable divisions, we repeat the process three times. The final confusion matrix used is calculated using the sum of the twelve confusion matrices (three times × four splits), as shown in Figure 7.

Based on the confusion matrices, the original 15 acoustic scenes are grouped into two to six super-classes. The division details are listed in Table 2. For example, using the confusion matrix generated by the STFT basic model (Figure 7a), the 15 acoustic scenes can be divided into two super-classes: the classes of bus, car, train, and tram are grouped as one super-class (represented as the blue squares); the other 11 classes are grouped as another super-class (represented as the red circles). Compared with the original super-classes (namely Indoor, Outdoor, and Vehicle categories, see Figure 1a), there are some interesting findings with the constructed divisions. First, for the two super-classes' divisions, each of them keeps one original super-class and merges the other two into another super-class. For the three super-classes' divisions, the results for CQT and log-Mel model are identical to the original Indoor, Outdoor, and Vehicle divisions. The four super-classes' divisions for STFT, CQT, and log-Mel models are the same. In addition, the five super-classes' divisions for CQT and log-Mel models are also identical. In fact, in most cases, the divisions are very similar to each other. The above results confirm the robustness of the class hierarchy construction method.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	11,739	75	55	250	351	310	35	182	97	2	19	491	1,167	134	69
2	99	10,748	218	254	16	1	53	81	28	9	5	91	227	1,111	2,035
3	198	328	7,912	45	581	57	2,489	555	326	853	142	191	291	576	432
4	197	235	3	13,609	2	0	0	49	68	0	27	48	86	334	318
5	260	45	268	50	12,177	8	74	11	1	522	2	145	1,326	46	41
6	123	1	46	8	12	12,816	95	266	66	8	89	291	1,095	48	12
7	90	34	1,219	2	49	40	12,231	50	214	776	48	86	24	48	65
8	158	142	210	26	7	231	94	11,203	1,893	24	823	175	121	90	67
9	513	18	133	18	8	194	239	2,056	9,925	49	1,116	184	191	200	132
10	4	14	96	0	62	23	514	14	141	14,041	21	3	15	19	9
11	23	31	17	13	2	81	36	722	777	70	12,546	451	147	35	25
12	766	82	170	30	407	542	93	53	250	13	177	9,533	2,684	89	87
13	567	158	119	64	1,306	718	40	64	252	33	119	2,292	9,146	52	46
14	156	899	578	233	316	79	26	91	79	24	18	90	995	9,528	1,864
15	54	1,013	630	415	233	18	39	35	61	15	21	71	173	940	11,258

(**a**)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	15,041	26	38	219	235	467	25	227	328	27	62	852	1,102	35	36
2	30	13,715	164	607	110	19	217	304	282	674	14	329	168	1,142	945
3	297	395	12,875	3	326	63	1,506	1,527	272	957	156	20	140	130	53
4	93	748	2	16,227	2	0	0	49	7	413	1	55	42	365	716
5	161	42	266	5	13,527	295	19	124	31	683	4	611	2,889	36	27
6	587	2	47	2	287	17,123	8	46	13	29	11	60	445	59	1
7	7	324	810	2	85	31	16,363	109	221	559	112	17	51	21	8
8	82	211	600	104	115	146	124	13,689	2,720	197	684	50	106	110	142
9	1,004	191	453	91	52	235	294	1,416	12,189	1,169	252	239	562	437	136
10	16	307	456	62	949	16	347	63	1,030	14,289	229	324	151	147	334
11	19	8	117	2	29	15	142	1,249	280	204	16,159	66	292	90	48
12	1,055	151	63	15	1,000	106	38	226	174	205	13	11,713	3,646	251	64
13	625	125	130	51	2,617	546	35	93	337	220	60	3,090	10,619	144	28
14	164	1,082	397	775	166	73	31	362	235	454	37	195	1,051	12,815	883
15	16	1,014	17	720	20	2	26	39	39	205	5	242	117	429	15,829

(**b**)

(**c**)

Figure 7. Confusion matrices on the TUT Acoustic Scenes 2017 dataset generated by the basic models (**a**) using STFT spectrograms; (**b**) using CQT spectrograms; and (**c**) using log-Mel spectrograms. The abscissa represents the predicted label, and the ordinate indicates the true label. Note: Beach (1), bus (2), café/restaurant (3), car (4), city center (5), forest path (6), grocery store (7), home (8), library (9), metro station (10), office (11), park (12), residential area (13), train (14), and tram (15).

Table 2. Clustering details of the constructed super-class for the TUT Acoustic Scenes 2017 dataset. The classes marked by the same shape with the same color are grouped into the same super-class.

Clustering Scheme	Supper-Class for STFT Model			Supper-Class for CQT Model			Supper-Class for Log-Mel Model	
		1	2	3	1	2	3	1	2
		4	5	6	4	5	6	4	5
Two Super-Classes		7	8	9	7	8	9	7	8
		10	11	12	10	11	12	10	11
		13	14	15	13	14	15	13	14
		1	2	3	1	2	3	1	2
		4	5	6	4	5	6	4	5
Three Super-Classes		7	8	9	7	8	9	7	8
		10	11	12	10	11	12	10	11
		13	14	15	13	14	15	13	14

In general, we rate the divisions for the log-Mel model as the best divisions. For example, its three super-classes' division is identical to the original division, where the two super-classes' division merges the Outdoor and Vehicle into one super-class, which seems more reasonable. It is believed that the superiority in divisions is due to the high performance of the log-Mel basic model (see Table 3). Hence, the basic classifier and the evaluated samples for confusion matrix creation should be well-chosen.

Table 3. Classification performance of the TUT Acoustic Scenes 2017 dataset using different models.

Model Type	Feature Type	Super-Class Number	15 Scenes Accuracy	Super-Class Accuracy
Basic Model	STFT	/	60.0 ± 0.5	/
Multitask	STFT	2	61.1 ± 0.3	93.1 ± 0.6
Multitask	STFT	3	62.0 ± 0.1	88.8 ± 0.4
Multitask	STFT	4	61.1 ± 0.1	85.9 ± 0.8
Multitask	STFT	5	61.5 ± 0.5	82.6 ± 0.2
Multitask	STFT	6	60.7 ± 0.2	77.0 ± 0.2
Basic Model	CQT	/	67.5 ± 0.2	/
Multitask	CQT	2	68.7 ± 0.2	96.2 ± 0.1
Multitask	CQT	3	68.4 ± 0.3	92.5 ± 0.4
Multitask	CQT	4	69.5 ± 0.8	89.3 ± 0.7
Multitask	CQT	5	68.5 ± 0.4	84.8 ± 0.4
Multitask	CQT	6	68.1 ± 0.6	84.0 ± 0.6
Basic Model	log-Mel	/	69.3 ± 0.1	/
Multitask	log-Mel	2	71.5 ± 1.0	94.9 ± 0.1
Multitask	log-Mel	3	72.1 ± 0.4	93.8 ± 0.1
Multitask	log-Mel	4	71.5 ± 1.5	91.0 ± 0.4

| Multitask | log-Mel | 5 | 72.8 ± 0.7 | 89.5 ± 0.5 |
| Multitask | log-Mel | 6 | 71.9 ± 1.1 | 85.2 ± 0.8 |

To demonstrate the effectiveness of the proposed multitask learning method, the performances of the multitask models with different super-class numbers using different spectrograms are given in Table 3. The super-class division used in the experiments here is presented in Table 2. The experiments are carried out three times and the results are then reported using the average and standard deviation in percentage. For the STFT models, the best-achieved accuracy is 62.0% obtained by the multitask model with three super-classes. This confirms an improvement of 2.0% in comparison with the basic model. Similarly, an accuracy of 69.5% is achieved by the multitask CQT model with four super-classes which is equivalent to an improvement of 2.0% over the basic model.

The accuracy of the multitask CQT model with three super-classes is 68.4%, which is identical to the original Indoor, Outdoor, and Vehicle division. As we can see, the performance of the model using original manually grouped division is outperformed by the one using super-classes generated by the proposed method. This means that even the datasets with hierarchical labels benefit from the proposed method.

The same situation can be observed in the log-Mel models. The accuracy of the multitask log-Mel model using three super-classes (they are the same as the original division) is 72.1%. However, the best accuracy among the multitask log-Mel models is 72.8%. This indicates an improvement of 3.5% over the basic model. All the multitask models (with two to six super-classes) have outperformed their corresponding basic models in Table 3.

A one-sided paired t-test was applied to obtain the statistical significant difference of the accuracy of 15-scenes between the basic model and the corresponding best-performed multitask model. The results revealed the statistical significance (significance level < 0.05) of the accuracy improvement on the STFT, CQT, and log-Mel models, respectively.

4.4. Evaluations on the LITIS Rouen Dataset

The confusion matrix used is similarly calculated using the sum of the twelve confusion matrices generated from the validation sets (Figure 8).

According to the above confusion matrices, the 19 scene classes in the LITIS Rouen dataset are grouped into two to six super-classes, as shown in Table 4. The divisions indicate the following findings: First, it is found that the outputs of the class hierarchy construction method are stable and robust. For example, the three super-classes' divisions for the STFT model and log-Mel model are the same. Likewise, the five super-classes' divisions and the six super-classes' divisions for the STFT model and CQT model, respectively, are identical. The two super-classes' divisions and four super-classes' devisions for the STFT model and log-Mel model are very similar as well. The difference only lies in the division of a single class. Second, the classes bus, train, metro Rouen, car, high-speed train, and metro Paris are grouped into one super-class in almost all cases. It is the equivalent of the Vehicle category in the TUT Acoustic Scenes 2017 dataset. However, the class plane is separated from the Vehicle super-class and divided as a one-element super-class, which seems more reasonable, as the plane is a kind of non-ground transportation. The classes restaurant, billiard pool hall, and student hall are also clustered as a fixed combination regularly. The features they have in common include their medium-sized indoor space and people's close-talk. These features may produce similar acoustical characteristics.

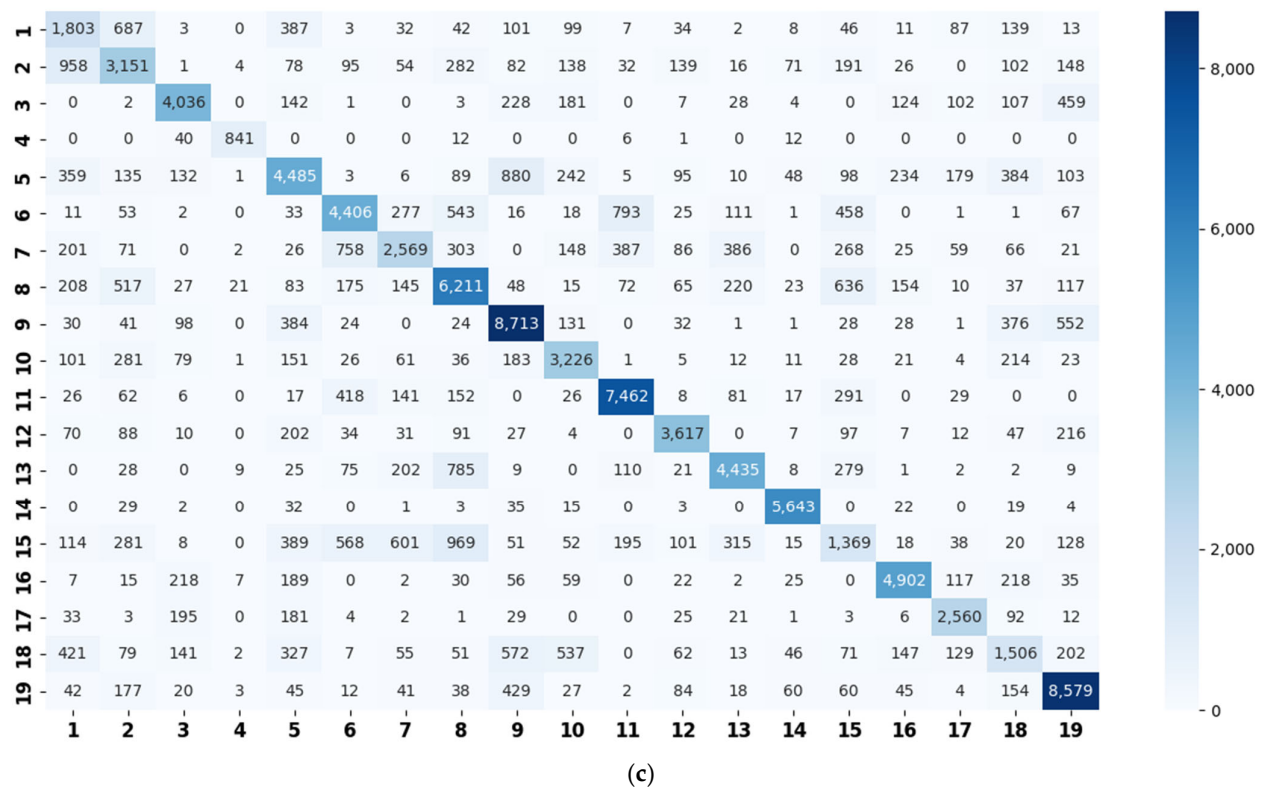|	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	2,215	1,015	18	2	626	30	49	90	70	234	31	83	15	17	56	28	196	302	179
2	1,405	4,393	9	4	121	128	108	574	183	185	105	235	29	153	277	44	4	152	243
3	1	3	5,321	0	481	7	4	59	441	276	0	39	4	5	52	766	295	110	272
4	0	2	6	1,165	0	0	5	24	0	0	52	14	7	6	3	15	0	0	69
5	389	235	401	0	6,078	31	16	93	1,385	285	5	194	25	76	136	248	562	857	216
6	62	166	11	16	42	6,059	514	1,277	123	66	797	42	323	12	585	23	1	13	92
7	624	181	15	43	51	1,010	2,930	766	15	156	565	75	565	130	317	19	318	149	135
8	210	1,102	72	65	226	645	305	7,979	87	84	285	225	273	59	883	179	17	29	451
9	43	156	273	0	704	105	5	75	12,505	173	2	81	6	16	185	53	35	657	622
10	234	338	455	4	180	97	125	73	272	4,145	17	27	57	25	106	65	11	430	35
11	120	101	1	0	10	761	294	336	0	26	11,047	7	164	15	154	0	62	2	4
12	219	255	29	5	255	43	59	180	79	40	0	4,806	15	40	146	125	39	156	349
13	23	25	2	26	62	303	562	1,187	17	1	661	11	5,768	17	211	33	28	30	33
14	20	144	2	0	69	0	7	25	22	20	8	7	3	8,217	46	74	0	11	37
15	134	626	73	29	354	861	550	1,473	210	160	385	241	367	58	1,882	109	43	59	234
16	4	119	177	2	379	39	20	146	431	138	2	48	39	270	86	6,485	100	45	326
17	36	19	424	0	395	3	9	12	23	41	8	26	25	0	0	11	3,578	88	54
18	271	208	139	7	929	21	61	58	693	829	4	111	42	65	71	167	128	2,386	362
19	54	285	107	26	211	51	54	115	617	125	2	176	19	152	130	424	14	405	11,793

(**a**)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	3,075	1,409	12	1	521	27	66	76	260	239	34	148	4	10	109	79	143	291	66
2	2,021	5,206	20	2	215	224	148	601	185	273	247	204	40	96	260	119	0	142	437
3	0	1	7,965	0	187	5	0	2	597	94	3	0	5	2	9	243	462	127	468
4	0	3	33	1,567	0	0	1	20	0	0	7	24	3	14	1	22	0	0	15
5	777	354	385	4	7,829	122	4	227	1,785	478	82	143	17	69	273	189	598	522	182
6	58	393	24	5	182	6,734	542	1,437	59	137	1,811	176	131	22	472	153	16	45	383
7	712	230	1	15	31	844	4,701	460	3	213	515	137	689	153	336	16	51	26	947
8	504	1,322	17	52	243	564	276	10,817	92	49	384	120	400	29	1,256	110	4	15	216
9	209	115	190	0	1,153	141	2	70	15,192	283	29	92	7	30	106	215	77	916	793
10	145	461	137	9	450	187	64	56	397	5,449	51	30	6	13	89	173	5	557	91
11	47	278	7	0	73	979	177	260	11	20	13,858	20	212	26	285	47	46	18	16
12	188	121	13	6	288	98	76	180	126	49	6	6,792	11	11	127	10	30	158	260
13	25	97	0	51	46	328	542	1,186	32	4	441	10	8,196	3	225	11	8	6	39
14	4	24	3	0	25	104	31	4	17	17	53	5	2	10,521	14	9	2	44	11
15	221	484	96	45	556	915	921	1,798	198	121	247	287	528	117	2,758	61	107	81	269
16	63	110	202	0	257	73	10	67	205	492	65	11	23	127	74	8,624	262	301	104
17	28	0	370	0	265	0	3	10	45	11	0	36	6	3	17	34	4,941	137	34
18	422	222	96	9	605	33	27	27	1,378	1,110	14	96	13	90	81	359	200	2,817	591
19	102	334	90	14	65	80	72	103	977	67	8	239	9	36	160	184	37	518	15,355

(**b**)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	1,803	687	3	0	387	3	32	42	101	99	7	34	2	8	46	11	87	139	13
2	958	3,151	1	4	78	95	54	282	82	138	32	139	16	71	191	26	0	102	148
3	0	2	4,036	0	142	1	0	3	228	181	0	7	28	4	0	124	102	107	459
4	0	0	40	841	0	0	0	12	0	0	6	1	0	12	0	0	0	0	0
5	359	135	132	1	4,485	3	6	89	880	242	5	95	10	48	98	234	179	384	103
6	11	53	2	0	33	4,406	277	543	16	18	793	25	111	1	458	0	1	1	67
7	201	71	0	2	26	758	2,569	303	0	148	387	86	386	0	268	25	59	66	21
8	208	517	27	21	83	175	145	6,211	48	15	72	65	220	23	636	154	10	37	117
9	30	41	98	0	384	24	0	24	8,713	131	0	32	1	1	28	28	1	376	552
10	101	281	79	1	151	26	61	36	183	3,226	1	5	12	11	28	21	4	214	23
11	26	62	6	0	17	418	141	152	0	26	7,462	8	81	17	291	0	29	0	0
12	70	88	10	0	202	34	31	91	27	4	0	3,617	0	7	97	7	12	47	216
13	0	28	0	9	25	75	202	785	9	0	110	21	4,435	8	279	1	2	2	9
14	0	29	2	0	32	0	1	3	35	15	0	3	0	5,643	0	22	0	19	4
15	114	281	8	0	389	568	601	969	51	52	195	101	315	15	1,369	18	38	20	128
16	7	15	218	7	189	0	2	30	56	59	0	22	2	25	0	4,902	117	218	35
17	33	3	195	0	181	4	2	1	29	0	0	25	21	1	3	6	2,560	92	12
18	421	79	141	2	327	7	55	51	572	537	0	62	13	46	71	147	129	1,506	202
19	42	177	20	3	45	12	41	38	429	27	2	84	18	60	60	45	4	154	8,579

(**c**)

Figure 8. Confusion matrices on the LITIS Rouen dataset generated by the basic models (**a**) using STFT spectrograms; (**b**) using CQT spectrograms; and (**c**) using log-Mel spectrograms. The abscissa represents the predicted label, and the ordinate indicates the true label. Note: Quiet street (1), busy street (2), restaurant (3), plane (4), shop (5), bus (6), train (7), metro Rouen (8), market (9), café (10), car (11), tubestation (12), high speed train (13), kid game hall (14), metro Paris (15), billiard pool hall (16), student hall (17), pedestrian street (18), and train station hall (19).

Table 4. Clustering details of the constructed super-class for the LITIS Rouen dataset. The classes marked by the same shape with the same color are grouped into the same super-class.

Clustering Scheme	Generated by STFT Model				Generated by CQT Model				Generated by Log-Mel Model			
Two Super-Classes	1	2	3	4	1	2	3	4	1	2	3	4
	5	6	7	8	5	6	7	8	5	6	7	8
	9	10	11	12	9	10	11	12	9	10	11	12
	13	14	15	16	13	14	15	16	13	14	15	16
	17	18	19		17	18	19		17	18	19	
Three Super-Classes	1	2	3	4	1	2	3	4	1	2	3	4
	5	6	7	8	5	6	7	8	5	6	7	8
	9	10	11	12	9	10	11	12	9	10	11	12
	13	14	15	16	13	14	15	16	13	14	15	16
	17	18	19		17	18	19		17	18	19	

Table 5 compares the accuracies of basic models, as well as their corresponding multitask learning models. The experiments are repeated three times and the average and standard deviation accuracies are provided. For the STFT models, the best accuracy is 76.0% which is achieved by the multitask model with four super-classes. It is equivalent to an improvement of 1.6% over the basic model. Additionally, the five multitask models all outperform the basic one. For the CQT models, the multitask model with two super-classes achieves the best accuracy, providing an improvement of 1.9% over the basic model. Again, the five multitask CQT models are all superior to the basic model. For the log-Mel models, the best accuracy of 78.1% is achieved by the multitask model with two super-classes. This is equivalent to an improvement of 1.8% over the basic one. Similarly, the results of the t-test had revealed the statistical significance (significance level < 0.05) of the accuracy improvement by the corresponding best-performed multitask model (over the basic model) on the STFT, CQT, and log-Mel models, respectively. According to the results, the classification performance of LITIS Rouen dataset has been significantly improved by constructing super-classes and integrating them into the multitask learning framework. Consequently, we can see that the proposed self-organized multitask learning method is also helpful for the acoustic scene datasets with single-level labels. According to the extended super-class labels for each sample, the super-class results predicted by the multitask learning models are also evaluated and shown in Table 5. High accuracies have been achieved on the super-class classification tasks. For instance, for the log-Mel models, an accuracy of 97.9% is achieved for the two super-classes classification and it is 94.9% for the six super-classes classification.

Table 5. Classification performance of the revised version of the LITIS Rouen dataset using different models.

Model Type	Feature Type	Super-Class Number	19 Scenes Accuracy	Super-Class Accuracy
Basic Model	STFT	/	74.4 ± 0.7	/
Multitask	STFT	2	75.7 ± 0.6	97.1 ± 0.2
Multitask	STFT	3	75.0 ± 0.6	96.6 ± 0.2
Multitask	STFT	4	76.0 ± 0.4	95.8 ± 0.3
Multitask	STFT	5	75.5 ± 0.9	96.1 ± 0.1
Multitask	STFT	6	74.7 ± 0.5	95.2 ± 0.3
Basic Model	CQT	/	75.6 ± 0.8	/
Multitask	CQT	2	77.5 ± 0.4	97.3 ± 0.1
Multitask	CQT	3	77.2 ± 0.5	97.7 ± 0.4
Multitask	CQT	4	75.8 ± 0.7	96.6 ± 0.0
Multitask	CQT	5	77.0 ± 0.3	96.7 ± 0.2
Multitask	CQT	6	76.5 ± 0.6	95.1 ± 0.3
Basic Model	log-Mel	/	76.3 ± 0.7	/
Multitask	log-Mel	2	78.1 ± 0.3	97.9 ± 0.3
Multitask	log-Mel	3	77.7 ± 0.9	97.9 ± 0.0
Multitask	log-Mel	4	76.2 ± 0.7	95.6 ± 0.3
Multitask	log-Mel	5	76.7 ± 0.3	95.6 ± 0.2
Multitask	log-Mel	6	77.1 ± 1.0	94.9 ± 0.4

4.5. Ensemble Results

Late fusion ensemble is commonly used in the domain of ASC (see, e.g., [31,51]). For instance, in [31], the linear logistic regression was performed on the classification scores of eight models to obtain fusion results by the winner of first place in the DCASE Challenge 2017. Here, we also use a late fusion ensemble. Specifically, the best multitask models are selected and combined with a simple majority voting scheme. For the TUT Acoustic Scenes 2017 dataset, these models include the multitask STFT model with three super-classes, the multitask CQT model with four super-classes, and the multitask log-Mel model with five super-classes. For the revised version of the LITIS Rouen dataset, these models include the multitask STFT model with four super-classes, the multitask CQT model with two super-classes, and the multitask log-Mel model with two super-classes. The ensemble results, as well as the state-of-the-art results for the two datasets, are displayed in Table 6.

Table 6. Comparison of accuracies with state-of-the-art works

Reference	Method	Accuracy	Dataset
[31]	GAN + SVM + FCNN	83.3	TUT
[51]	Background subtraction	80.4	TUT
[52]	Late fusion of CNN and ensemble classifiers	80.0	TUT
[53]	Embedded filters + DCT-based temporal module	79.2	TUT
[17]	Multi-spectrogram fusion	77.7	TUT
[18]	Mixup + multi-channel	76.7	TUT
[54]	Sound texture enhancement	75.7	TUT
[55]	Multi-spectrogram	72.6	TUT

	encoder-decoder		
Ensemble of three basic models	Ensemble	77.8	TUT
Our method	Super-class construction + multitask learning	81.4	TUT
[48]	Supervised nonnegative matrix factorization	81.8	Rouen-15
Ensemble of three basic models	Ensemble	78.1	Rouen-revised
Our method	Super-class construction + multitask learning	83.9	Rouen-revised

Note: TUT is referred to the TUT Acoustic Scene 2017 dataset (evaluation set); Rouen-15 for the LITIS Rouen-15 dataset; and Rouen-revised for the LITIS Rouen dataset (revised version).

As shown in Table 6, the ensemble result using the three best multitask models for the TUT Acoustic Scenes 2017 is 81.4%. This is higher than the accuracies of most of the state-of-the-art techniques listed in Table 6 except for the model in [31] which used GAN data augmentation. The ensemble result using the three best multitask models for our revised version of LITIS Rouen is 83.9%. For a rough comparison, the accuracy of Rouen-15 (81.8%) [48] is listed here. This is as the evaluated datasets are different. To comprehensively evaluate the multitask learning method, the three basic models are also late fused using the same ensemble method. The ensemble result on the TUT Acoustic Scenes 2017 dataset is 77.8% and the one on the LITIS Rouen dataset is 78.1%. Hence, our proposed ensemble results outperform the corresponding basic ensemble results by 3.6% and 5.8%, respectively, on the two considered datasets.

5. Discussion

5.1. Similarity Relationship of Acoustic Scenes

The experiment results confirm our assumption that the similarity relation amongst the acoustic scenes can be reflected by the classification errors. For example, according to the confusion matrices (Figure 7), in most cases, the classes of beach, city center, and park are, respectively, misclassified as residential areas (for example, there are 1167 beach samples misclassified as residential areas in Figure 7a). Additionally, the residential areas are misclassified as parks; grocery stores are misclassified as cafés; and homes are misclassified as libraries and vice versa. It is also seen that a train is misclassified as a tram, a car is misclassified as a bus, and so on. Similarly, for the LITIS Rouen dataset (Figure 8), in most cases, the class of shop is misclassified as market; metro Rouen is misclassified as metro Paris, and the train station hall is misclassified as a market. The above scenes are similar and understandable from the human point of view, hence seem convincing to us that the similarity relationship among scenes can be learnt from the confusion matrix.

5.2. Advantages of the Super-Class Construction Method

Deviating from the classical acoustic scene/event taxonomy methods [7,8], the proposed super-class construction method only depends on the classification results by a basic classifier. It does not need any feature embedding. The method is simple and effective. As shown in Tables 2 and 3, identical class hierarchies can be achieved by using classifiers on different spectrograms. Furthermore, the method does not limit the type of basic classifier. It can be SVM, random forest, and other models, although the CNN model is applied in this paper. In this sense, the proposed method has good robustness and general applicability. On the other hand, the super-class is constructed based on the similarity relations among scenes. It is more explainable compared to the embedding-based results. Finally, the proposed method is also capable to construct a multi-level class hierarchy.

Although the construction method is proposed for ASC, it can be easily extended into acoustic event clustering and other audio taxonomy.

5.3. Foundation of Self-Organized Multitask Learning

As shown in the experiments, the proposed self-organized multitask learning method can improve the ASC performance compared to the corresponding basic models. The achieved improvement comes in three ways. First, the constructed super-class labels provide more information in supervised learning. Second, teaching the model to classify the fine-grained category along with the coarse category accords with the cognitive law of "learning the easy things first" [56]. Third, according to [57], the performance of the harder task can be improved by using the information obtained from easier tasks, where predicting the super-class is an easier task. As shown in Table 5, high super-class accuracies are achieved by our models. For example, the two to six super-classes accuracies of multitask learning models using log-Mel spectrograms in the LITIS Rouen dataset are 97.9%, 97.9%, 95.6%, 95.6%, and 94.9%, respectively. Consequently, to keep the auxiliary task easy, the number of super-class should not be too large. This also justifies why competitive results are not obtained by the models with larger super-class numbers (see Figures 5 and 6).

5.4. Regularization by Similarity Relation

The relevance between super-class and original class is expressed as a regularization item in the multitask learning loss function (see Equation (5)). According to our experiments, this regularization slightly improves the performance. The evaluation experiments are only performed on the multitask learning models with three super-classes using STFT, CQT, and log-Mel spectrograms in both datasets. The improvements in the TUT Acoustic Scene 2017 dataset are 0.3%, 0.6%, and 0.8% and those in the LITIS Rouen dataset are 0.4%, 0.4%, and 0.6%, for STFT, CQT, and log-Mel models, respectively.

6. Conclusions

In this paper, the similarity relation among acoustic scenes is utilized to construct a two-level class hierarchy. The class hierarchy is further incorporated into a self-organized multitask learning framework. The experimental results show that the proposed multitask learning method can improve the classification performance effectively using different spectrograms.

The best improvements of the STFT, CQT, and log-Mel multitask models over their corresponding basic models are 2.0%, 2.0%, and 3.5%, respectively, on the TUT Acoustic Scenes 2017 dataset. Corresponding to two to six super-classes, respectively, the coarse category classification accuracies range from 93.1 to 77.0% for STFT models, from 96.2% to 84.0% for CQT models, and from 94.9% to 85.2% for log-Mel models. By applying the late fusion strategy, a fine-grained category accuracy of 81.4% is achieved on the dataset. On the LITIS Rouen dataset, the best improvements over the corresponding basic models are 1.6%, 1.9%, and 1.8% for the STFT, CQT, and log-Mel multitask models, respectively. The super-class accuracies range from 97.1% to 95.2%, from 97.7% to 95.1%, and from 97.9% to 94.9% corresponding to two to six super-classes, for STFT, CQT, and log-Mel models, respectively. An ensemble fine-grained category of 83.9% is achieved.

According to the experiments, the following conclusions can be drawn:

(1) The similarity relation based class hierarchy construction method is effective and reasonable.

(2) The constructed class hierarchy can be utilized to improve the ASC performance effectively in multitask learning.

(3) For a hierarchically arranged dataset, there may exist a hierarchy that is automatically constructed by our method. This may perform better than the original hierarchy in ASC.

(4) In self-organized multitask learning, the number of super-class should be chosen carefully. The multitask models with large super-class numbers would not obtain competitive results.

(5) The relevance between coarse and fine-grained classes can be utilized as regularization to improve the ASC performance.

(6) By arranging the class hierarchy, the self-organized multitask learning method provides a feasible way to promote the performance of a certain model.

In future work, we will extend this confusion matrix based super-class construction method into the domain of acoustic event taxonomy. Furthermore, to improve the performance of scene identification, the multimodal fusion method will be explored. Specifically, image and sensor data, etc., can be employed to enhance the audio data in the scene identification task.

Author Contributions: Conceptualization, W.Z. and Z.M.; methodology, W.Z.; software, Z.M.; validation, Z.M. and W.Z.; formal analysis, W.Z.; investigation, W.Z.; resources, G.Z.; data curation, Z.M.; writing—original draft preparation, W.Z.; writing—review and editing, Z.M.; visualization, Z.M.; supervision, G.Z.; project administration, G.Z.; funding acquisition, W.Z. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Barchiesi, D.; Giannoulis, D.; Stowell, D.; Plumbley, M.D. Acoustic Scene Classification: Classifying environments from the sounds they produce. *IEEE Signal Process. Mag.* **2015**, *32*, 16–34.
2. Stowell, D.; Giannoulis, D.; Benetos, E.; Lagrange, M.; Plumbley, M.D. Detection and classification of acoustic scenes and events. *IEEE Trans. Multimed.* **2015**, *17*, 1733–1746.
3. Hossain, M.S.; Muhammad, G. Environment classification for urban big data using deep learning. *IEEE Commun. Mag.* **2018**, *56*, 44–50.
4. Imoto, K. Incorporating Intra-Class Variance to Fine-Grained Visual Recognition. *Acoust. Sci. Technol*. **2018**, *39*, 182–188.
5. Em, Y.; Gag, F.; Lou, Y.; Wang, S.; Huang, T.; Duan, L.-Y. Incorporating Intra-Class Variance to Fine-Grained Visual Recognition. In Proceedings of the 2017 IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 1452–1457.
6. Mesaros, A.; Heittola, T.; Diment, A.; Elizalde, B.; Shah, A.; Vincent, E.; Raj, B.; Virtanen, T. DCASE 2017 challenge setup: Tasks, datasets and baseline system. In Proceedings of the DCASE 2017-Workshop on Detection and Classification of Acoustic Scenes and Events, Munich, Germany, 16–17 November 2017; pp. 85–92.
7. Ye, J.; Kobayashi, T.; Wang, X.; Tsuda, H.; Murakawa, M. Audio Data Mining for Anthropogenic Disaster Identification: An Automatic Taxonomy Approach. *IEEE Trans. Emerg. Top. Comput*. **2017**, *8*, 126–136.
8. Li, Y.; Liu, M.; Wang, W.; Zhang, Y.; He, Q. Acoustic Scene Clustering Using Joint Optimization of Deep Embedding Learning and Clustering Iteration. *IEEE Trans. Multimed*. **2019**, *22*, 1385–1394.
9. Tonami, N.; Imoto, K.; Niitsuma, M.; Yamanishi, R.; Yamashita, Y. Joint analysis of acoustic events and scenes based on multitask learning. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; pp. 338–342.
10. Abrol, V.; Sharma, P. Learning Hierarchy Aware Embedding from Raw Audio for Acoustic Scene Classification. *IEEEACM Trans*. *Audio Speech Lang. Process*. **2020**, *28*, 1964–1973.
11. Rakotomamonjy, A.; Gasso, G. Histogram of gradients of time-frequency representations for audio scene classification. *IEEE/ACM Trans. Audio Speech Lang. Process*. **2014**, *23*, 142–153.
12. Geiger, J.T.; Schuller, B.; Rigoll, G. Large-scale audio feature extraction and SVM for acoustic scene classification. In Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20–23 October 2013; pp. 1–4.
13. Ma, L.; Milner, B.; Smith, D. Acoustic environment classification. *ACM Trans. Speech Lang. Process. TSLP* **2006**, *3*, 1–22.
14. Chakrabarty, D.; Elhilali, M. Exploring the role of temporal dynamics in acoustic scene classification. In Proceedings of the 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 18–21 October 2015; pp. 1–5.
15. Yu, D.; Li, J. Recent progresses in deep learning based acoustic models. *IEEE/CAA J. Autom. Sin*. **2017**, *4*, 396–409.
16. Eghbal-Zadeh, H.; Lehner, B.; Dorfer, M.; Widmer, G. CP-JKU submission for DCASE-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks. In Proceedings of the IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), Budapest, Hungary, 3 September 2016; pp. 5024–5028.

17. Weiping, Z.; Jiantao, Y.; Xiaotao, X.; Xiangtao, L.; Shaohu, P. Acoustic scene classification using deep convolutional neural network and multiple spectrograms fusion. In Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE), Munich, Germany, 16–17 November 2017; pp. 133-137.
18. Xu, K.; Feng, D.; Mi, H.; Zhu, B.; Wang, D.; Zhang, L.; Cai, H.; Liu, S. Mixup-based acoustic scene classification using multi-channel convolutional neural network. In Proceedings of the Pacific Rim Conference on Multimedia, Hefei, China, 21–22 September 2018; pp. 14–23.
19. Hershey, S.; Chaudhuri, S.; Ellis, D.P.; Gemmeke, J.F.; Jansen, A.; Moore, R.C.; Plakal, M.; Platt, D.; Saurous, R.A.; Seybold, B. CNN architectures for large-scale audio classification. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 131–135.
20. Phan, H.; Koch, P.; Katzberg, F.; Maass, M.; Mazur, R.; Mertins, A. Audio scene classification with deep recurrent neural networks. In *Proceedings of the INTERSPEECH*; INTERSPEECH: Stockholm, Sweden, 2017.
21. Bae, S.H.; Choi, I.; Kim, N.S. Acoustic scene classification using parallel combination of LSTM and CNN. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), Budapest, Hungary, 3 September 2016; pp. 11–15.
22. Xu, Y.; Huang, Q.; Wang, W.; Plumbley, M.D. Hierarchical learning for DNN-based acoustic scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016), Budapest, Hungary, 3 September 2016; pp. 110–114.
23. Guo, J.; Xu, N.; Li, L.-J.; Alwan, A. Attention based CLDNNs for short-duration acoustic scene classification. In Proceedings of the INTERSPEECH; INTERSPEECH, Stockholm, Sweden, 20–24 August 2017; pp. 469–473.
24. Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process*. **2010**, *19*, 788–798.
25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput*. **1997**, *9*, 1735–1780.
26. Zhang, X.; Zhou, F.; Lin, Y.; Zhang, S. Embedding label structures for fine-grained feature representation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1114–1123.
27. Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In Proceedings of the Advances in Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; pp. 1857–1865.
28. Xie, S.; Yang, T.; Wang, X.; Lin, Y. Hyper-class augmented and regularized deep learning for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2645–2654.
29. Wu, H.; Merler, M.; Uceda-Sosa, R.; Smith, J.R. Learning to make better mistakes: Semantics-aware visual food recognition. In Proceedings of the 24th ACM International Conference on Multimedia; Association for Computing Machinery, New York, NY, USA, 1 October 2016; pp. 172–176.
30. Imoto, K.; Tonami, N.; Koizumi, Y.; Yasuda, M.; Yamanishi, R.; Yamashita, Y. Sound event detection by multitask learning of sound events and scenes with soft scene labels. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 621–625.
31. Mun, S.; Park, S.; Han, D.K.; Ko, H. Generative adversarial network based acoustic scene training set augmentation and selection using SVM hyper-plane. In Proceedings of the DCASE 2017–Detection and Classification of Acoustic Scenes and Events Workshop, Munich, Germany, 16–17 November 2017; pp. 93–97.
32. Salamon, J.; Bello, J.P. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283.
33. Lu, R.; Duan, Z.; Zhang, C. Metric learning based data augmentation for environmental sound classification. In Proceedings of the 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 15–18 October 2017; pp. 1–5.
34. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Proc. Adv. Neural Inf. Process. Syst.* **2014**, *3*, 2672–2680.
35. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. *AAAI* **2020**, *34*, 13001–13008.
36. Gharib, S.; Derrar, H.; Niizumi, D.; Senttula, T.; Tommola, J.; Heittola, T.; Virtanen, T.; Huttunen, H. Acoustic scene classification: A competition review. In Proceedings of the 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), Aalborg, Denmark, 17–20 September 2018; pp. 1–6.
37. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond empirical risk minimization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
38. Nawa, S.; Quatieri, T.; Lim, J. Signal Reconstruction from Short-Time Fourier Transform Magnitude. *IEEE Trans. Acoust. Speech Signal Process.* **1983**, *31*, 986–998.
39. Brown, C.J.; Puckette, S.M. An efficient algorithm for the calculation of a constant Q transform. *J. Acoust. Soc. Am.* **1992**, *92*, 2698–2701.
40. Logan, B. Mel frequency cepstral coefficients for music modeling. In Proceedings of the In International Symposium on Music Information Retrieval, Montréal, QC, Canada, 23–25 October 2000; pp. 1–11.
41. Boashash, B.; Khan, N.A.; Ben-Jabeur, T. Time-frequency features for pattern recognition using high-resolution TFDs: A tutorial review. *Digit. Signal Process.* **2015**, *40*, 1–30.

42. Paseddula, C.; Gangashetty, S.V. Late fusion framework for Acoustic Scene Classification using LPCC, SCMC, and log-Mel band energies with Deep Neural Networks. *Appl. Acoust.* **2021**, *172*, 107568.
43. Mcdonnellk, M.D.; Gao, W. Acoustic Scene Classification Using Deep Residual Networks with Late Fusion of Separated High and Low Frequency Paths. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Conference, 4–8 May 2020; pp. 141–145.
44. Ng, A.; Jordan, M.; Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2002; pp. 849–856.
45. Gopal, S.; Yang, Y. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In Proceedings of the 19th ACM SIGKDD–International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA 11–14 August 2013; pp. 257–265.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
47. Szegedy, C.; Liu, W.; Jia, Y.; et al Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
48. Rakotomamonjy, A. Supervised representation learning for audio scene classification. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1253–1265.
49. Abadi, M.; Agarwal, A.; Barham, P.; et al. TensorFlow: Large-scale machine learning on heterogeneous systems. In Proceedings of the Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
50. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
51. Han, Y.; Park, J.; Lee, K. Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification. In Proceedings of the DCASE 2017–Detection and Classification of Acoustic Scenes and Events Workshop, Munich, Germany, 16–17 November 2017; pp. 46–50.
52. Alamir, M.A. A novel acoustic scene classification model using the late fusion of convolutional neural networks and different ensemble classifiers. *Appl. Acoust.* **2021**, *175*, 107829.
53. Chen, H.; Zhang, P.; Yan, Y. An audio scene classification framework with embedded filters and a DCT-based temporal module. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 835–839.
54. Wu, Y.; Lee, T. Enhancing sound texture in CNN-based acoustic scene classification. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 815–819.
55. Pham, L.; Phan, H.; Nguyen, T.; Palaniappan, R.; Mertins, A.; McLoughlin, I. Robust acoustic scene classification using a multi-spectrogram encoder-decoder framework. *Digit. Signal Process*. **2021**, *110*, 102943.
56. Lee, Y.J.; Grauman, K. Learning the easy things first: Self-paced visual category discovery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 1721–1728.
57. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75.