Learning and Fusing Multimodal Deep Features for Acoustic Scene Categorization

Yifang Yin
National University of Singapore
Singapore, Singapore
idsyin@nus.edu.sg

Rajiv Ratn Shah
IIIT-Delhi
Delhi, India
rajivratn@iiitd.ac.in

Roger Zimmermann
National University of Singapore
Singapore, Singapore
rogerz@comp.nus.edu.sg

ABSTRACT

Convolutional Neural Networks (CNNs) have been widely applied to audio classification recently where promising results have been obtained. Previous CNN-based systems mostly learn from two-dimensional time-frequency representations such as MFCC and spectrogram, which may tend to emphasize more on the background noise of the scene. To learn the key acoustic events, we introduce a three-dimensional CNN to emphasize on the different spectral characteristics from neighboring regions in spatial-temporal domain. A novel acoustic scene classification system based on multimodal deep feature fusion has been proposed in this paper, where three CNNs have been presented to perform 1D raw waveform modeling, 2D time-frequency image modeling, and 3D spatial-temporal dynamics modeling, respectively. The learnt features have been shown to be highly complementary to each other, which are next combined in a feature fusion network to obtain significantly improved classification predictions. Comprehensive experiments have been conducted on two large-scale acoustic scene datasets, namely the DCASE16 dataset and the LITIS Rouen dataset. Experimental results demonstrate the effectiveness of our proposed approach, as our solution achieves the state-of-the-art classification rates and improves the average classification accuracy by 1.5% ∼ 8.2% compared to the top ranked systems in the DCASE16 challenge.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Information systems** → *Multimedia databases*;

KEYWORDS

Acoustic scene classification; spatial-temporal dynamics modeling; multimodal deep features; CNN fusion

ACM Reference Format:
Yifang Yin, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Learning and Fusing Multimodal Deep Features for Acoustic Scene Categorization. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240631

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240631

1 INTRODUCTION

Acoustic scene classification aims at automatically recognizing the environments based on an audio recording of the scene. It has been an important yet challenging problem in audio processing, which enables a wide range of subsequent applications including surveillance, robotic navigation, and context-aware services [31]. Some examples of such applications can be analyzing human activity for surveillance, or tracking traffic in urban area [32]. A recognized scene can also be used as priors to improve the performance of sound event detection [13].

An acoustic scene usually involves various foreground sounds and background noise, which makes it highly challenging to extract a descriptive representation for classification. To address this problem, a great number of signal processing and machine learning techniques have been investigated such as Gaussian mixture models [1], matrix factorization [6], and most recent deep neural networks (DNNs) [12]. In particular, DNN-based methodologies recognize acoustic scenes through computer vision techniques. Audios are processed as a set of images where 2D time-frequency representations are extracted for classification. When applied to spectrogram-like inputs, convolutional neural networks (CNNs) can effectively capture energy modulation patterns across time and frequency, and thus obtain promising results in various audio applications [34]. However, such two-dimensional CNNs tend to emphasize more on the background noise rather than on the acoustic event occurrences [44], the performance of which can be limited due to the requirement on the availability of large quantities of training data. Though data augmentation techniques [34] can be applied, it is difficult to achieve the state-of-the-art classification accuracy based on CNNs alone. Various fusion approaches have been proposed to boost the system performance by combining features or scores learnt by different classifiers [8, 26].

We therefore present a novel acoustic scene classification system based on multi-CNN fusion. The system overview is illustrated in Figure 1. We introduce three CNNs of different dimensions to learn from multimodal features extracted from audios: 1) a 1D CNN for raw waveform modeling, 2) a 2D CNN for time-frequency image modeling, and 3) a 3D CNN for spatial-temporal dynamics modeling. Previous studies mostly use recurrent neural networks (RNNs) for temporal modeling of audios [10, 22]. However, one shortcoming of such methods is the tendency to overemphasize the temporal information. To solve this issue, we innovatively introduce a 3D CNN, which is capable of simultaneously learning features from both spatial and temporal dimensions through capturing the correlations between three-dimensional signals [21]. Inspired by 3D motion of videos, we generate a 3D signal capturing

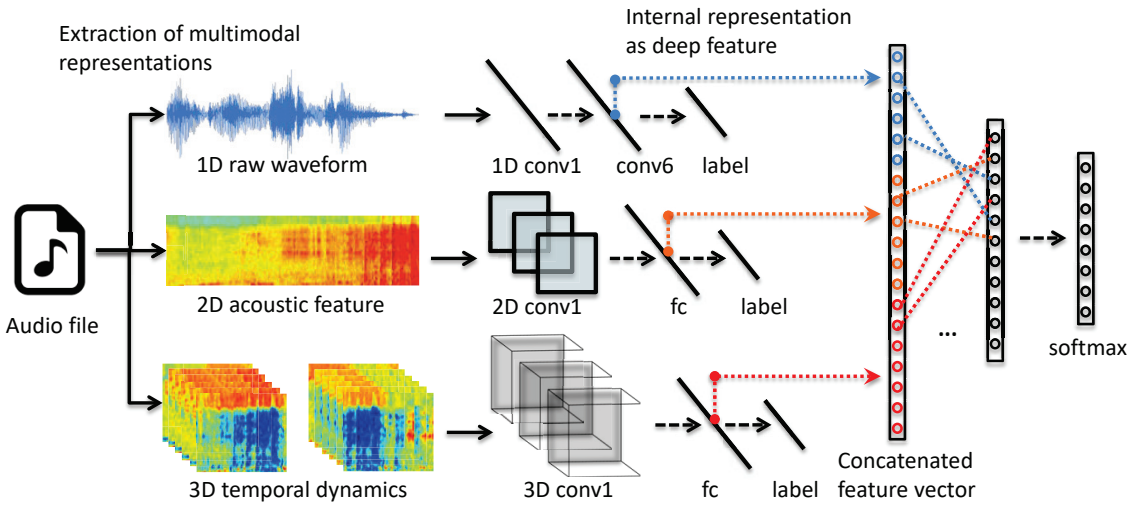

Figure 1: Overview of our proposed acoustic scene classification system based on multimodal deep fusion.

the change in 2D time-frequency representations of consecutive audio frames. The extraction of the 3D signal emphasizes different spectral characteristics from neighboring regions in spatial-temporal domain, which makes it easier to detect key acoustic events that help with scene classification.

The 1D and 3D CNNs require longer training time, but the learnt features capture quite different acoustic characteristics compared to the 2D time-frequency based CNN. This enables significant improvements in classification by applying feature fusion techniques. Moreover, we apply an effective segment-level score aggregation method and neural network ensemble to further boost the system's performance. Thus far, to the best of our knowledge, no studies have leveraged 3D CNNs to model the spatial-temporal dynamics of audios. Furthermore, our proposed method obtains the best classification results on both the DCASE16 dataset and the LITIS Rouen dataset. It is also worth mentioning that our method is able to outperform the top ranked systems in the DCASE16 acoustic scene classification challenge without applying any data augmentation or transfer learning techniques.

The key contributions of this work are summarized as follows:

- We are the first to utilize a 3D convolutional neural network for the spatial-temporal dynamics modeling of an audio. Most of the earlier CNN-based methods only learn from 2D time-frequency representations such as MFCC and spectrogram for classification.
- We present a robust acoustic scene classification system based on multimodal feature fusion. The internal representations of the three proposed CNNs are leveraged as deep features, which are next combined in a fusion network to obtain more robust predictions.
- Extensive experiments have been conducted on both of the DCASE16 and the LITIS Rouen datasets for acoustic scene classification. Our proposed method obtains the state-of-the-art classification rates without applying any data augmentation or transfer learning techniques. The average classification accuracy has been improved by 1.5% ~ 8.2% compared to the top ranked systems in the DCASE16 challenge .

The rest of the paper is organized as follows. The important related work is reported in Section 2. Section 3 introduces the proposed CNN architectures for multimodal deep feature learning. Section 4 presents our score aggregation, CNN fusion, and network ensemble techniques for performance improvements. Experimental results on model verification and comparison with the state-of-the-art methods are reported in Section 5. Finally, Section 6 concludes and suggests future work.

2 RELATED WORK

Acoustic scene classification has drawn great research attention in recent years. One traditional way to model the auditory perception of natural and human environments is the bag-of-frames approach that adopts a Gaussian mixture model with acoustic features such as Mel-frequency cepstral coefficients (MFCCs) [1]. This approach has been proven to be effective and till today is still considered as a reasonable baseline system for the Detection and Classification of Acoustic Scenes and Events (DCASE) challenges [23, 35]. Various acoustic features, including MFCCs, log-mel spectrograms, histogram of gradients *etc.* [6, 12, 32, 33, 48], have been investigated and fused for performance improvement. Classifiers based on Gaussian mixture models (GMMs), hidden Markov models (HMMs), non-negative matrix factorization (NMF), and support vector machines (SVMs) [6, 31, 47] are widely adopted for acoustic scene classification historically. For instance, Bisot *et al.* investigated various matrix factorization methods to generate better features from time-frequency representations for acoustic scene classification [6]. Rakotomamonjy and Gasso proposed to extract features based on histogram of gradients from time-frequency representations, which are next fed to a multi-class linear SVM for classification [33].

Table 1: Architecture configuration of the proposed 1D CNN for raw waveform modeling.

Layer	conv1	pool1	conv2	pool2	conv3	conv4	conv5	pool5	conv6	conv7	conv8	output
No. of Filters	64	4	64	4	128	128	256	4	256	512	15	15
Filter Size	16	16	16	16	16	16	8	8	8	8	8	–
Stride	2	4	2	4	2	2	2	4	2	2	2	–

Table 2: Architecture configuration of the proposed 2D CNN for log-mel spectrogram modeling.

Layer	conv1	pool1	conv2	pool2	conv3	conv4	fc	output
No. of Filters	64	64	64	64	128	128	256	15
Filter Size	(5×5)	(2×2)	(5×5)	(2×2)	(5×5)	(5×5)	–	–
Stride	(2×2)	(2×2)	(2×2)	(2×2)	(2×2)	(2×2)	–	–

Inspired by the great success of Deep Neural Networks (DNNs) on image classification [36, 49], Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been gaining increased attention in audio processing as state-of-the-art performances have been achieved in the field of, *e.g.*, acoustic event detection and acoustic scene classification [2, 12, 14, 34]. For instance, Han *et al.* applied CNNs consisting of eight convolution layers to both mono and stereo sounds to effectively learn different acoustic characteristics from audio recordings [12]. Aytar *et al.* proposed a novel deep convolutional architecture, which consists of a series of one-dimensional convolutions followed by nonlinearities, for learning sound representations from raw audio waveforms [2]. Guo *et al.* presented a noval attention-based DNN framework to take advantages of both frequency modeling with CNNs and temporal modeling with RNNs [10]. Eghbal-Zadeh *et al.* proposed a CNN architecture trained on spectrograms of audio excerpts, which was fused with scores predicted based on their proposed i-vector representations [8]. Recently, Hershey *et al.* [14] presented a large-scale audio dataset and investigated the classification performance of various CNN architectures including AlexNet [16], Inception [40] *etc.* Their experiments show that state-of-the-art image networks are capable of producing excellent results on audio classification as well.

As pointed out by Hershey *et al.* [14], the use of large training and label sets can help improve the classification performance. Data augmentation [42] and transfer learning [18] are two of the widely adopted techniques to increase the number of training samples for better performance. Audio deformations such as time stretching, pitch shifting, and background noise mixing are commonly applied for data augmentation [34]. Transfer learning focuses on transferring knowledge from other data sources, which has been successful in computer vision due to the availability of large dataset such as the ImageNet. Recently, a very large audio dataset named AudioSet has been released in public [14], which motivates the application of transfer learning in audio analysis. The pre-trained models on such dataset can be adapted for new tasks by introducing adaptation layers while keeping the parameters of the pre-trained models unchanged [18]. Additionally, multimodal analysis also helps obtain effective classification results by applying early or late fusion strategies to combine multiple features [19, 37, 38]. More advanced fusion techniques, *e.g.*, bilinear CNN models [20] and multiplicative fusion methods [25], have also been proposed recently where significant improvements were reported. However, existing DNN-based audio classifiers mostly use two-dimensional time-frequency representations as the network input. The three-dimensional spatial-temporal dynamics modeling of acoustic features has not been studied yet when compared to video analysis [46].

3 DEEP FEATURE LEARNING

We present three convolutional neural networks to learn deep features from different sound representations. The features learnt are complementary to each other, and therefore improved acoustic scene classification results can be obtained by applying feature fusion techniques.

3.1 1D Raw Waveform Modeling

Recently, deep convolutional networks that learn directly from raw audio waveforms have been proposed for acoustic scene classification and automatic music generation [2, 7, 45]. We follow this path and find that raw waveform based CNNs are capable of learning quite different acoustic characteristics in supplementary to traditional time-frequency based CNNs. Here we present a deep convolutional network that consists of eight one-dimensional convolutional layers followed by nonlinear transformations, three max-pooling layers, and one softmax output layer. The architecture configuration is illustrated in Table 1.

The input to the network is raw audio waveforms sampled at 22 kHz. We also scale the waveforms to be in the range [-256, 256], so that we do not need to subtract the mean as the data are naturally near zero already. To obtain better classification accuracy, batch normalization (BN) and rectified linear unit (ReLU) are employed after each convolutional layer. Additionally, dropout regularization is applied to convolutional layers conv6, conv7, and conv8. Instead of the fully connected layers that are commonly used in 2D CNNs for image classification, we alternatively adopt a single global max-pooling layer [24, 50] at the output followed by a softmax activation function to generate the prediction scores. The advantages of this modification are two folds: (1) for weakly labeled audios that are only associated with class labels, the global max-pooling explicitly searches for the best candidate position of representative acoustic features for each class in the audio; and

Table 3: Architecture configuration of the proposed 3D CNN for spatial-temporal dynamics modeling.

Layer	conv1	pool1	conv2	pool2	conv3	conv4	fc	output
No. of Filters	64	64	64	64	128	128	256	15
Filter Size	(5×5×5)	(2×2×2)	(5×5×5)	(2×2×2)	(5×5×5)	(5×5×5)	–	–
Stride	(1×2×2)	(2×2×2)	(1×2×2)	(2×2×2)	(1×2×2)	(1×2×2)	–	–

(2) the global max-pooling downsamples variable length inputs to a fixed dimensional vector, which makes the network capable of handling audios that vary in temporal length [2]. Finally, the mean squared error is adopted as the cost function for network optimization.

3.2 2D Time-frequency Image Modeling

Instead of using raw waveforms directly as the input, it is more typical to learn from manually-crafted time-frequency representations (*e.g.*, MFCC, spectrogram, *etc.*) in CNN-based classifiers: audios are divided into fixed length segments, followed by feature extraction, segment-wise classification, and score aggregation [12, 14, 34]. Subsequently, an audio file after feature extraction is represented by a set of images (time-frequency representations) where two-dimensional CNNs for image classification can be directly applied.

The 2D acoustic feature we choose for our system is the log-mel spectrogram. For classification we present a CNN architecture in Table 2, which consists of four two-dimensional convolutional layers, two max-pooling layers, one fully-connected layer, and one softmax output layer. We apply BN to convolutional layers, dropout regularization to the fully-connected layer, and ReLU to all layers for better classification accuracy. The cross-entropy loss is adopted for network optimization.

In terms of log-mel spectrogram extraction, we used a full 44.1 kHz without downsampling and extracted the spectrograms with 64 bin mel-scale. The window size for short-time Fourier transform was 1024 samples with a hop size of 512 samples. The resulting mel-spectrogram was next converted into logarithmic scale, and standardised by subtracting the mean value and dividing by the standard deviation. Standardisation are obtained only from training data to scale both of training and testing data.

3.3 3D Spatial-temporal Dynamics Modeling

As introduced in Section 3.2, audios are traditionally modeled as a set of time-frequency images before applying CNNs. But if we consider the order of the images, it is also possible to model an audio file as a "video" (a sequence of 2D time-frequency representations) where "motion features" can be extracted [39]. Let A denotes an audio file, and $\tilde{I} = \{I_1, I_2, ..., I_n\}$ where I_k represents the 2D acoustic feature extracted from the k-th segment of A. Our first solution to capture the changing dynamics of A in spatial-temporal domain is to compute the difference between the acoustic features of consecutive segments: $\tilde{J} = \{J_1, J_2, ..., J_{n-1}\}$ where $J_k = I_{k+1} - I_k$. Thereafter, the CNN applied to $\tilde{I}$ (see Table 2) can be directly applied to $\tilde{J}$ without architecture change to learn more descriptive deep features from $\tilde{J}$ for classification. However, as an audio file can be segmented with arbitrary overlaps, it is difficult to find the optimal hop size for audio segmentation when computing $\tilde{J}$. To be more general, we define $J_{k,l} = I_k - I_l$, and propose to generate a 3D signal to capture the spatial-temporal dynamics of audios by adding a third dimension to incorporate the variability of the segmentation hop size. The generation of the proposed 3D signal is illustrated in Figure 2. Let $\tilde{D} = \{D_1, D_2, ..., D_m\}$, and D_k represents the proposed 3D signal extracted from the k-th segment of A. Let P, Q, and R represent the number of elements in each dimension of D_k, respectively, then we have,

$$\begin{aligned} D_k^{p,q,r} &= J_{k+r,k}^{p,q} \\ &= I_{k+r}^{p,q} - I_k^{p,q} \end{aligned} \tag{1}$$

where $D_k^{p,q,r}$ represents the element at position (p, q, r) in D_k, $I_k^{p,q}$ and $J_k^{p,q}$ represent the element at position (p, q) in I_k and J_k, respectively. The number of features in $\tilde{D}$, represented by m, equals to $n - R$. For the 2D acoustic features $\tilde{I}$, we use the same log-mel spectrogram extracted as introduced in Section 3.2.

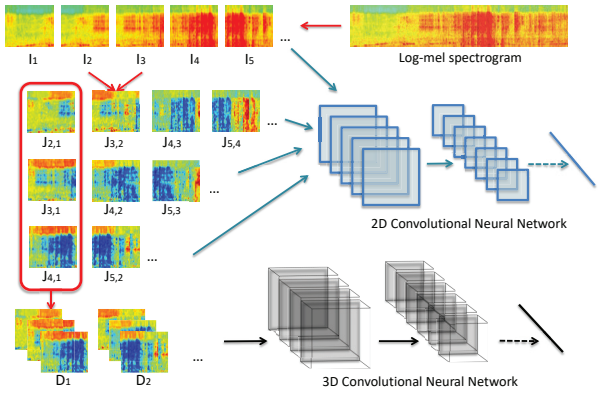

Figure 2: Illustration of the proposed 3D spatial-temporal dynamics modeling.

Next, we adopt a 3D convolutional neural network to model the spatial-temporal dynamics of signal $\tilde{D}$. As shown in Table 3, it consists of four three-dimensional convolutional layers, two max-pooling layers, one fully-connected layer, and one softmax output layer. Similar to the 2D CNN introduced in Section 3.2, BN, ReLU, and dropout regularization are also employed to obtain better classification accuracy. The cross-entropy loss is utilized for network optimization. 3D CNNs have been first proposed for video applications such as action recognition and scene categorization, and have been reported to be more suitable for spatial-temporal feature learning compared to 2D CNNs [43]. To the best of our knowledge, this is among the first attempt to apply 3D CNNs to the spatial-temporal dynamics modeling in acoustic scene classification.

4 MULTIMODAL DEEP FUSION

The trained CNNs can be directly applied to categorize audio files. Alternatively, we can ignore the output layer of the networks and use the internal representation as deep features to train a better classifier. Next, we will discuss how to improve the classification accuracy based on CNN fusion.

4.1 Segment Score Aggregation

As CNNs generally work well with short audio chunks, we split audio files into fixed length segments for training and testing. Segment-wise classifications are first performed, followed by local classification scores aggregated into a global prediction per audio. More specifically, we splitted the original audio files into 1-s chunks without overlapping (*i.e.*, hop size set to 1 s) for the 1D and 2D CNNs, while setting the hop size for the 3D CNN to be 0.25 s. A small value of the hop size allows the 3D signal $\tilde{D}$ capture the audio's spatial-temporal dynamics more effectively. To make the number of audio segments consistent for fusion, we set the size of the third dimension of $\tilde{D}$ to 4 (*i.e.*, R=4), and only kept the features with indices of $4k+1$, which is $\{D_1, D_5, D_9, ...\}$, as training and testing samples.

Let $\tilde{S} = \{S_1, S_2, ..., S_n\}$ denote the segment-wise prediction scores for an audio. The global prediction score S is computed based on $\tilde{S}$ as follows,

$$S = \frac{1}{n}\sum_{k=1}^{n} S_k + \max_k S_k \tag{2}$$

where $\frac{1}{n}\sum_{k=1}^{n} S_k$ and $\max_k S_k$ aggregate the scores by average and max functions, respectively. Average pooling and max pooling are two of the widely used pooling strategies utilized in many state-of-the-art classification applications [11, 27]. Here we define the global prediction score as the sum of $\frac{1}{n}\sum_{k=1}^{n} S_k$ and $\max_k S_k$, which obtains stable improvements with CNN fusion compared to applying average pooling or max pooling alone.

4.2 CNN Fusion

The fusion of different CNN classifiers can significantly improve the classification accuracy. We ignore the output layer of the 1D, 2D, and 3D neural networks and concatenate the internal representations into a multimodal feature vector to train a more robust classifier. More specifically, a 768 dimensional deep feature descriptor is generated from the three CNNs introduced in Section 3 as the representation of each audio segment. For the 2D and 3D CNNs, we take the output of the fully-connected layer and obtain a 256 dimensional feature descriptor for each input audio segment. For the 1D CNN, we take the output of conv6 and apply an additional max-pooling step to downsample the output of conv6 to a fixed 256 dimensional feature descriptor. For classification, we adopt a multi-layer neural network, which consists of two hidden layers followed by ReLU activation and one output layer with softmax function. The number of neurons in the hidden layers are 1024 and 512, respectively. Finally, we compute the audio-level fusion score by aggregating the output of the neural network using Eq. 2.

4.3 Network Ensemble

The results generated by the same CNN trained with the same dataset may still differ slightly due to the randomness in neural networks. To solve this problem, network ensembles [17] have been investigated, which combine the scores of individually trained neural networks to obtain improved classification accuracy. In our experiments, we train each of the CNNs introduced in Section 3 three times with random seeds and take their average as the final prediction scores for each class. Please note that as three slightly different internal representations can be generated from each modal, there are a total of $3\times3\times3 = 27$ types of feature combinations in terms of the feature concatenation. Therefore, we train the neural network in CNN fusion 27 times and again take the average of their outputs as the final prediction scores for each class.

5 EVALUATION

We describe the experimental setup in Section 5.1, and then proceed with the evaluations in two steps. First, we perform a step-by-step model justification to verify the effectiveness of our proposed methods. Next, we compare our system with the state-of-the-art approaches in acoustic scene classification.

5.1 Experimental Setup

We evaluated our proposed methods based on two large-scale audio datasets, namely the DCASE16 [23] and the LITIS Rouen [33] acoustic scene datasets. The DCASE16 dataset consists of 15 classes with 1170 samples for training and 390 samples for testing. The LITIS Rouen dataset consists of 3026 samples of 19 scene categories, which are further divided into 20-fold 80%-20% splits for evaluation. The samples in both datasets are 30 seconds in duration. The metric of classification accuracy, *i.e.*, the number of correctly classified samples among the total number of samples, was adopted as the evaluation criteria. For the DCASE16 dataset, we report the average classification accuracy over the 15 classes. For the LITIS dataset, we report the mean of the average classification accuracy[1] over the 20 folds.

The network architectures have been trained using stochastic mini-batch gradient descent based on back-propagation with momentum. The mini-batch size and the momentum were set to 32 and 0.9, respectively. The learning rate was set to 0.01 for the 1D CNN and 0.001 for the 2D and 3D CNNs with exponential decay. The dropout was set to 0.7 for the 1D CNN and 0.5 for the 2D and 3D CNNs. We implemented all the models using the TensorFlow library. For early stopping, we randomly selected 15% of the training data for validation and the network training was stopped if the average classification accuracy on the validation set did not increase by more than 20 epochs.

5.2 Step-By-Step Model Justification

Our proposed multimodal system includes two main components: segment-level score aggregation and multi-CNN fusion. To demonstrate the effectiveness of our proposed approaches in each step, we replace our method by a functionally reduced counterpart and

[1]This is also referred to as the average class-wise precision [29].

Table 4: Average classification accuracy comparison of different audio segment-level score aggregation methods on the DCASE16 dataset.

Classifiers	Average	Max	Ours
1D CNN	0.818	0.797	**0.821**
2D CNN	0.854	**0.856**	0.854
3D CNN	0.808	0.797	**0.810**
CNN Fusion	0.903	**0.910**	**0.910**

Table 5: Average classification accuracy comparison of different audio segment-level score aggregation methods on the LITIS Rouen dataset.

Classifiers	Average	Max	Ours
1D CNN	**0.895**	0.843	0.877
2D CNN	0.925	0.898	**0.926**
3D CNN	**0.845**	0.772	0.833
CNN Fusion	**0.964**	0.946	**0.964**

compare the corresponding classification accuracy. Next, we compare the performance of our proposed CNNs on video-level classification and segment-level classification, and discuss the characteristics of the proposed CNNs at the end of this section.

Evaluation on Segment Score Aggregation. We compared our segment-level score aggregation method based on Eq. 2 to average aggregation and max aggregation, and reported the average classification accuracy in Tables 4 and 5 with the **best** results highlighted. For individual classifiers, average aggregation generally outperformed max aggregation on both of the datasets. This is because the average aggregation is more robust to less representative segments that usually only take a small part of an audio. On the other hand, CNN fusion significantly improved the descriptiveness of audio representations, which enabled max aggregation to obtain competitive or even better results compared to average aggregation. Our method takes the advantages of both average aggregation and max aggregation. It obtained the best average classification accuracy with CNN fusion on both of the DCASE16 and the LITIS Rouen datasets.

Next, we compare the classification results obtained by individual classifiers and their fusion. The following observations hold on both datasets and regardless of which segment-level score aggregation method to use: (1) in terms of individual classifiers, the 2D CNN outperformed the 1D CNN and the 3D CNN, and (2) the fusion of the three CNNs significantly improved the average classification accuracy compared to individual classifiers. The experimental results are consistent with our expectations, as the 2D CNN based on traditional time-frequency sound representations is one of the most widely used classifiers in nowadays state-of-the-art acoustic scene classification systems [8, 22, 26]. The 1D and 3D CNNs are less effective if being compared individually, but significant improvements have been obtained by applying simple fusion techniques. While the 2D CNN based on time-frequency representations tends to emphasize more on the background noise, the 3D CNN based on spatial-temporal dynamics is more suitable for capturing the acoustic events. By using our proposed segment-level score aggregation method, CNN fusion outperformed the second best method, 2D CNN, by 6.6% and 4.1% on DCASE16 and LITIS Rouen, respectively. This indicates the deep features learnt from raw waveforms, time-frequency representations, and spatial-temporal dynamics are highly complementary to each other, and thus justifies the effectiveness of our proposed CNN fusion architecture based on multimodal features.

Table 6: Average classification accuracy comparison of CNN fusions on the DCASE16 and the LITIS Rouen datasets.

Classifiers	DCASE16	LITIS Rouen
1D+2D CNNs	0.862	0.946
1D+3D CNNs	0.874	0.951
2D+3D CNNs	0.890	0.952
1D+2D+3D CNNs	0.910	0.964

Evaluation on CNN Fusion. We aggregated segment scores based on Eq. 2 and compared the classification accuracy obtained by different combinations of CNN fusion. As can be seen from Table 6, CNN fusion significantly improved the average classification accuracy compared to individual CNN classifiers in all cases. The fusion of 1D and 2D, 1D and 3D, 2D and 3D CNNs outperformed the individual 2D CNN by 0.9%, 2.3%, 4.2% on the DCASE16 dataset, and by 2.2%, 2.7%, 2.8% on the LITIS Rouen dataset, respectively. This indicates that the features generated by our proposed 1D, 2D, and 3D CNNs are highly complementary to each other, as little benefits can be obtained by fusion if the features are correlated. More importantly, the fusion of 3D with 1D or 2D CNN outperformed the fusion of 1D and 2D CNNs on both datasets, which indicates that our proposed 3D CNN complements the existing 1D and 2D approaches and plays a key role to perform beyond the current state-of-the-art methods.

Segment-level Classification vs. Audio-level Classification. We have evaluated our proposed system on audio-level classification above. Here we also reported the classification accuracy on 1-s audio clips after segmentation in Tables 7 and 8. From the results we can see that classification on audios is much more accurate as audios are 30 seconds long and contain rich acoustic information for scene recognition. The time-frequency based 2D CNN performed the best on segment classification, which may indicate the 2D CNN tends to emphasize more on the background noise rather than on the acoustic event occurrences [44]. Comparatively, the extraction of changing dynamics emphasizes different spectral characteristics from neighboring regions in spatial-temporal domain, which makes it easier to detect acoustic events from audios. This explains why the 3D CNN performs less effective on segment-level classification. Audio clips of one second may only contain the background noise rather than the key events giving help with scene classification. The 3D CNN obtains an average classification accuracy of 81% and 83.3% on audio-level classification, which indicates the learnt acoustic events appear in the majority of audio clips as the 3D CNN is able to recognize scenes generally well from 30-s audios.

Table 7: Average classification accuracy comparison of segment-level classification and audio-level classification on the DCASE16 dataset.

Classifiers	Segment (1-s)	Audio (30-s)
1D CNN	0.710	0.821
2D CNN	0.788	0.854
3D CNN	0.667	0.810
CNN Fusion	0.819	0.910

Table 8: Average classification accuracy comparison of segment-level classification and audio-level classification on the LITIS Rouen dataset.

Classifiers	Segment (1-s)	Audio (30-s)
1D CNN	0.783	0.877
2D CNN	0.812	0.926
3D CNN	0.618	0.833
CNN Fusion	0.885	0.964

5.3 Comparison with the State-of-the-art

We first compared our proposed method to the state-of-the-art systems that ranked top ten in the DCASE challenge of acoustic scene classification 2016 [23]. The DCASE16 acoustic scene classification dataset is composed of 15 classes, namely bus, cafe/restaurant, car, city center, forest path, grocery store, home, beach, library, metro station, office, residential area, train, tram, and park. The average classification accuracy comparison of the systems is reported in Table 9 with the **best** result highlighted. We also illustrated the accuracy per class comparison to the top five systems in the DCASE16 acoustic scene classification challenge in Figure 3.

We can see that DNN-based classifiers have been widely used in nowadays state-of-the-art audio classification systems. Some researchers built their systems based on single classifiers. For example, Valenti *et al.* used a single CNN to classify short sequences of audio, represented by the log-mel spectrogram [44]. Others fused the scores obtained by different classifiers to boost the performance of their systems [8, 22, 26]. However, the features for classification in these systems are mostly traditional manually-crafted time-frequency representations such as MFCC and spectrogram. Eghbal-Zadeh *et al.* innovatively applied i-vector representation, which has been first introduced in the field of speaker verification, to acoustic scene classification [8]. They achieved the best classification accuracy of 89.7% by combining the i-vector based classifier with a CNN trained on spectrograms. Marchi *et al.* utilized a deep RNN for temporal modeling and obtained an average classification accuracy of 86.4% [22]. In comparison, we innovatively introduced a 3D CNN to model the spatial-temporal dynamics of audio excerpts, and outperformed the RNN-based system proposed by Marchi *et al.* by 5.3%. Moreover, our system focuses on the modeling of complementary features in different dimensions, which successfully improved the average classification accuracy by 1.5% $\sim$ 8.2% compared to the top ten systems in the DCASE16 acoustic scene classification challenge. From Figure 3 we can see that our proposed method

Table 9: Average classification accuracy comparison to the state-of-the-art approaches on the DCASE16 dataset.

Methods	Classifier	Accuracy
Bae *et al.* [3]	CNN-RNN	0.841
Lee *et al.* [12]	CNN	0.846
Lee *et al.* [15]	CNN ensemble	0.854
Takahashi *et al.* [41]	DNN-GMM	0.856
Kumar *et al.* [9]	SVM	0.859
Valenti *et al.* [44]	CNN	0.862
Marchi *et al.* [22]	fusion	0.864
Ko *et al.* [26]	fusion	0.872
Bisot *et al.* [5]	NMF	0.877
Eghbal-Zadeh *et al.* [8]	fusion	0.897
Ours	fusion	**0.910**

Table 10: Average classification accuracy comparison to the state-of-the-art approaches on the LITIS Rouen dataset.

Methods	Accuracy
HOG [33]	0.917
DNN + MFCC [28]	0.922
HOG + SPD [4]	0.933
Scene-LTE + Speech-LTE [31]	0.959
CNN-LTE [29]	0.963
Multimodal Fusion, Ours	**0.964**

obtained the top-two accuracy scores on 11 out of the 15 classes. It indicates the performance of our proposed method is more stable among different classes compared to the other systems. While the majority of the classes are quite easy to be recognized, there are exceptions such as library, train, and park, where the performance of different classifiers varies a lot. Such classes usually contain sounds of the same nature, which confuses classifiers and considerably affects their prediction results.

The confusion matrix of our proposed method on the DCASE16 dataset is illustrated in Figure 4, where X-axis indicates the predicted label and Y-axis indicates the true label. We can see that most confusions are between classes with similar backgrounds or containing acoustic events of the same nature. For example, our system confuses train with tram, library with forest path and home. The rest of the classes are classified rather easily. As can be seen that our proposed system obtained 100% accuracy on bus, car, forest path, grocery store, and metro station, and 96.2% accuracy on home, office,and tram where only one instance has been mistakenly classified.

Next, we compared our method to the state-of-the-arts on the LITIS Rouen audio scene dataset. This dataset is composed of 19 scene categories, namely plane, busy street, bus, cafe, car, train station hall, kid game hall, market, metro-paris, metro-rouen, billiard pool hall, quiet street, student hall, restaurant, pedestrian street, shop, train, high-speed train, and tubestation, but is less challenging compared to the DCASE16 dataset. As can be seen from Table 10, our proposed method outperformed all the competitors and obtained the best average classification accuracy. The LTE-based

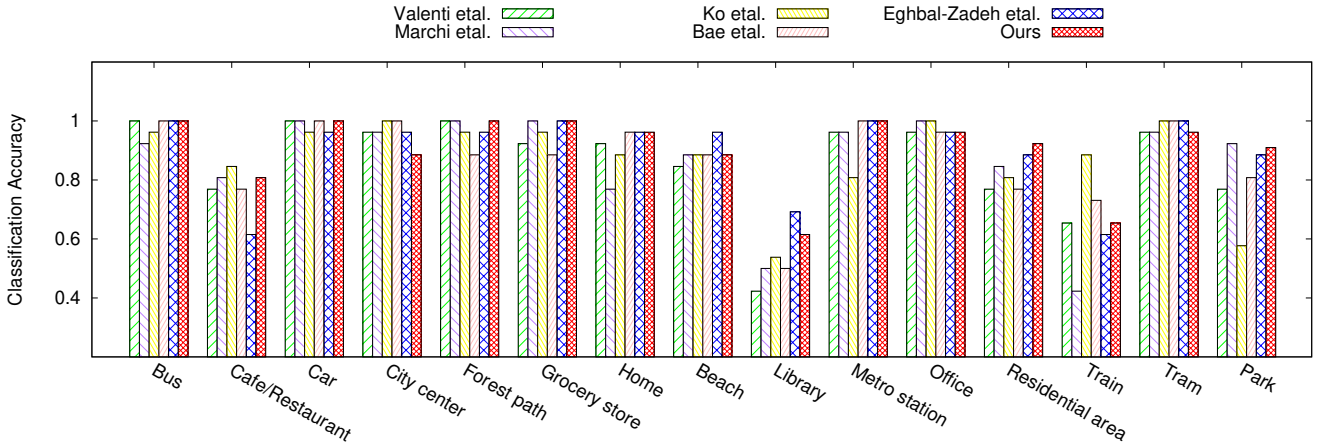

Figure 3: Classification accuracy comparison per class between our proposed method and the state-of-the-art approaches on the DCASE16 dataset.

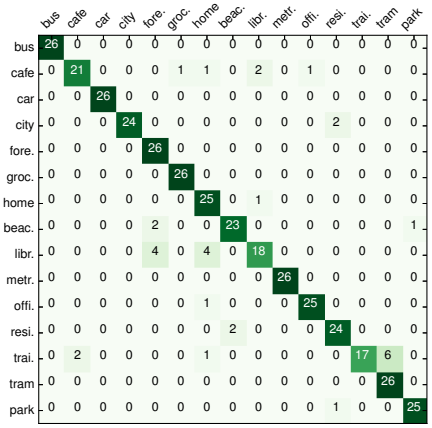

Figure 4: Confusion matrix of our proposed multimodal acoustic scene classification method on the DCASE16 dataset.

CNN fusion approach [29] proposed by Phan *et al.* achieved a competitive classification result compared to ours on the LITIS Rouen dataset. However, the accuracy of their method dropped when being applied to the more challenging DCASE16 dataset. Our method outperformed the CNN-LTE by 9.2% on the DCASE16 test set, as their method only obtained an average classification accuracy of 83.3% in this case [30]. Moreover, our method outperformed the rest of the state-of-the-art methods by 0.5% ∼ 5.1% on the LITIS Rouen dataset, which demonstrates the effectiveness of our proposed approach.

6 CONCLUSIONS

We have presented a novel acoustic scene classification system based on multimodal deep feature fusion. Three convolutional neural networks have been introduced to learn features from raw audio waveforms, time-frequency representations and spatial-temporal dynamic features, respectively. For score fusion, we concatenate internal representations of the three CNNs into a multimodal feature vector and adopt an additional three-layer neural network for classification. To deal the randomness in neural networks, CNN ensembles have been applied to further boost our system's performance. We have conducted extensive experiments on the DCASE16 dataset and the LITIS Rouen dataset for acoustic scene classification. The experimental results show that our proposed method obtains the state-of-the-art average classification accuracy, which outperforms existing audio classification systems by 1.5% ∼ 8.2% and 0.5% ∼ 5.1% on the two datasets respectively without applying any data augmentation techniques. As generally large dataset can help improve classification accuracy, potential improvements can be obtained by increasing the number of training samples based on audio deformation such as time stretching, pitch shifting, and background noise mixing.

Currently, we convert audios from stereo to mono before processing. In the future, we plan to investigate binaural representations for acoustic scene classification. The difference between the sounds recorded in two stereo tracks emphasizes the feature change in geospatial domain, which contains supplementary information and thus can be integrated as an additional feature in our proposed system. Moreover, several pre-trained models have become publicly available for deep acoustic feature extraction such as AudioSet and SoundNet. These models have been trained based on very large dataset with rich and diverse information. We also plan to take advantage of pre-trained models for classification by investigating transfer learning techniques.

ACKNOWLEDGEMENT

This research has been supported in part by Singapore's Ministry of Education (MOE) Academic Research Fund Tier 1, grant number T1 251RES1713. We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research.

REFERENCES

[1] J.-J. Aucouturier, B. Defreville, and F. Pachet. 2007. The Bag-of-frame Approach to Audio Pattern Recognition: A Sufficient Model for Urban Soundscapes But Not For Polyphonic Music. *Journal of the Acoustical Society of America* (2007), 881–91.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning Sound Representations from Unlabeled Video. In *Advances in Neural Information Processing Systems*. 892–900.

[3] Soo Hyun Bae, Inkyu Choi, and Nam Soo Kim. 2016. *Acoustic Scene Classification Using Parallel Combination of LSTM and CNN.* Technical Report.

[4] V. Bisot, S. Essid, and G. Richard. 2015. HOG and Subband Power Distribution Image Features for Acoustic Scene Classification. In *European Signal Processing Conference*. 719–723.

[5] Victor Bisot, Romain Serizel, Slim Essid, and Gaël Richard. 2016. *Supervised Nonnegative Matrix Factorization for Acoustic Scene Classification.* Technical Report.

[6] V. Bisot, R. Serizel, S. Essid, and G. Richard. 2017. Feature Learning With Matrix Factorization Applied to Acoustic Scene Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2017), 1216–1229.

[7] W. Dai, C. Dai, S. Qu, J. Li, and S. Das. 2017. Very Deep Convolutional Neural Networks for Raw Waveforms. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 421–425.

[8] Hamid Eghbal-Zadeh, Bernhard Lehner, Matthias Dorfer, and Gerhard Widmer. 2016. *CP-JKU Submissions for DCASE-2016: a Hybrid Approach Using Binaural I-Vectors and Deep Convolutional Neural Networks.* Technical Report.

[9] Benjamin Elizalde, Anurag Kumar, Ankit Shah, Rohan Badlani, Emmanuel Vincent, Bhiksha Raj, and Ian Lane. 2016. *Experiments on The DCASE Challenge 2016: Acoustic Scene Classification and Sound Event Detection in Real Life Recording.* Technical Report.

[10] Jinxi Guo, Ning Xu, Li-Jia Li, and Abeer Alwan. 2017. Attention Based CLDNNs for Short-Duration Acoustic Scene Classification. In *Interspeech*. 469–473.

[11] Yoonchang Han and Kyogu Lee. 2016. *Convolutional Neural Network with Multiple-Width Frequency-Delta Data Augmentation for Acoustic Scene Classification.* Technical Report. DCASE2016 Challenge.

[12] Yoonchang Han and Jeongsoo Park. 2017. Convolutional Neural Networks with Binaural Representations and Background Subtraction for Acoustic Scene Classification. In *Detection and Classification of Acoustic Scenes and Events Workshop*.

[13] Toni Heittola, Annamaria Mesaros, Antti Eronen, and Tuomas Virtanen. 2013. Context-dependent Sound Event Eetection. *EURASIP Journal on Audio, Speech, and Music Processing* (2013).

[14] 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. 2017. CNN Architectures for Large-scale Audio Classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 131–135.

[15] Jaehun Kim and Kyogu Lee. 2016. *Empirical Study on Ensemble Method of Deep Neural Networks for Acoustic Scene Classification.* Technical Report.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *International Conference on Neural Information Processing Systems*. 1097–1105.

[17] Anders Krogh and Jesper Vedelsby. 1994. Neural Network Ensembles, Cross Validation and Active Learning. In *International Conference on Neural Information Processing Systems*. 231–238.

[18] Anurag Kumar, Maksim Khadkevich, and Christian Fügen. 2017. Knowledge Transfer from Weakly Labeled Audio using Convolutional Neural Network for Sound Events and Scenes. *CoRR* (2017). http://arxiv.org/abs/1711.01369

[19] Zhen-zhong Lan, Lei Bao, Shoou-I Yu, Wei Liu, and Alexander G. Hauptmann. 2012. Double Fusion for Multimedia Event Detection. In *Advances in Multimedia Modeling*. 173–185.

[20] T. Y. Lin, A. RoyChowdhury, and S. Maji. 2015. Bilinear CNN Models for Fine-Grained Visual Recognition. In *IEEE International Conference on Computer Vision*. 1449–1457.

[21] Hong Liu, Juanhui Tu, and Mengyuan Liu. 2017. Two-Stream 3D Convolutional Neural Network for Skeleton-Based Action Recognition. *CoRR* (2017). http://arxiv.org/abs/1705.08106

[22] Erik Marchi, Dario Tonelli, Xinzhou Xu, Fabien Ringeval, Jun Deng, Stefano Squartini, and BjÃűrn Schuller. 2016. *The Up System for The 2016 DCASE Challenge Using Deep Recurrent Neural Network and Multiscale Kernel Subspace Learning.* Technical Report.

[23] A. Mesaros, T. Heittola, and T. Virtanen. 2016. TUT Database for Acoustic Scene Classification and Sound Event Detection. In *European Signal Processing Conference*. 1128–1132.

[24] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. 2015. Is Object Localization for Free? - Weakly-supervised Learning with Convolutional Neural Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*. 685–694.

[25] E. Park, X. Han, T. L. Berg, and A. C. Berg. 2016. Combining Multiple Sources of Knowledge in Deep CNNs for Action Recognition. In *IEEE Winter Conference on Applications of Computer Vision*. 1–8.

[26] Sangwook Park, Seongkyu Mun, Younglo Lee, and Hanseok Ko. 2016. *Score Fusion of Classification Systems for Acoustic Scene Classification.* Technical Report.

[27] Otávio A. B. Penatti, Lin Tzy Li, Jurandy Almeida, and Ricardo da S. Torres. 2012. A Visual Approach for Video Geocoding Using Bag-of-scenes. In *ACM International Conference on Multimedia Retrieval*. 53:1–53:8.

[28] Y. Petetin, C. Laroche, and A. Mayoue. 2015. Deep Neural Networks for Audio Scene Recognition. In *European Signal Processing Conference*. 125–129.

[29] Huy Phan, Lars Hertel, Marco Maass, Philipp Koch, Radoslaw Mazur, and Alfred Mertins. 2017. Improved Audio Scene Classification Based on Label-Tree Embeddings and Convolutional Neural Networks. *IEEE/ACM Transactions on Audio, Speech and Language Processing* (2017), 1278–1290.

[30] Huy Phan, Lars Hertel, Marco Maass, Philipp Koch, and Alfred Mertins. 2016. *CNN-LTE: a Class of 1-X Pooling Convolutional Neural Networks on Label Tree Embeddings for Audio Scene Recognition.* Technical Report. DCASE2016 Challenge.

[31] Huy Phan, Lars Hertel, Marco Maass, Philipp Koch, and Alfred Mertins. 2016. Label Tree Embeddings for Acoustic Scene Classification. In *ACM on Multimedia Conference*. 486–490.

[32] A. Rakotomamonjy. 2017. Supervised Representation Learning for Audio Scene Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2017), 1253–1265.

[33] A. Rakotomamonjy and G. Gasso. 2015. Histogram of Gradients of Time-Frequency Representations for Audio Scene Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2015), 142–153.

[34] J. Salamon and J. P. Bello. 2017. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters* (2017), 279–283.

[35] J. Schröder, N. Moritz, J. Anemüller, S. Goetze, and B. Kollmeier. 2017. Classifier Architectures for Acoustic Scenes and Events: Implications for DNNs, TDNNs, and Perceptual Features from DCASE 2016. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2017), 1304–1314.

[36] Rajiv Shah and Roger Zimmermann. 2017. *Multimodal Analysis of User-generated Multimedia Content.* Springer.

[37] Rajiv Ratn Shah, Yi Yu, Akshay Verma, Suhua Tang, Anwar Dilawar Shaikh, and Roger Zimmermann. 2016. Leveraging Multimodal Information for Event Summarization and Concept-level Sentiment Analysis. *Knowledge-Based Systems* (2016), 102–109.

[38] Rajiv Ratn Shah, Yi Yu, and Roger Zimmermann. 2014. Advisor: Personalized Video Soundtrack Recommendation by Late Fusion with Heuristic Rankings. In *ACM on Multimedia Conference*. 607–616.

[39] Karen Simonyan and Andrew Zisserman. 2014. Two-stream Convolutional Networks for Action Recognition in Videos. In *International Conference on Neural Information Processing Systems*. 568–576.

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition*. 2818–2826.

[41] Gen Takahashi, Takeshi Yamada, Shoji Makino, and Nobutaka Ono. 2016. *Acoustic Scene Classification Using Deep Neural Network and Frame-Concatenated Acoustic Feature.* Technical Report.

[42] Naoya Takahashi, Michael Gygli, and Luc Van Gool. 2017. AENet: Learning Deep Audio Features for Video Analysis. *CoRR* (2017). http://arxiv.org/abs/1701.00599

[43] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *IEEE International Conference on Computer Vision*. 4489–4497.

[44] Michele Valenti, Aleksandr Diment, Giambattista Parascandolo, Stefano Squartini, and Tuomas Virtanen. 2016. *DCASE 2016 Acoustic Scene Classification Using Convolutional Neural Networks.* Technical Report.

[45] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. *CoRR*. https://arxiv.org/abs/1609.03499

[46] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. 2015. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification. In *ACM on Multimedia Conference*. 461–470.

[47] W. Yang and S. Krishnan. 2017. Combining Temporal Features by Local Binary Pattern for Acoustic Scene Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2017), 1315–1321.

[48] Jiaxing Ye, Takumi Kobayashi, Masahiro Murakawa, and Tetsuya Higuchi. 2015. Acoustic Scene Classification Based on Sound Textures and Events. In *ACM on Multimedia Conference*. 1291–1294.

[49] Y. Yin, Z. Liu, Satyam, and R. Zimmermann. 2016. Laplacian Sparse Coding of Scenes for Video Classification. In *IEEE International Symposium on Multimedia*. 499–506.

[50] Y. Yin, Z. Liu, and R. Zimmermann. 2017. Geographic Information Use in Weakly-supervised Deep Learning for Landmark Recognition. In *IEEE International Conference on Multimedia and Expo*. 1015–1020.