Attentional Graph Convolutional Network for Structure-aware Audio-Visual Scene Classification

Liguang Zhou, *Student Member, IEEE,* Yuhongze Zhou, Xiaonan Qi, Junjie Hu,
Tin Lun Lam, *Senior Member, IEEE,* and Yangsheng Xu, *Fellow, IEEE*

***Abstract*—Audio-Visual scene understanding is a challenging problem due to the unstructured spatial-temporal relations that exist in the audio signals and spatial layouts of different objects in the visual images. Recently, many studies have focused on abstracting features from convolutional neural networks while the learning of explicit semantically relevant frames of sound signals and visual images has been overlooked. To this end, we present an end-to-end framework, namely attentional graph convolutional network (AGCN), for structure-aware audio-visual scene representation. First, the spectrogram of sound and input image is processed by a backbone network for feature extraction. Then, to build multi-scale hierarchical information of input signals, we utilize an attention fusion mechanism to aggregate features from multiple layers of the backbone network. Notably, to well represent the salient regions and contextual information, the salient acoustic graph (SAG) and contextual acoustic graph (CAG), salient visual graph (SVG), and contextual visual graph (CVG) are constructed for the audio-visual scene representation. Finally, the constructed graphs pass through a graph convolutional network for structure-aware audio-visual scene recognition. Extensive experimental results on the audio, visual and audio-visual scene recognition datasets show that promising results have been achieved by the AGCN methods. Visualizing graphs on the spectrograms and images have been presented to show the effectiveness of proposed CAG/SAG and CVG/SVG that could focus on the salient and semantic relevant regions. The visualization results indicate that AGCN can focus on the semantic relevant regions.**

***Index Terms*—Attentional graph convolutional network, salient acoustic graph, contextual acoustic graph, salient visual graph, contextual visual graph, structure-aware audio-visual representation**

I. Introduction

Audio-Visual scene understanding is an important research direction for human-robot interaction. Via ambient audio and visual information of the environment, the robot is capable of knowing the environment that it is traveling by. For instance, audio-visual scene understanding has been applied in various applications, including robot hearing [1], relative bearing estimation of robots [2], smart homecare surveillance system [3], place recognition for mobile robots [4], image quality assessment [5], human activity recognition [6], elderly care [7], automatic construction [8], robot discovery of the auditory scene [9], environmental understanding for self-driving cars [10], and heart sound classification [11].

The audio-visual understanding is challenging due to the intrinsic complex nature of audio signals and visual images. For example, the acoustic scene is composed of dynamic and unstructured patterns, which are not semantically meaningful based on its log-Mel spectrogram as shown in Figure 1. The visual scene contains various objects spatially distributed over different locations, and different repetitive textures that exist in different scenes as shown in Figure 2. To sum up, it is challenging to design an algorithm that effectively captures the texture and characteristics for audio-visual scene representations.

Various efforts have been made to address this challenge for acoustic scene classification (ASC). For example, some works use features like the local binary pattern (LBP) [12], Log-Mel [13], wavelet [14], [15], and Mel-frequency cepstral coefficients (MFCCs) [16]. State-of-the-art solutions on ASC mostly utilize the Log-Mel features [17]. To conduct the ASC task, the handcrafted features are first extracted from the audio signal waveform. Then, the trained discriminative models such as support vector machine (SVM) [18] and Gaussian mixture models (GMM) [19] classify environmental sounds.

Benefiting from the remarkable progress of deep learning, recent works have exploited the usage of the convolutional neural network (ConvNet) to classify the spectrograms of sounds [20], [17]. However, ConvNet only captures local spatial features since convolutional filters focus on the small region of interest (ROI). They neglect the global connection between regions of interest that lies at different locations of visual images or acoustic signals. Further, to capture the salient features that are more relevant to the sound events, attention mechanisms have been proposed [21] [22]. However, this pooling mechanism only considers the strength of the sound signal but neglects the spatial connection of different sound spectrogram patches.

For visual scene classification (VSC), there has been a longer history and many approaches been proposed, such as object-based methods [23], [24], [25], [26], bag-of-visual-words [27], global image features [28], color descriptors [29], VGGNet [30], ConvNet [31]. Until recently, graph convolu-

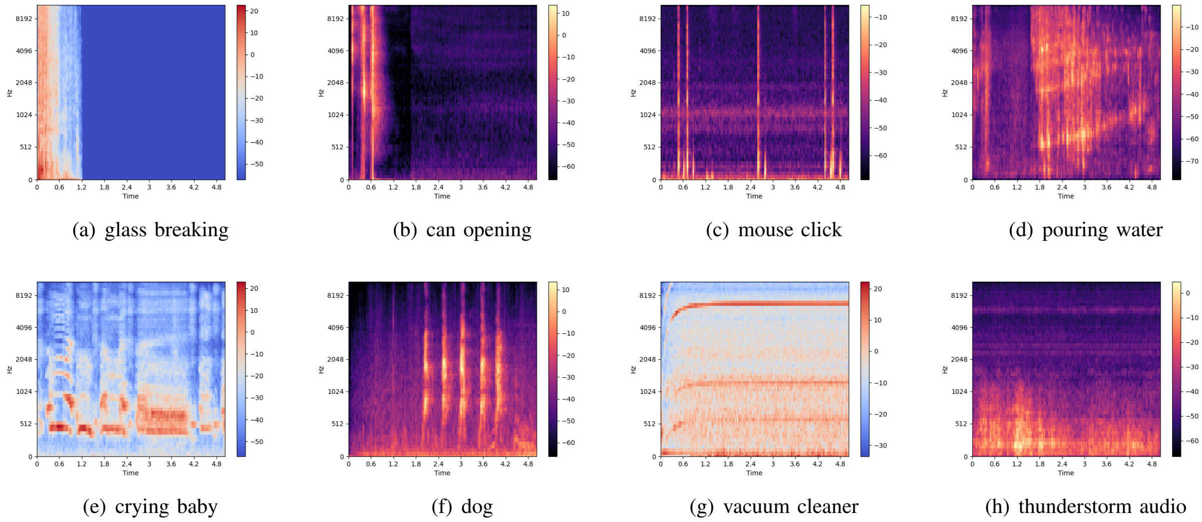
Manuscript received September 20, 2021; revised XXXXX XX, XXXX. This work was partly supported by the National Natural Science Foundation of China under Grant 62073274, the Shenzhen Science and Technology Program under the Grant JCYJ20220818103000001, and the Shenzhen Institute of Artificial Intelligence and Robotics for Society under Grant AC01202101103. (Corresponding Author: Tin Lun Lam.)

L. Zhou, J. Hu, T. Lam, and Y. Xu are with Shenzhen Institute of Artificial Intelligence and Robotics for Society, The Chinese University of Hong Kong, Shenzhen, China, (email:liguangzhou@link.cuhk.edu.cn, hujunjie@cuhk.edu.cn, tllam@cuhk.edu.cn, ysxu@cuhk.edu.cn)

L. Zhou, T. Lam, and Y. Xu are with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China

Y. Zhou is with McGill University, Montréal, Canada, (email:yuhongze.zhou@mail.mcgill.ca)

X. Qi is with Carnegie Mellon University, Pittsburgh, PA, USA,(email:xiaonanq@andrew.cmu.edu)



(a) glass breaking (b) can opening (c) mouse click (d) pouring water

(e) crying baby (f) dog (g) vacuum cleaner (h) thunderstorm audio

Fig. 1. Examples of log-Mel spectrograms of various environmental sounds in the ESC-50 dataset.

(a) balcony (b) bathroom (c) bedroom (d) closet (e) dining room (f) garage (g) home office

(h) home theater (i) kitchen (j) laundromat (k) living room (l) playroom (m) staircase (n) wet bar

Fig. 2. Samples of each class of Places365-14 dataset.

tional network (GCNs) has been proposed to improve scene recognition [32], [33], [34]. These methods have shown the merits of GCNs over ConvNets. However, they only model the most distinctive features or deep semantic/object features without considering the contextual background information of backgrounds, which plays a vital role for VSC.

Different from past works, we present an AGCN for the audio-visual scene classification (AVSC), as shown in Figure 3. To represent the audio-visual scene in a structure-aware manner, we leverage GCNs, which could be a suitable but less exploited choice. There are three steps to achieving this. First, to extract feature representation of the audio-visual scene, we use ResNet as a backbone network. Then, to augment the obtained features with multi-scale information, we aggregate the feature maps of ResNet at multiple layers with an attention fusion module [35]. Last, to achieve structure-aware audio-visual scene understanding, we construct graphs that can focus on the regions of interest and pay attention to the context information. Specifically, we construct the SAG/SVG that takes the maximum nodes selected from the pyramid feature map and the CAG/CVG that takes the medium nodes selected from the pyramid feature map. The SAG/SVG is regarded as the distinctive region of interest (ROIs) of audio-visual signals while the contextual acoustic graph is viewed as contextual information of audio-visual signals. To refine the graph features, these obtained graphs are processed by the GCNs. Finally, the classification results are estimated through the fully connected layer.

To demonstrate the effectiveness of the proposed method, the experiments are conducted on ASC tasks including ESC-10 and ESC-50 datasets, VSC tasks including Places365-7, Places365-14, and SUN-RGBD datasets, and AVSC task with DCASE2021 datasets, which proves the proposed method reaches comparable performance on these tasks. To further show the difference between the proposed method and the baseline method, the selected CAG/SAG and CVG/SVG are visualized among those audio spectrograms and visual images, which indicates both the semantically relevant regions and contextual regions are captured during the learning process.

To summarize, our main contribution to this work can be summarized in four folds:

- We propose a simple and effective attentional graph convolutional neural network (AGCN) that achieves structure-aware audio-visual scene classification.

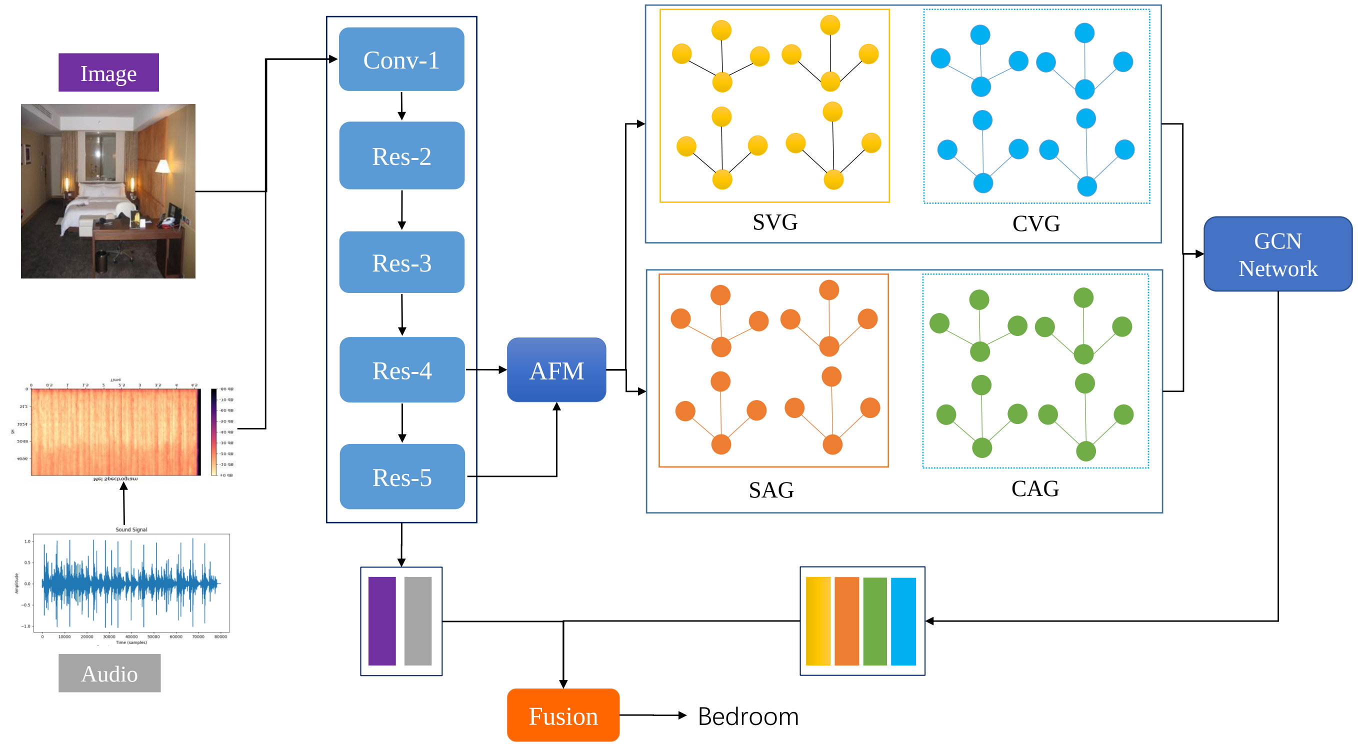


Fig. 3. The proposed audio-visual scene recognition framework. First, the input images and sound spectrograms are processed by a backbone network and intermediate feature maps are obtained. These features are fused by AFM module to expand its feature maps and receptive fields. Then, we constructs the SAG (in orange), CAG (in green), SVG (in yellow), and CVG (in blue) for audio-visual scene representation with a graph convolutional network. Hence, the refined features are obtained. Last, the refined graph features and backbone features are fused for audio-visual scene recognition.

- We construct and combine the salient acoustic graph (SAG), contextual acoustic graph (CAG), salient visual graph (SVG), and contextual visual graph (CVG) to represent the discriminative regions and background contextual regions of audio-visual scene inputs.
- Extensive experimental results of our method on the ASC, VSC, and AVSC datasets show that our method has achieved comparable performance.
- Visualization results show the proposed AGCN can focus on the semantically relevant regions and contextual regions.

The rest of the paper is organized as follows. The related works are presented in Section II. The AGCN network is described in Section III. The experimental results are shown in Section IV. The conclusion and future work about our research comes in Section V.

II. Related Works

A. Audio-Visual Scene Classification

AVSC is to establish the feature representation of audio-visual input signals and classify it. Hu et al. [36] create the ADVANCE for aerial scene recognition with audio-visual scene data. Wang et al. [37] introduce a dataset of urban scenes for audio-visual scene analysis. The AVSC task consists of two modules including the audio module and visual module. For the audio module, there are various networks designed to tackle ASC tasks. In particular, with the rapid development of deep learning in image classification, numerous works have exploited the usage of ConvNet to classify the spectrograms of sounds[20]. To capture the salient features that are more relevant to the sound events, attention mechanisms have been introduced [21] [22]. Temporal attention is proposed to capture the temporal structure of sound [21]. A sparse key-point encoding and efficient multispike learning framework are proposed for this problem [38]. Parallel temporal-spectral attention is proposed to enhance the representation of both the temporal and spectral features by capturing the importance of different time frames and frequency bands [22]. Zhou et al. [39] shows a feature pyramid attention structure is effective for acoustic scene recognition. These attention mechanisms enhance the feature representations in the channel and spatial level but lack preserving the multi-level feature representation and neglect the relative position of sound features. To this end, we utilize a latent feature fusion attention module to incorporate the multi-level feature for multi-scale feature extraction.

Even though extensive research has been done to improve the performance of ASC, the construction of salient acoustic graphs and contextual acoustic graphs based on graph neural networks is missing from the literature. This is important because the acoustic time-frequency features contain both local and global informative patches for ASC while the convolutional neural network based methods are incompetent of effectively associating these features.

VSC is the task of recognizing the scene with visual inputs. For the VSC task, there are additional semantic or depth information being incorporated as an additional branch for VSC. For instance, the semantic aided scene recognition become popular in research area. Ni et al. [10] propose an improved faster RCNN to extract representative objects as local features for scene recognition. Zhou [25] et al. present a Bayesian object pair modeling method to enhance the scene representation ability of network with additional semantic information. Miao [26] et al. develop a branch of scene embedding network to improve the scene representation. These methods has achieved comparable performance over single branch methods in some extend. However, the semantic branch requires cost expensive computations which limits the applications of these algorithms. To model scene representation with GCNs, an adaptive cross-modal GCNs has been proposed [32], where the most distinctive features are selected for visual scene representation. Besides, the semantic region masks are selected for graph modeling with GCNs to model the scene representation [34]. Segmented regions are utilized to model the remote scene representation in [33]. These methods have shown the merits of GCNs over ConvNets on scene representation. However, the background contextual information, which is important for scene representation as well, is abandoned, leading to information loss for scene representation.

To unite the audio and visual modules, we present a novel AGCN method based on the graph convolutional network. The proposed method first augments the feature representation of the backbone network by utilizing an attention-based attention fusion module [35]. Especially, salient graphs and contextual graphs are proposed to focus on the most distinctive regions and contextual regions of both audio and visual features.

B. Graph Convolutional Networks

Graph convolutional network is designed for non-Euclidean and graph-structured data, such as text documents, bioinformatic data, and scene signal [40][41]. GCNs have been successfully applied in various tasks, including scene recognition [33], [32], action recognition [42], image recognition [43], and emotion recognition [44]. However, the application of GCNs in audio signal analysis remains underexplored [45], [46]. There are two major types of data processing methods in graph convolutional networks, spatial [42] and spectral [32]. For the spatial graph, the convolutional filters are directly applied to the nodes and neighbors, which is efficient but ignores locality information. For the spectral graph, the convolutional filters are performed on the spectral domain, where the data nodes are processed with the Laplacian matrix and convolutional filters. In this paper, we use the spectral graph for graph feature refining.

III. Methodology

As shown in Figure 3, our proposed AGCN network mainly contains four steps. The first step is the backbone network for feature extraction where we use ResNet50 as the backbone network. The second step is the graph feature construction, where the attention fusion module (AFM) that refines and aggregates the different levels of feature maps from the backbone network is utilized [35]. The third step is graph construction and processing, where we construct salient and contextual graphs for audio-visual feature representation, including the SAG/CAG and the SVG/CVG. Specifically, we utilize the most distinctive and average nodes from the feature maps as a mimic of salient regions and contextual regions for feature abstracting. The detailed graph construction procedure is presented in Algorithm 1. Sequentially, these constructed graphs are fed into a spectral graph convolutional network for feature extraction. Last, with a fully connected layer attached, the classification results are obtained.

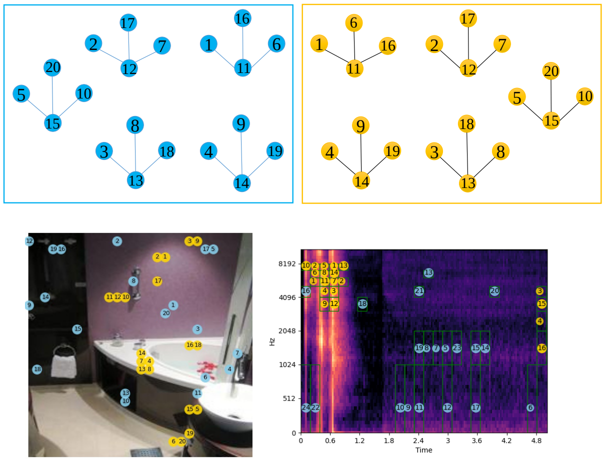


Fig. 4. In the first row, the left column shows the CAG/CVG in blue, and the right column shows the SAG/SVG in yellow. In the second row, the visualization of CVG and SVG of bedroom is shown in the left column, while the visualization of CAG and SAG of can opening acoustic spectrogram are shown in right column.

A. Graph Feature Construction

Consider the acoustic signals as x_a. The weights of the ConvNet backbone network and AFM module are represented as W_{convs} and W_{afm} respectively. The input feature x_v will be processed by these two modules consecutively.

$$F_{m_2}, F_{m_3}, F_{m_4} = W_{convs}(x_a) \tag{1}$$

$$F_{ffr} = W_{afm}(F_{m_3}, F_{m_4}) \tag{2}$$

where F_{ffr} is the fused feature representation extracted by the ResNet50 backbone and AFM network.

To select nodes from refined feature maps for graph construction, we first sum the fused feature maps F_{ffr} along the channel dimension as:

$$F_{rei} = \sum_{i=1}^{C} F_{ffr}(N, i, H, W) \tag{3}$$

The tensor shape of F_{ffr} is (N, C, H, W), where N denotes batch size, C represents the number of channels, H is the height of F_{ffr}, and W is the width of F_{afm}. We propose to select K most distinctive and K average feature vectors to form salient scene graph and contextual scene graph for the graph nodes V.

Then the reshaped intensity map F_{rei} is reshaped to $(N, H * W)$, where each pixel shows the intensity of features. As acoustic signal contains natural spatial-temporal relationships, these relationships are essential to ASC. Similarly, visual images contain the intrinsic spatial layouts of various objects and textures, which is important for visual scene representation. To preserve the spatial information, we constructed graphs for audio-visual scene representation with geometric relations.

B. Audio-Visual Graph Construction

The unstructured spatial-temporal relationships of acoustic spectrograms in acoustic scenes are difficult to learn. The spatial layouts of various objects in visual scenes are hard to capture. To address these issues, we decompose the audio-visual signals into two novel graphs, SAG/SVG to represent the most distinctive and semantic relevant regions of audio spectrograms and visual images, and CAG/CVG to represent the background contextual information of spectrograms and visual images. Since we only select the distinctive regions and contextual regions of features, this modeling method can skip the silent and semantically irrelevant information of audio-visual scenes.

Let's take the composition of SAG as an example. The SAG/CAG can be represented as $G_{sag} = (V, E, A)$ and $G_{cag} = (V, E, A)$. V represents the representative features selected from F_{rei}. E denotes the edges of the graph. A represents the adjacency matrix of the graph, which preserves the relative position of the acoustic signals.

1) Nodes Selection: The detailed construction of scene graphs are illustrated in the Algorithm 1. K denotes the nodes number of graph.

Algorithm 1 Algorithm for constructing SAG, CAG, SVG, and CVG.

```
Input: The reshaped intensity map $F_{rei}$;
Output: The indexes of $K$ selected feature vectors $ind_{sel}$;
Sort N reshaped intensity maps $F_{rei}$ with descending order, and select the top $K$ and medium $K$ indexes to $ind_F$;
$m_{left} = \frac{W \times H}{2} - \frac{K}{2}$;
$m_{right} = \frac{W \times H}{2} + \frac{K}{2}$;
for $i = 0 \rightarrow N$ do
    $index = sort(F_{rei}(i))$;
    $ind_{sag}(i) = index(1 : K) + index(m_{left} : m_{right})$;
end for
Sort the $N$ indexes $ind_F$ with ascending order to keep the acoustic signals with original spatial order;
for $i = 0 \rightarrow N$ do
    $ind_{sel}(i) = sort(ind_F(i))$;
end for
return $ind_{sel}$;
```

The nodes of acoustic graphs $V = \{v_i | i = 1, \ldots, K\}$ are associated with the selected features from $2K$ different locations of F_{ffr}, with K locations for SAG and the other K locations for CAG. The nodes selection method is shown in Algorithm 1. The graph nodes assignment can be formulated as $V_{sag} = \{v_i = F_{ffr}(N, C, i) | i = ind_{sel}(1), \ldots, ind_{sel}(K)\}$ and $V_{cag} = \{v_i = F_{ffr}(N, C, i) | i = ind_{sel}(K+1), \ldots, ind_{sel}(2K)\}$, where F_{ffr} is reshaped to $(N, C, H * W)$, and the shape of graph nodes V is (N, K, C).

2) Subgraph Construction: The total number of nodes of SAG or CAG is K. The SAG or CAG is made of N/4 subgraphs (sub-textures), that is, each subgraph is composed of 4 nodes. The subgraph is automatically constructed followed by the subgraph set G_{sub} and the center set of subgraph G_{csub}:

$$\begin{aligned} G_{sub} =& \{i, i + \frac{1}{4}K, i + \frac{2}{4}K, i + \frac{3}{4}K \mid \\ & i \in [1, K/4], i \in \mathbb{Z}, K \in \mathbb{Z}, K/4 == 0\} \\ G_{csub} =& \{i + \frac{2}{4}K \mid \\ & i \in [1, K/4], i \in \mathbb{Z}, K \in \mathbb{Z}, K/4 == 0\} \end{aligned} \tag{4}$$

Figure 4 shows the case when the node number of the graph is 20 and there are 5 subgraphs with the center of each subgraph assigned 11, 12, 13, 14, and 15. In the first row, 40 distinct nodes are selected from the feature map to construct the SAG and CAG, where 20 most distinctive nodes construct a SAG/SVG in gold and 20 average nodes construct a CAG/CVG in blue. In the second row, the proposed graphs are visualized on a bedroom image and a can opening acoustic signal. It is obvious that the nodes of SVG/SAG mostly locate in the most distinctive area of visual image and sound signals, which indicates the importance of this region. Besides, the nodes of the CVG/CAG are scatted around the whole image and spectrograms to preserve the small details of the contextual background of the visual image and sound event. The combination of these two graphs can well represent the visual image and acoustic signals across the entire spectrograms of both representative objects and background contextual information.

3) Geometric Relation: Since the spatial-temporal relationship of acoustic signals and spatial layouts of visual images are essential for AVSC, to utilize this relationship, we construct the adjacency matrix A according to the geometric relations $\mathbf{E_{ij}}$ among adjacent nodes.

$$A_{ij} = \begin{cases} \left|x_i - x_j\right| + \left|y_i - y_j\right|, E_{ij} \neq 0 \\ 0, E_{ij} = 0 \end{cases} \tag{5}$$

where (x_i, y_i) and (x_j, y_j) represent the position of i-th node and j-th node. $E_{ij} \neq 0$ means that there is a connection between i-th node and j-th node.

C. Spectral Graph Convolution

Given the acoustic graphs $G(V, E, A)$, the Laplacian matrix L is obtained by L=D-A, where D is degree matrix of graph G, and A is the adjacency matrix that records the geometric relations of nodes. To normalize the Laplacian matrix, we use the symmetric normalized Laplacian matrix,

$$L^{sym} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I_k - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{6}$$

where I_K is the identity matrix. With the added self-connections, $\hat{A}$ is denoted as $\hat{A} = A + I$. Then, the GCNs follows the same setting as [41]:

$$Y = (D + I_k)^{-\frac{1}{2}} \hat{A} (D + I_k)^{-\frac{1}{2}} X\Theta \tag{7}$$

$$Y = L_{norm} X \Theta_j \tag{8}$$

Each Θ_j is implemented with the vanilla convolution layer. To reduce the dimension of acoustic graphs, we employ a 1x1 convolution layer with 256 channels and then ReLU is added as the activation function.

IV. Experiments and Discussions

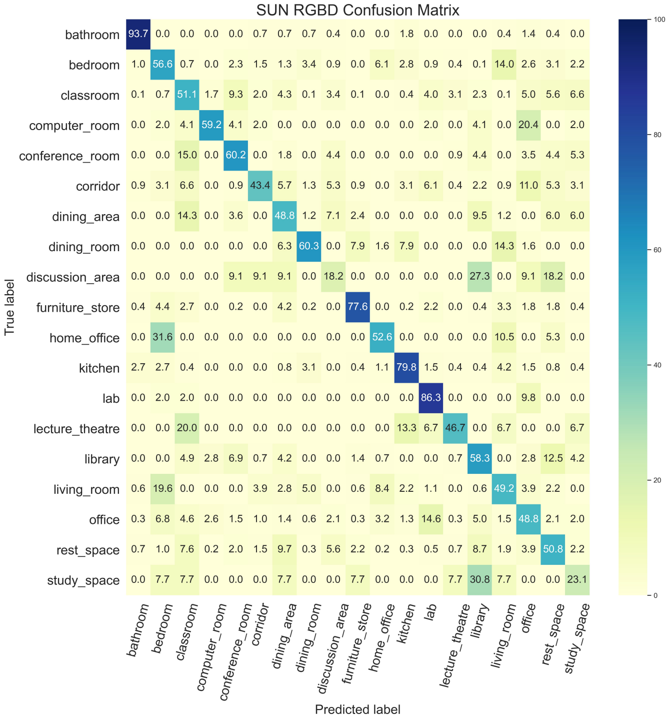
In this section, we examine the performance of AGCN on VSC, ASC, and AVSC tasks. Then, the visualization of the proposed SAG/CAG and SVG/CVG are analyzed. These detailed evaluations with extensive results are shown to prove the effectiveness of the proposed AGCN.

A. Visual Scene Recognition

1) Datasets: We use the Places365-7 and Places365-14 datasets for the evaluation of the proposed AGCN. The dataset split is the same as BORM [25]. For the Places365-7 dataset, there are 7 classes, 35000 images as the training set, and 701 images as the test set. For the Places365-14 dataset, there are 14 classes, and 75000 images for the training set, 1500 images for the test set. The SUN RGB-D is also utilized. Following the same settings as in [47]. This dataset is composed of 19 classes, in which 4845 images as training set and 4659 images as test set. Besides, it contains depth module information for RGB-D scene recognition.

2) Implementation Details: The proposed model is implemented in Pytorch [48]. All the baseline CNN models are pretrained on the ImageNet [49]. The SGD optimizer is used for training in experiments. The initial learning rate is 0.01 and decreased by 10 at every 20 epochs. The momentum of the optimizer is 0.9. The total train epochs are 60. The network structure for visual scene recognition in detail is listed in Table II.

3) Comparison with Stat-of-the-arts: First, our method is evaluated on two datasets, Places365-7 and Places365-14, the most commonly used datasets for scene recognition. We compare our proposed model with existing state-of-the-art methods. Table III demonstrates the performance of the AGCN algorithm on the Places365-7. The proposed AGCN reaches the competitive performance on both datasets without the semantic head. Instead, the BORM and OTS are semantic-based methods, i.e., these methods require the strong semantic priors obtained from the semantic segmentation network. In contrast, our AGCN has no prior information and still reaches the competitive performance compared with the BORM and OTS. Besides, though AGCN is based on the ResNet50 backbone, it still outperforms ResNet50 with 9% accuracy on the Places365-7 dataset. Besides, as indicated in able VII, the proposed AGCN only has around one-fifth of parameters and about one-third GFlops as BORM, showing the efficiency of the proposed AGCN over the semantic-based methods.



	bathroom	bedroom	classroom	computer_room	conference_room	corridor	dining_area	dining_room	discussion_area	furniture_store	home_office	kitchen	lab	lecture_theatre	library	living_room	office	rest_space	study_space
bathroom	93.7	0.0	0.0	0.0	0.0	0.7	0.7	0.7	0.4	0.0	0.0	1.8	0.0	0.0	0.0	0.4	1.4	0.4	0.0
bedroom	1.0	56.6	0.7	0.0	2.3	1.5	1.3	3.4	0.9	0.0	6.1	2.8	0.9	0.4	0.1	14.0	2.6	3.1	2.2
classroom	0.1	0.7	51.1	1.7	9.3	2.0	4.3	0.1	3.4	0.1	0.0	0.4	4.0	3.1	2.3	0.1	5.0	5.6	6.6
computer_room	0.0	2.0	4.1	59.2	4.1	2.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	4.1	0.0	20.4	0.0	2.0
conference_room	0.0	0.0	15.0	0.0	60.2	0.0	1.8	0.0	4.4	0.0	0.0	0.0	0.0	0.9	4.4	0.0	3.5	4.4	5.3
corridor	0.9	3.1	6.6	0.0	0.9	43.4	5.7	1.3	5.3	0.9	0.0	3.1	6.1	0.4	2.2	0.9	11.0	5.3	3.1
dining_area	0.0	0.0	14.3	0.0	3.6	0.0	48.8	1.2	7.1	2.4	0.0	0.0	0.0	0.0	9.5	1.2	0.0	6.0	6.0
dining_room	0.0	0.0	0.0	0.0	0.0	0.0	6.3	60.3	0.0	7.9	1.6	7.9	0.0	0.0	0.0	14.3	1.6	0.0	0.0
discussion_area	0.0	0.0	0.0	0.0	9.1	9.1	9.1	0.0	18.2	0.0	0.0	0.0	0.0	0.0	27.3	0.0	9.1	18.2	0.0
furniture_store	0.4	4.4	2.7	0.0	0.2	0.0	4.2	0.2	0.0	77.6	0.0	0.2	2.2	0.0	0.4	3.3	1.8	1.8	0.4
home_office	0.0	31.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	52.6	0.0	0.0	0.0	0.0	10.5	0.0	5.3	0.0
kitchen	2.7	2.7	0.4	0.0	0.0	0.0	0.8	3.1	0.0	0.4	1.1	79.8	1.5	0.4	0.4	4.2	1.5	0.8	0.4
lab	0.0	2.0	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	86.3	0.0	0.0	0.0	9.8	0.0	0.0
lecture_theatre	0.0	0.0	20.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	13.3	6.7	46.7	0.0	6.7	0.0	0.0	6.7
library	0.0	0.0	4.9	2.8	6.9	0.7	4.2	0.0	0.0	1.4	0.7	0.0	0.0	0.7	58.3	0.0	2.8	12.5	4.2
living_room	0.6	19.6	0.0	0.0	0.0	3.9	2.8	5.0	0.0	0.6	8.4	2.2	1.1	0.0	0.6	49.2	3.9	2.2	0.0
office	0.3	6.8	4.6	2.6	1.5	1.0	1.4	0.6	2.1	0.3	3.2	1.3	14.6	0.3	5.0	1.5	48.8	2.1	2.0
rest_space	0.7	1.0	7.6	0.2	2.0	1.5	9.7	0.3	5.6	2.2	0.2	0.3	0.5	0.7	8.7	1.9	3.9	50.8	2.2
study_space	0.0	7.7	7.7	0.0	0.0	0.0	7.7	0.0	0.0	7.7	0.0	0.0	0.0	7.7	30.8	7.7	0.0	0.0	23.1

Fig. 5. Confusion Matrix for SUN RGB-D dataset. The results are shown in percentage (%).

In addition, the proposed AGCN method has been evaluated on the SUN RGB-D dataset. The confusion matrix is shown in Figure 5. It can be observed that AGCN is capable of recognizing the bathroom, bedroom, furniture store, kitchen, office, and rest space, and is not able to recognize the study space and discussion area. Moreover, home office is misclassified as bedroom, lecture theater and conference rooom are misclassified as classroom.

Table V shows the results on the SUN RGB-D dataset, where our method reaches competitive performance compared with other SOTA methods without additional depth information, suggesting that our method is superior to others on various scene recognition datasets. In comparison with these methods, the IOD [26] utilizes the object detection results as an additional branch to boost the scene recognition performance. The MAPNet [55], G+L+SOOR [56], TRecgNet Aug [57], and ASK [58] are using the RGB-D information and additional branch to represent depth information, which is more complex.

4) Network Performance Analysis: From Table VI, the AGCN with 20 nodes reaches best results at Places365-7 dataset. Generally speaking, increasing the number of nodes is helpful for the performance improvements of visual scene recognition. The BORM, which utilizes the semantic segmentation methods, requires strong priors of the semantic information, as well as the statistical analysis of datasets for training. Our method is end-to-end training, which is more convenient for training, and has fewer parameters and calculations for both network training and inference.

TABLE I
INTRODUCTION OF THE DATASETS USED IN THIS ARTICLE.

Databases	Description	Total Scenarios	Training	Testing
Places365-7 [25]	Indoor scene recognition dataset with 7 categories.	35701	35000	701
Places365-14 [25]	Indoor scene recognition dataset with 14 categories	76500	75000	1500
SUN RGB-D [50]	RGB-D Scene Recognition datasets with 19 categories.	9504	4845	4659
ESC-10 [51]	Environmental Sound Classification dataset with 10 categories	400	320	80
ESC-50 [51]	Environmental Sound Classification dataset with 50 categories	2000	1600	400
ADVANCE [36]	Audio-Visual Scene Recognition Dataset	2696	2147	549
TAU Audio-Visual Scenes 2021 dataset [37]	Aerial Audio-Visual Scene Recognition Dataset	12291	8646	3645

TABLE II
NETWORK DETAILS OF AGCN FOR VISUAL SCENE RECOGNITION. OMIT THE BATCH SIZE.

Layer	Input/Input Size	Output Size
Conv1	3x224x224	64x112x112
Res-2	64x112x112	256x112x112
Res-3	256x112x112	512x56x56
Res-4	512x56x56	1024x28x28
Res-5	1024x28x28	2048x14x14
fc	2048x14x14	2048
AFM	Res-4, Res-5	1024x28x28
GCN	AFM	#Node*256
classification fc	GCN, fc	#Class

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE REDUCED PLACES365-7 DATASET

Method	Config	Semantic Head	Acc
PlacesCNN [52]	ResNet18	✗	80.4
	ResNet50	✗	82.7
Deduce [53]	Φ_{obj} (OM)	✓	62.6
	Φ_{scene}	✗	87.3
	$\Phi_{comb.}$	✓	88.1
BORM [25]	IOM	✓	82.4
	BORM	✓	83.1
	CBORM	✓	90.1
OTS [26]	OTS	✓	90.1
Ours	AGCN	✗	**91.7**

TABLE IV
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE REDUCED PLACES365-14 DATASET OF SCENE RECOGNITION ACCURACY, THE * INDICATES THE RE-IMPLEMENT OF THE METHOD.

Method	Config	Semantic Head	Acc
PlacesCNN [52]	ResNet18	✗	76.0
	ResNet50	✗	80.0
Word2Vec [54]	ResNet50+Word2Vec	✓	83.7
*Deduce [53]	Φ_{obj} (OM)	✓	47.0
BORM [25]	IOM	✓	74.1
	BORM	✓	74.9
	CBORM	✓	85.8
OTS [26]	OTS	✓	85.9
Ours	AGCN	✗	**86.0**

TABLE V
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE SUN RGB-D DATASET OF RECOGNITION ACCURACY

Methods	RGB	Depth	RGB-D
CNN+SVM [59]	37.0	-	41.5
MCAF [60]	40.4	36.3	48.1
MSMM [61]	41.5	40.1	52.3
RGB-D-CNN [62]	42.7	42.4	52.4
DF^2Net [63]	-	-	54.6
GAN [64]	42.6	53.3	53.3
ACM Graph [32]	45.7	-	55.1
G+L+SOOR [56]	50.5	44.1	55.5
MAPNet [55]	-	-	56.2
MSN [65]	-	-	56.2
ASK [58]	-	-	57.3
TRecgNet Aug [57]	53.8	49.3	58.5
IOD [47]	58.2	-	58.2
AGCN	**58.7**	-	-

TABLE VI
COMPARISON OF THE PROPOSED AGCN ALGORITHM WITH DIFFERENT NUMBER OF NODES ON PLACES365-7 DATASET.

Model	Places365-7
AGCN (8 nodes)	91.3
AGCN (12 nodes)	91.4
AGCN (16 nodes)	91.0
AGCN (20 nodes)	**91.7**
AGCN (24 nodes)	91.3

As Table VII shows, the proposed AGCN is an end-to-end training network with a much smaller number of parameters compared with the semantic segmentation-based BORM methods. BORM has 6x parameters and 3x GFlops compared with AGCN. In comparison with OTS, it has 3x parameters and similar GFlops. These results demonstrated our proposed methods feature fewer parameters.

5) Visualization of Visual Scene Graph: We show the visualization of automatically constructed SVG and CVG, as

TABLE VII
PARAMS COMPARISON

Model	Params	GFlops
BORM [25]	226.89	138.011
Semantic Head	253.30	46.724
Total	480.19	184.735
OTS [26]	3.50	0.214
Semantic Head	253.30	46.724
Total	256.80	**46.938**
AGCN (8 nodes)	**79.17**	54.729
AGCN (12 nodes)	81.27	54.733
AGCN (16 nodes)	83.37	54.738
AGCN (20 nodes)	85.46	54.742
AGCN (24 nodes)	87.56	54.746

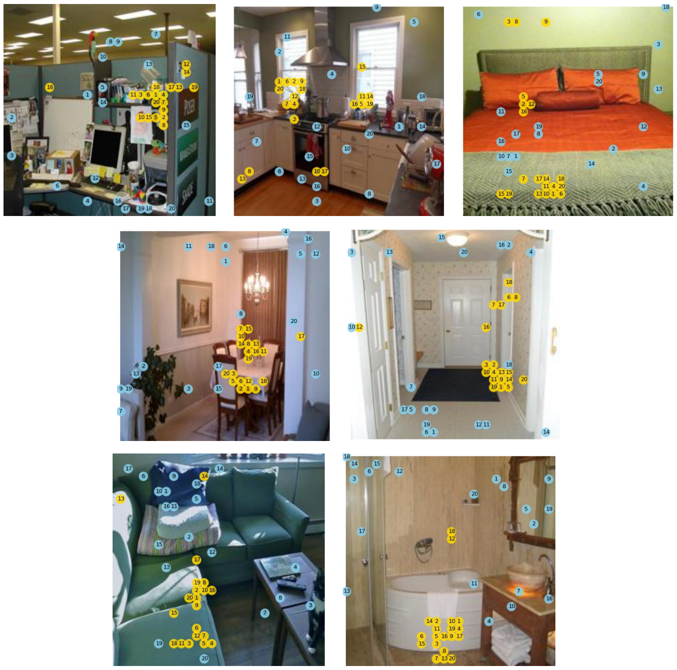


Fig. 6. The scene graph visualization samples for test set of Places365-7 with 20 nodes configuration. The first row are office and kitchen, the second row are bedroom and dining room, the third row are corridor and living room. The last row is bathroom. The gold circle represent the nodes of salient acoustic graph while blue circle represent the nodes of contextual acoustic graph.

shown in Figure 6. The visualization analysis of the graphs shows how our method focuses on the important regions and contextual information of the scene, respectively. It is worth noticing that the selected SVG concentrates more on the crucial areas, while the CVG concentrates on the contextual information, leaving out the less relevant information for scene recognition.

For example, in the second row of Figure 6, for the left one, an image of a bedroom, the SVG is mainly focused on the bed, especially the area of sheets and pillow, while fewer nodes are on the wall. The selected CVG is scattered around the entire image to evenly capture more detailed information on sheets, pillows, and walls. Similarly, for the right one, a picture of the Kitchen, the selected SVG is focused more on the dining table area while paying less attention to the wall and floor. At the same time, the CVG is located around each area evenly for detail preservation.

We have visualized the scene graph during the training process. It is worth mentioning that as the network training converges, a better distribution of scene graphs is obtained. The scene graph is better at mimicking the spatial layout of the scene objects or important regions of the scene. The visualization of the training process is shown in Figure 7. The first row shows the visualization of a bathroom with training epochs increases. It is worth noticing that the SVG focuses on several important elements of a bedroom like a bathtub, washbasin, walls, and floors instead of solely focusing on the bathtub. The second row shows the visualization of an office. In the beginning, the SVG is located in less essential regions like the area of a wooden panel of a desk. Later on, the SVG is moved to the working area near the printing machine, computer screen, and office chair, showing the excellent adaptivity of the proposed algorithm during training. The third row is a dining room, where the SVG focuses on the region of the dining table, and the CVG is scatted around the dining room to capture more detailed surrounding information of the dining room as additional cues. The last row is a living room. At first, the SVG is mainly located around the big glass area and floor and gradually moves the attention to the sofa, coffee table, and floor, showing the superior learning ability to find essential regions of a scene.

TABLE VIII
NETWORK DETAILS OF AGCN FOR ACOUSTIC SCENE RECOGNITION. OMIT THE BATCH SIZE.

Layer	Input/Input Size	Output Size
conv1	1x201x64	64x101x32
Res-2	64x101x32	256x101x32
Res-3	256x101x32	512x51x16
Res-4	512x51x16	1024x26x8
Res-5	1024x26x8	2048x13x4
fc	2048x13x4	2048
AFM	Res-4, Res-5	1024x26x8
GCN	AFM	#Node*256
classification fc	GCN, fc	#Class

B. Acoustic Scene Classification

1) Datasets: We use Environmental Sound Classification (ESC) dataset that contains 50 acoustic scene classes for evaluation [51]. The official dataset splition settings ESC-10 and ESC-50 are utilized, where the 5-fold cross validation is applied. The ESC-50 dataset contains 2000 labeled environmental sounds with a duration of 5 seconds, with each categories are equally distributed. The ESC-50 is split as training set with 1600 samples and test set with 400 samples. The ESC-10 is a subsect of ESC-50 that contains 320 samples as training set, and 80 samples as test set.

2) Implementation Details: The raw videos are resampled to 16.0kHz and fixed to the certain length, 5s for ESC-10 and ESC-50. Then, to analyse the signal in frequency domain, the short time Fourier transform (STFT) is applied on the audio signals, where the spectrograms is obtained with a hop length of 400 and window length of 1024. Last, to obtain the Log-Mel spectrograms, the 64 mel filter banks are applied on the previous spectrograms and followed by a logarithmic operation. During the training stage, we adopts the ResNet50 pretrained on the AudioSet for sound information extraction [66]. The proposed model is implemented in Pytorch [48]. The SGD optimizer is used for training in experiments. The initial learning rate is 0.01 and decreased by 10 at every 20 epochs. The momentum of optimizer is 0.9. The total train epochs is 60. The network structure in details are listed on the Table VIII. The input feature of sound signal after Fourier transform is of shape 1x201x64. Therefore, we adjust the input channel of ResNet50 from 3 to 1.

3) Comparison with stat-of-the-arts: Our method is evaluated on two ESC datasets, ESC-10 and ESC-50, which are the most commonly used datasets for ESC, where the Log-Mel spectrograms are extracted from the audio signals as the

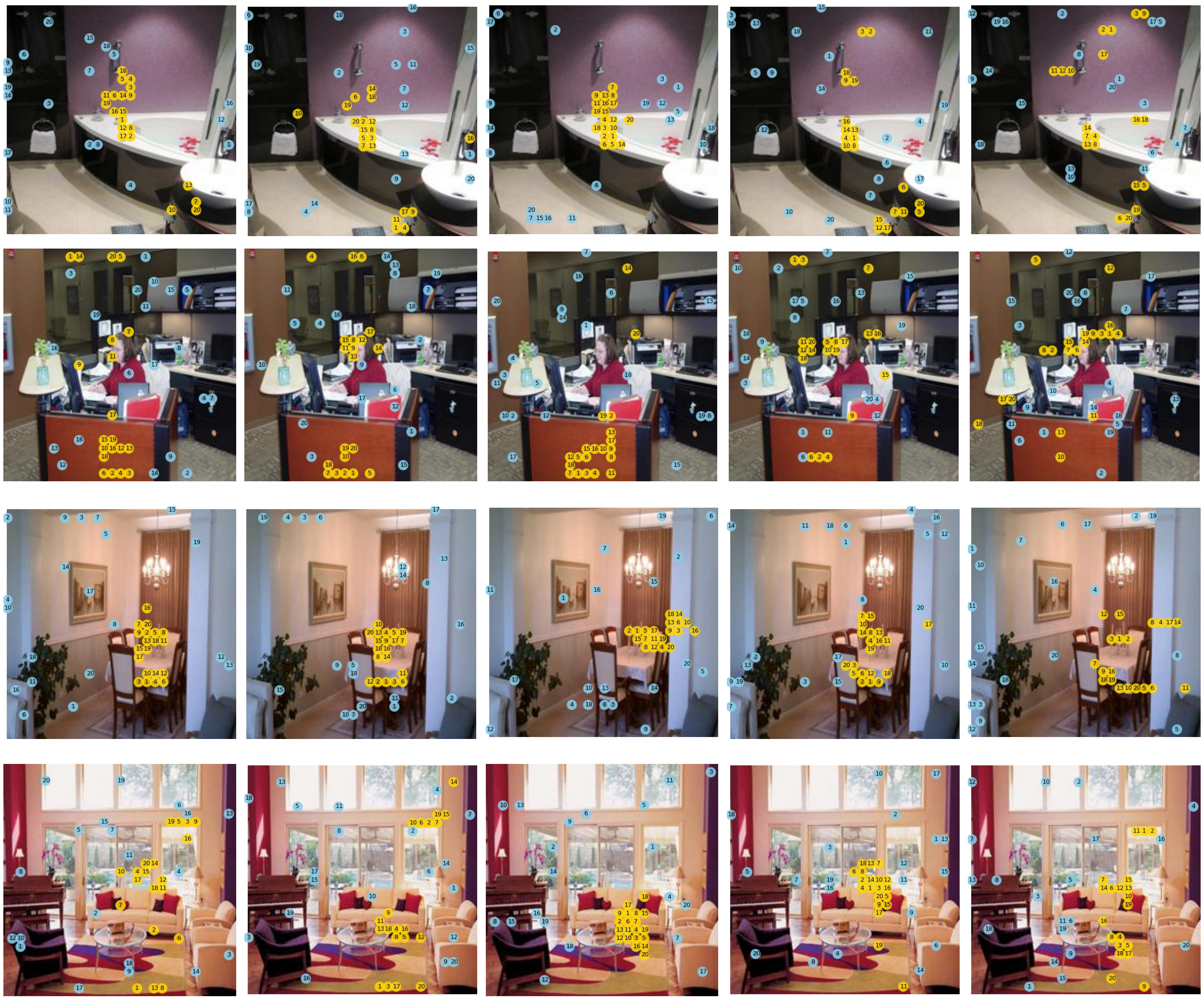


Fig. 7. The scene graph visualization of Places365-7 under different training epoch number. From first row to last row are bathroom, office, dining room and living room, respectively. As the training epochs increases, the salient acoustic graph is more concentrated on the important region of interests of the visual scene, while the contextual acoustic graph is scatted around the entire images for details preservation.

TABLE IX
COMPARISON OF THE AGCN AND EXISTING STATE-OF-THE-ART METHODS ON THE ESC-10, ESC-50 DATASETS. WE PERFORM 5-FOLD CROSS-VALIDATION FOLLOWED BY OFFICIAL FOLD SETTINGS. THE AVERAGE ACCURACY OF 5-FOLD CROSS-VALIDATION ARE REPORTED.

Model	ESC-10	ESC-50
KNN [51]	66.7	32.3
SVM [51]	67.5	39.6
Random Forest [51]	72.7	44.3
AlexNet [67]	78.4	78.7
Google Net [67]	63.2	67.8
PiczakCNN [51]	80.5	64.9
SoundNet [68]	93.7	79.1
Multi-Stream CNN [21]	94.2	84.0
MelFB+LGTFB+EN-CNN [69]	93.7	88.1
Attention Network[21]	94.2	84.0
Human[20]	95.7	81.3
TS-CNN10 [22]	95.8	88.6
ARNN [70]	92.0	-
DA-KL [71]	92.5	-
ACRNN [72]	93.7	86.1
AGCN (Ours)	**98.0**	**90.8**

TABLE X
COMPARISON OF THE PROPOSED AGCN WITH DIFFERENT NUMBER OF NODES

Model	ESC-10	ESC-50
AGCN (8 nodes)	96.3	88.3
AGCN (12 nodes)	97.5	90.5
AGCN (16 nodes)	97.5	89.6
AGCN (20 nodes)	97.0	91.2
AGCN (24 nodes)	**98.5**	**90.8**

input of the AGCN. We compare our proposed model with existing state-of-the-art methods. Table IX demonstrates the performance of AGCN on the ESC-10 and ESC-50 datasets. The proposed AGCN reaches state-of-the-art performance on both datasets. In the ESC-10 dataset, our method reaches 98.5% classification accuracy, which is higher than TS-CNN10 about 2.7% accuracy. The detailed confusion matrix of ESC-10 results are shown in Figure 10. In the ESC-50 dataset, our method reaches 90.8% classification accuracy, which is 2.2% higher than TS-CNN10.

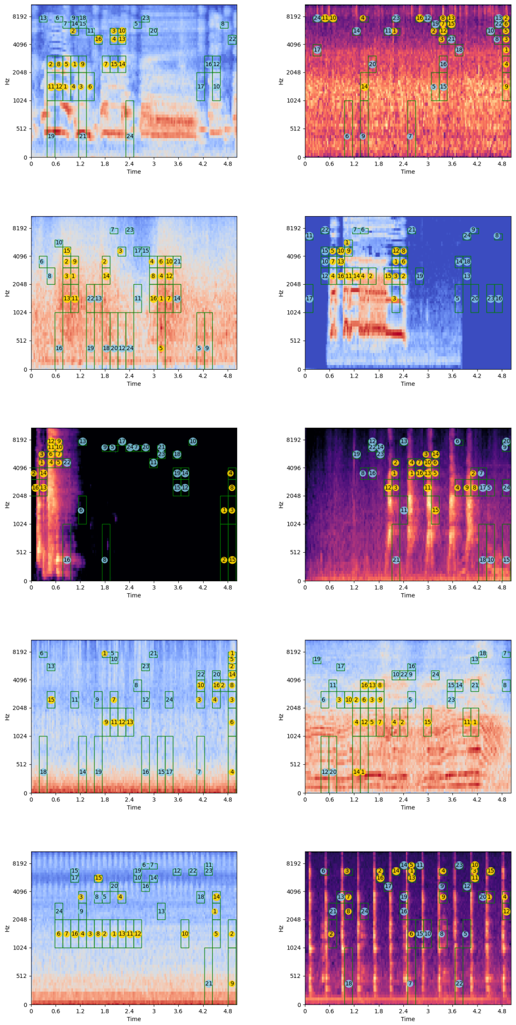


Fig. 8. Acoustic graphs visualization samples of ESC-10 dataset with 20 nodes configuration. From left to right column, first row to last row order, the acoustic scene classes are crying baby, rain, sea waves, roster, sneezing, dog, crackling fire, chainsaw, helicopter, and clock tick, respectively. The gold circle represents nodes of SAG, and blue circle represents the nodes of CAG.

4) Effective of Node Number: From the Table X, we observed that the influence of nodes numbers to the recognition accuracy is near linearly correlated. The best accuracy is with the 24 number of nodes, showing the **98.0%** and **90.8%** classification accuracy on the ESC-10, and ESC-50 datasets, respectively.

5) Visualization of Acoustic Graphs: We show the visualization of automatically constructed SAG and CAG, as shown in Figure 8. The visualization analysis of SAG shows how our method focuses on the crucial regions of sound signal spectrograms. Besides, it shows CAG captures the detailed patterns of sound signal spectrograms. AGCN well exploited the sound signal textures and frequency bands textures and abandoned the less relevant information for acoustic representation. For instance, in the third row, the left column shows the spectrograms of a sneezing sound spectrogram, where the SAG focuses on the essential region with a relatively strong signal, and the CAG is distributed over the entire signal frequency bands to preserve the details of acoustic signal information. The right column shows a sound signal spectrogram of the dog. It can be seen that the SAG makes use of the most discriminative signal bands while the CAG utilizes the signals of different frequencies, and strengths for a details representation of the acoustic scene.

We have visualized the acoustic graphs during the training process. It is obvious that as the network training converges, optimized distribution of acoustic graphs is obtained. The visualization of the training process is shown in Figure 9. The first three rows are samples of the ESC-10 dataset. The first row shows an auditory scene of the dog. In the first column, the nodes of the SAG are diversified between the contextual regions and salient regions of signal, then rapidly converge to the salient regions at the following epochs. The CAG is scatted around the entire sound signal but then filters out a small portion of the sound signal between 0-1 second, only to preserve the more useful information. The second row shows an acoustic scene of sea waves. At the early stage, like the first two columns, the SAG is diverged and distributed around the entire spectrograms and then gradually concerted on two regions, showing a convergence ability. The CAG is scatted around the frequency bands to capture the details of each frequency level, ensuring the signal is well exploited. The third row shows the sneezing scene. The SAG moves from one place to another in the first two columns, gradually focusing on the most discriminative signal regions and forming a distinctive SAG. The last row is the sheep scene from the ESC-50 dataset, viewed as an unseen acoustic scene. In the beginning, there are two locations for the SAG, which are not representative. Then, the SAG is gradually converging to the most apparent regions. The CAG is distributed on the high-frequency bands at the beginning and then spread over the entire frequency bands for detail preservation.

C. Audio-Visual Scene Classification

1) Datasets: TAU Audio-Visual Scenes 2021 dataset is used to test the effectiveness of the proposed AGCN. Both the audio and video samples have a duration of 10 seconds. Based on the official settings, there are 86K samples in the training set, and 36K samples in the test set. The audio processing is the same as the previous environmental sound classification. To ensure the efficiency of AGCN, we only sample one image from each video as the input for visual modalities.

ADVANCE dataset is an aerial audio-visual scene recognition datasets.

2) Results: Figure 11 shows the comparison between the baseline methods, which shows a substantial improvement has been achieved. Besides, we compared with the OpenL3-based methods, where we achieved a relative improvement in VSC and AVSC tasks. Figure 12 shows the per-class accuracy of AGCN on three tasks, compared with OpenL3 based method [37], our proposed method has better performance on the VSC task and AVSC tasks. On the VSC and AVSC tasks, we achieved 89.5 and 91.6 percent accuracy, respectively. The relative improvement of our proposed AGCN over the OpenL3-based method on the VSC task is about 31.1%, which is tremendously surprising. Overall, on the AVSC task, our

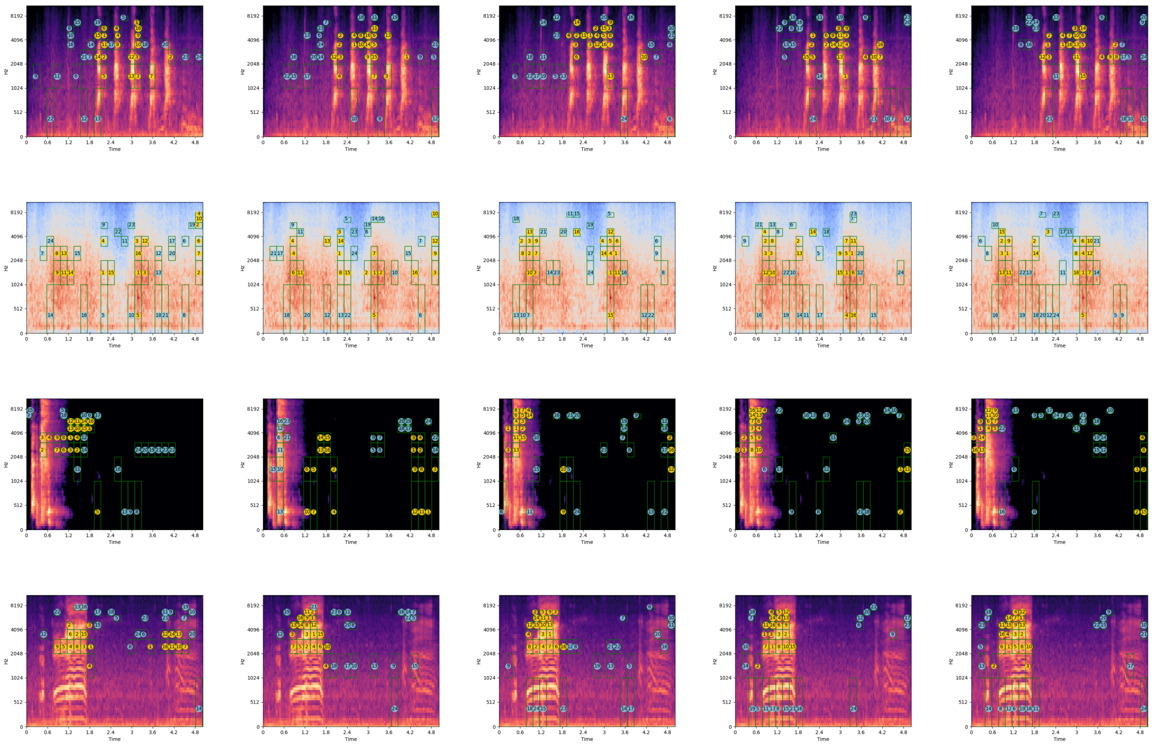


Fig. 9. The visualization of acoustic graphs under different training epochs. The first three rows are roster, sea waves, and sneezing of ESC-10 dataset. The last row is from ESC-50, which can be viewed as unseen acoustic signal. From first row to last row are rain, roster, sea waves, and helicopter, respectively. As the training epoch increases, the salient acoustic graph is more concentrated on the important regions of the acoustic scene, while the contextual acoustic graph is evenly distributed over the entire frequency bands for details preservation.

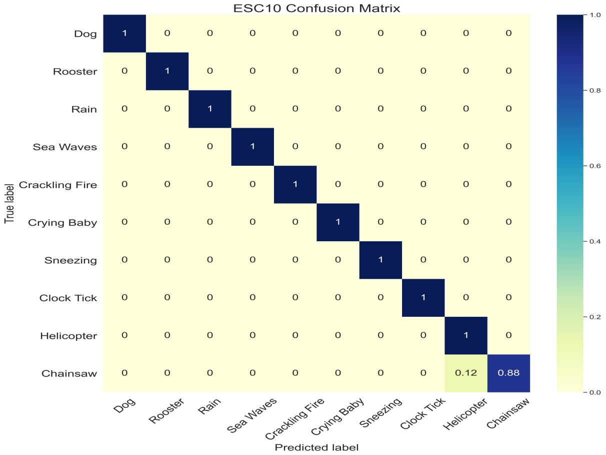


Fig. 10. Confusion Matrix for ESC-10 dataset. The results are shown in percentage (%/100).

relative improvement over the OpenL3-based method is about 8.0%, which suggests a substantial improvement is achieved by the AGCN. These results suggest the proposed AGCN is suitable for the AVSC task.

V. Conclusion

In this paper, we propose an AGCN network for audio-visual scene representation that exploits the SAG/SVG and CAG/CVG for audio-visual scene classification. We explicitly construct the SAG/SVG that focuses on essential regions of audio-visual scene inputs with the most distinctive nodes selected from feature maps. Meanwhile, to utilize the meaningful background contextual information for better audio-visual signal modeling, we construct CAG/CVG with average nodes selected from fused feature attention maps. In addition, the spatial layouts of scene nodes are preserved with an adjacency matrix. Extensive experiments have been conducted on ASC, VSC, and AVSC tasks. These results showed the

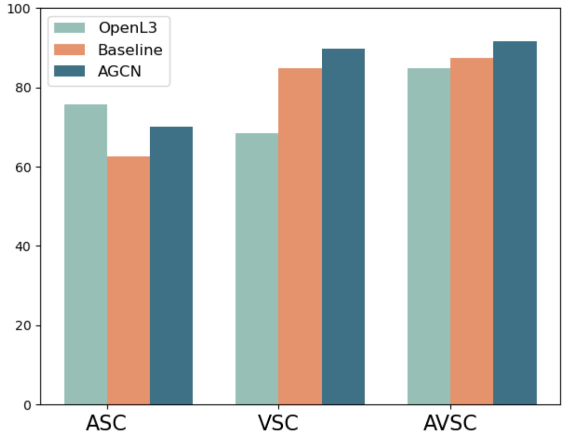


Fig. 11. The performance of OpenL3 [37] based method and AGCN on the ASC, VSC, and AVSC tasks.

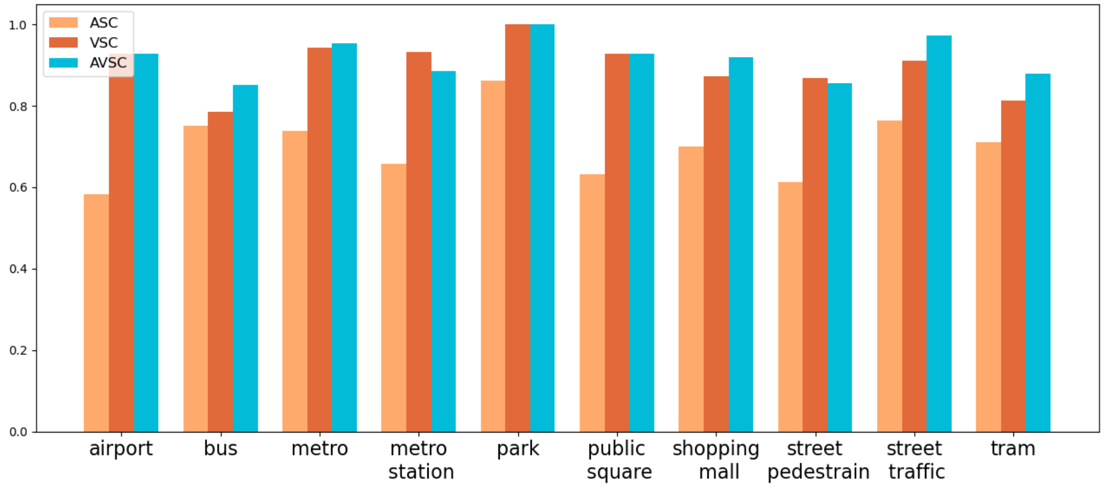


Fig. 12. The per-class accuracy performances of AGCN on the TAU Audio-Visual Scenes 2021 dataset, the ASC, VSC, and AVSC tasks are reported.

TABLE XI
THE UNIMODAL AERIAL SCENE RECOGNITION RESULTS ON THE ADVANCE DATASET WITH 5 DIFFERENT EXPERIMENTS.

Modality	Metrics	CTTA [36]	AGCN (Ours)
Sound	Precision	31.14 ± 0.30	**48.09 ± 2.22**
	Recall	33.80 ± 1.03	**53.55 ± 3.28**
	F1-score	29.66 ± 0.13	**50.40 ± 2.52**
Visual	Precision	73.97 ± 0.39	**89.88 ± 0.63**
	Recall	73.47 ± 0.52	**89.47 ± 0.69**
	F1-score	73.44 ± 0.45	**89.34 ± 0.68**

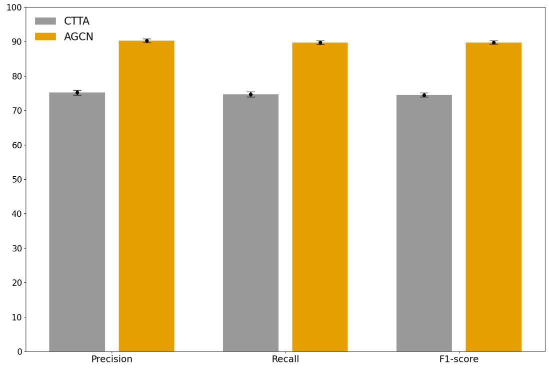


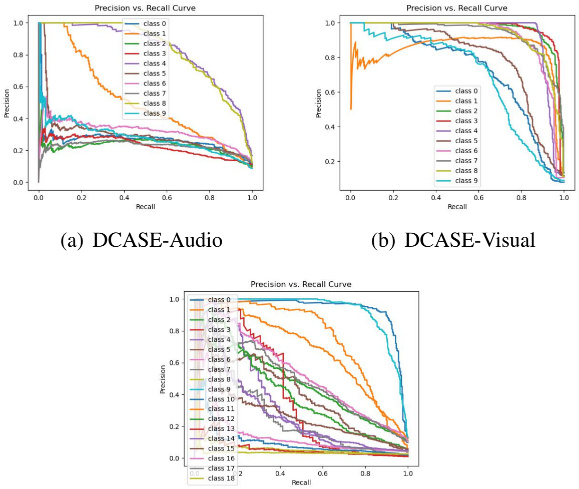
Fig. 13. The comparison between AGCN and CTTA on the AVSC task with ADVANCE dataset.

superiority of AGCN over the previous CNN-based methods. Moreover, the visualized graphs show the proposed SAG/SVG and CAG/CVG are semantic meaningful.

REFERENCES

[1] J. Ren, X. Jiang, J. Yuan, and N. Magnenat-Thalmann, "Sound-event classification using robust texture features for robot hearing," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 447–458, 2016.

[2] M. Basiri, F. Schill, P. Lima, and D. Floreano, "On-board relative bearing estimation for teams of drones using sound," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 820–827, 2016.



(a) DCASE-Audio (b) DCASE-Visual

(c) SUN RGB-D

Fig. 14. The comparison between audio and visual modality in various datasets.

[3] B. Chen, C. Chen, and J. Wang, "Smart homecare surveillance system: Behavior identification based on state-transition support vector machines and sound directivity pattern analysis," *and Cybernetics: Systems IEEE Transactions on Systems, Man*, vol. 43, no. 6, pp. 1279–1289, 2013.

[4] Y. Zhuang, N. Jiang, H. Hu, and F. Yan, "3-d-laser-based scene measurement and place recognition for mobile robots in dynamic indoor environments," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 2, pp. 438–450, 2012.

[5] X. Yang, F. Li, L. Li, K. Gu, and H. Liu, "Study of natural scene categories in measurement of perceived image quality," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–12, 2022.

[6] H. M. Do, K. C. Welch, and W. Sheng, "Soham: A sound-based human activity monitoring framework for home service robots," *IEEE Transactions on Automation Science and Engineering*, 2021.

[7] H. M. Do, M. Pham, W. Sheng, D. Yang, and M. Liu, "Rish: A robot-integrated smart home for elderly care," *Robotics and Autonomous Systems*, vol. 101, pp. 74–92, 2018.

[8] E. Akbal and T. Tuncer, "A learning model for automated construction site monitoring using ambient sounds," *Automation in Construction*, vol. 134, p. 104094, 2022.

[9] E. Martinson and A. Schultz, "Robotic discovery of the auditory scene," in *Proceedings 2007 IEEE International Conference on Robotics and*

Automation. IEEE, 2007, pp. 435–440.

[10] J. Ni, K. Shen, Y. Chen, W. Cao, and S. X. Yang, "An improved deep network-based scene classification method for self-driving cars," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[11] S. Li, F. Li, S. Tang, and F. Luo, "Heart sounds classification based on feature fusion using lightweight neural networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–9, 2021.

[12] O. K. Toffa and M. Mignotte, "Environmental sound classification using local binary pattern and audio features collaboration," *IEEE Transactions on Multimedia*, p. 1, 2020.

[13] Xinyu Wu, Haitao Gong, Pei Chen, Zhong Zhi, and Yangsheng Xu, "Intelligent household surveillance robot," in *Proc. IEEE Int. Conf. Robotics and Biomimetics*, 2009, pp. 1734–1739.

[14] X. Valero and F. Alías, "Gammatone wavelet features for sound classification in surveillance applications," in *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. IEEE, 2012, pp. 1658–1662.

[15] A. Meintjes, A. Lowe, and M. Legget, "Fundamental heart sound classification using the continuous wavelet transform and convolutional neural networks," in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 409–412.

[16] X. Wu, H. Gong, P. Chen, Z. Zhong, and Y. Xu, "Surveillance robot utilizing video and audio information," *Journal of Intelligent & Robotic Systems*, vol. 55, no. 4-5, pp. 403–421, 2009.

[17] S. Wang, T. Heittola, A. Mesaros, and T. Virtanen, "Audio-visual scene classification: analysis of dcase 2021 challenge submissions," *arXiv preprint arXiv:2105.13675*, 2021.

[18] A. Tripathi, D. Baruah, and R. D. Baruah, "Acoustic event classification using ensemble of one-class classifiers for monitoring application," in *2015 IEEE Symposium Series on Computational Intelligence*. IEEE, 2015, pp. 1681–1686.

[19] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of audio signals using aann and gmm," *Applied soft computing*, vol. 11, no. 1, pp. 716–723, 2011.

[20] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2015, pp. 1–6.

[21] X. Li, V. Chebiyyam, and K. Kirchhoff, "Multi-Stream Network with Temporal Attention for Environmental Sound Classification," in *Proc. Interspeech 2019*, 2019, pp. 3604–3608. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-3019

[22] H. Wang, Y. Zou, D. Chong, and W. Wang, "Environmental Sound Classification with Parallel Temporal-Spectral Attention," in *Proc. Interspeech 2020*, 2020, pp. 821–825. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1219

[23] P. Espinace, T. Kollar, A. Soto, and N. Roy, "Indoor scene recognition through object detection," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 1406–1413.

[24] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Objects as attributes for scene classification," in *European conference on computer vision*. Springer, 2010, pp. 57–69.

[25] L. Zhou, C. Jun, X. Wang, Z. Sun, T. L. Lam, and Y. Xu, "Borm: Bayesian object relation model for indoor scene recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2021.

[26] B. Miao, L. Zhou, A. S. Mian, T. L. Lam, and Y. Xu, "Object-to-scene: Learning to transfer object knowledge to indoor scene recognition," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2069–2075.

[27] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, 2007, pp. 197–206.

[28] X. Meng, Z. Wang, and L. Wu, "Building global image features for scene recognition," *Pattern recognition*, vol. 45, no. 1, pp. 373–380, 2012.

[29] K. Van De Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1582–1596, 2009.

[30] L. Wang, S. Guo, W. Huang, and Y. Qiao, "Places205-vggnet models for scene recognition," *arXiv preprint arXiv:1508.01667*, 2015.

[31] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[32] Y. Yuan, Z. Xiong, and Q. Wang, "Acm: Adaptive cross-modal graph convolutional neural networks for rgb-d scene recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9176–9184.

[33] N. Khan, U. Chaudhuri, B. Banerjee, and S. Chaudhuri, "Graph convolutional network for multi-label vhr remote sensing scene recognition," *Neurocomputing*, vol. 357, pp. 36–46, 2019.

[34] H. Zeng, X. Song, G. Chen, and S. Jiang, "Amorphous region context modeling for scene recognition," *IEEE Transactions on Multimedia*, 2020.

[35] Q. Song, K. Mei, and R. Huang, "Attanet: Attention-augmented network for fast and accurate scene parsing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 2567–2575.

[36] D. Hu, X. Li, L. Mou, P. Jin, D. Chen, L. Jing, X. Zhu, and D. Dou, "Cross-task transfer for geotagged audiovisual aerial scene recognition," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer, 2020, pp. 68–84.

[37] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 626–630.

[38] Q. Yu, Y. Yao, L. Wang, H. Tang, J. Dang, and K. C. Tan, "Robust environmental sound recognition with sparse key-point encoding and efficient multispike learning," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 2, pp. 625–638, 2020.

[39] L. Zhou, Y. Zhou, X. Qi, J. Hu, T. L. Lam, and Y. Xu, "Feature pyramid attention based residual neural network for environmental sound classification," *arXiv preprint arXiv:2205.14411*, 2022.

[40] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[41] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2017.

[42] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[43] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5177–5186.

[44] W. Nie, M. Ren, J. Nie, and S. Zhao, "C-gcn: Correlation based graph convolutional network for audio-video emotion recognition," *IEEE Transactions on Multimedia*, 2020.

[45] Y. Sun and S. Ghaffarzadegan, "An ontology-aware framework for audio event classification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 321–325.

[46] H. Wang, Y. Zou, D. Chong, and W. Wang, "Modeling label dependencies for audio tagging with graph convolutional network," *IEEE Signal Processing Letters*, vol. 27, pp. 1560–1564, 2020.

[47] R. Pereira, L. Garrote, T. Barros, A. Lopes, and U. J. Nunes, "A deep learning-based indoor scene classification approach enhanced with inter-object distance semantic features," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 32–38.

[48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[50] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

[51] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[52] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[53] A. Pal, C. Nieto-Granda, and H. I. Christensen, "Deduce: Diverse scene detection methods in unseen challenging environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4198–4204.

[54] B. X. Chen, R. Sahdev, D. Wu, X. Zhao, M. Papagelis, and J. K. Tsotsos, "Scene classification in indoor environments for robots using context based word embeddings," in *2018 IEEE International Conference on Robotics and Automation (ICRA) Workshop*, 2018.

[55] Y. Li, Z. Zhang, Y. Cheng, L. Wang, and T. Tan, "Mapnet: Multi-modal attentive pooling network for rgb-d indoor scene classification," *Pattern Recognition*, vol. 90, pp. 436–449, 2019.

[56] X. Song, S. Jiang, B. Wang, C. Chen, and G. Chen, "Image representations with spatial object-to-object relations for rgb-d scene recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 525–537, 2019.

[57] D. Du, L. Wang, Z. Li, and G. Wu, "Cross-modal pyramid translation for rgb-d scene recognition," *International Journal of Computer Vision*, vol. 129, no. 8, pp. 2309–2327, 2021.

[58] Z. Xiong, Y. Yuan, and Q. Wang, "Ask: Adaptively selecting key local features for rgb-d scene recognition," *IEEE Transactions on Image Processing*, vol. 30, pp. 2722–2733, 2021.

[59] H. Zhu, J.-B. Weibel, and S. Lu, "Discriminative multi-modal feature fusion for rgbd indoor scene recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2969–2976.

[60] A. Wang, J. Cai, J. Lu, and T.-J. Cham, "Modality and component aware feature fusion for rgb-d scene classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5995–6004.

[61] X. Song, S. Jiang, and L. Herranz, "Combining models from multiple sources for rgb-d scene recognition." in *IJCAI*, 2017, pp. 4523–4529.

[62] X. Song, L. Herranz, and S. Jiang, "Depth cnns for rgb-d scene recognition: Learning from scratch better than transferring from rgb-cnns," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[63] Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan, "Df 2 net: Discriminative feature learning and fusion network for rgb-d indoor scene classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[64] D. Du, X. Xu, T. Ren, and G. Wu, "Depth images could tell us more: Enhancing depth discriminability for rgb-d scene recognition," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[65] Z. Xiong, Y. Yuan, and Q. Wang, "Msn: Modality separation networks for rgb-d scene recognition," *Neurocomputing*, vol. 373, pp. 81–89, 2020.

[66] Y. Wang, "Polyphonic sound event detection with weak labeling," *PhD thesis*, 2018.

[67] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Procedia computer science*, vol. 112, pp. 2048–2056, 2017.

[68] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in neural information processing systems*, vol. 29, pp. 892–900, 2016.

[69] H. Park and C. D. Yoo, "Cnn-based learnable gammatone filterbank and equal-loudness normalization for environmental sound classification," *IEEE Signal Processing Letters*, vol. 27, pp. 411–415, 2020.

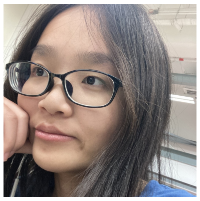
[70] A. M. Tripathi and A. Mishra, "Environment sound classification using an attention-based residual neural network," *Neurocomputing*, vol. 460, pp. 409–423, 2021.

[71] A. M. Tripathi and K. Paul, "Data augmentation guided knowledge distillation for environmental sound classification," *Neurocomputing*, vol. 489, pp. 59–77, 2022.

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

Liguang Zhou received the B.Eng. degree in Electrical Engineering from China Jiliang University, Hangzhou, China, in 2016, and he is pursing the Ph.D. degree in Computer Information Engineering at The Chinese of Hong Kong, Shenzhen, China.

His research interests include robotic vision, scene understanding, and computational photography.



Yuhongze Zhou received the B.Eng. degree in Digital Media Technology from Zhejiang University, Hangzhou, China, in 2020, and she is pursing the M.Sc. degree in Computer Science at McGill University, Montréal, Canada.

Her research interests include computational photography, machine learning, and computer graphics.

Xiaonan Qi received the B.Eng. degree in Computer Science and Engineering at the Chinese University of Hong Kong, Shenzhen, China, and she is pursuing the Master degree in Software Engineering at Carnegie Mellon University, Pittsburgh, PA, USA.

Her research interests include machine learning, signal processing, and computer graphics.

Junjie Hu (Member, IEEE) received the M.S. and Ph.D. degrees from the Graduate School of Information Science, Tohoku University, Sendai, Japan, in 2017 and 2020, respectively.

He is currently a Research Scientist with the Shenzhen Institute of Artificial Intelligence and Robotics for Society. His research interests include machine learning, computer vision, and robotics.

Tin Lun Lam (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in automation and computer-aided engineering from Chinese University of Hong Kong, Hong Kong, in 2006 and 2010, respectively. He is an Assistant Professor with the School of Science Engineering, The Chinese University of Hong Kong. His research interests include multi-robot systems, field robotics and human-robot interaction.

He has been granted over 60 patents, published 2 monographs, and over 60 international journal and conference papers. Most of them were published in top-tier international journals and conference proceedings in robotics and automation such as TRO, T-MECH, RA-L, JFR, ICRA, and IROS. Based on his research, he has received an IEEE/ASME T-MECH Best Paper Awardin 2011 and IROS Best Paper Award on Robot Mechanisms and Design in 2020. His research outcomes have also been reported by many internationally renowned media including Reuters, Discovery Channel, and IEEE Spectrum.

Yangsheng Xu (Fellow, IEEE) received the B.S.E. and M.S.E. degrees from Zhejiang University, Hangzhou, China, in 1982 and 1984, respectively, and the Ph.D. degree in 1989 from University of Pennsylvania, Philadelphia, PA, USA, all in robotics.

He has been a Professor with the Department of Automation and Computer-Aided Engineering, The Chinese University of Hong Kong, Hong Kong, since 1997. He is the President of The Chinese University of Hong Kong, Shenzhen. His research interests include robotics, intelligent systems, and electric vehicles.