

# Audio Event-Relational Graph Representation Learning for Acoustic Scene Classification

Yuanbo Hou , *Student Member, IEEE*, Siyang Song , Chuang Yu , *Member, IEEE*,  
Wenwu Wang , *Senior Member, IEEE*, and Dick Botteldooren , *Senior Member, IEEE*

**Abstract**—Most deep learning-based acoustic scene classification (ASC) approaches identify scenes based on acoustic features converted from audio clips containing mixed information entangled by polyphonic audio events (AEs). However, these approaches have difficulties in explaining what cues they use to identify scenes. This letter conducts the first study on disclosing the relationship between real-life acoustic scenes and semantic embeddings from the most relevant AEs. Specifically, we propose an event-relational graph representation learning (ERGL) framework for ASC to classify scenes, and simultaneously answer clearly and straightly which cues are used in classifying. In the event-relational graph, embeddings of each event are treated as nodes, while relationship cues derived from each pair of nodes are described by multi-dimensional edge features. Experiments on a real-life ASC dataset show that the proposed ERGL achieves competitive performance on ASC by learning embeddings of only a limited number of AEs. The results show the feasibility of recognizing diverse acoustic scenes based on the audio event-relational graph.

**Index Terms**—Acoustic scene classification, event-relational graph, multi-dimensional edge, graph representation learning.

## I. INTRODUCTION

ACOUSTIC scene classification (ASC) aims to classify an audio clip into a pre-defined semantic category, indicating the acoustic environment where the clip is captured (e.g., park, mall, or bus) [1]. ASC provides a broad description of the acoustic environment, which can assist intelligent agents in quickly understanding their surrounding environment. As a result, it is useful for various applications, such as sound source recognition [2], [3], [4], [5], [6], well-being assistance [7], [8], [9], and audio-visual scene recognition [10], [11], [12], [13], [14].

Manuscript received 3 August 2023; accepted 17 September 2023. Date of publication 26 September 2023; date of current version 10 October 2023. This work was supported by WAVES Research Group, Flemish Government through “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” programme. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Rosangela Coelho. (*Corresponding author: Yuanbo Hou.*)

Yuanbo Hou and Dick Botteldooren are with the WAVES Research Group, Ghent University, 9000 Ghent, Belgium (e-mail: yuanbo.hou@ugent.be; dick.botteldooren@ugent.be).

Siyang Song is with the School of Computing and Mathematical Science, University of Leicester, LE17RH Leicester, U.K. (e-mail: ss1535@leicester.ac.uk).

Chuang Yu is with the UCL Interaction Centre, University College London, WC1E6EA London, U.K. (e-mail: chuang.yu@ucl.ac.uk).

Wenwu Wang is with the Centre for Vision, Speech, and Signal Processing, University of Surrey, GU27XH Guildford, U.K. (e-mail: w.wang@surrey.ac.uk).

Visualizations of graph representations learned by ERGL are available here (<https://github.com/Yuanbo2020/ERGL>).

Digital Object Identifier 10.1109/LSP.2023.3319233

Typical deep learning-based ASC methods consist of three steps: 1) Converting a time-domain audio signal to a time-frequency spectrogram, which is used as input acoustic features; 2) Feeding these features to neural networks to obtain high-level representations; 3) Recognizing acoustic scenes based on such high-level representations. For example, Ren et al. [15] utilize a CNN-based model with mel features, where attention-based pooling layers are used to reduce the dimension of representations. The spatial pyramid pooling approach is used by CNN in [16] to provide various resolutions for ASC. Apart from mel-based features, wavelet-based deep scattering spectrum [17] is introduced for ASC. To explore the instance-level information of audio clips, multiple-instance learning [18] is used in ASC. The higher-order temporal information of acoustic features is exploited by convolutional recurrent neural networks (CRNN) with bidirectional recurrent layers [19] and with spatio-temporal attention pooling [20]. In addition, attentional graph convolutional networks are used for audio-visual scene classification [21]. Given the intrinsic relationship between acoustic scenes and audio events (AEs), some studies jointly analyze scenes and events based on multi-task learning (MTL) [22], [23]. Relation-guided ASC [24] is proposed to exploit the implicit relations between coarse-grained scenes and fine-grained AEs.

The features used in the methods above often contain polyphonic AEs information. These features capture both useful and irrelevant information, as well as noise, which are then used for ASC by recognition systems. However, it is difficult to explain what cues in audio clips are used by these approaches to recognize acoustic scenes (ASs). In real life, it is natural for humans to recognize ASs based on the semantically meaningful AEs contained in them, despite variations in relations among the occurring AEs in ASs [25]. This letter proposes audio event-relational graph representation learning (ERGL) to classify scenes and simultaneously clearly answer which cues are used in classifying. Inspired by the multi-dimensional edge learning in graph-based image analysis [26], [27], we introduce it into the proposed audio-based ERGL to enhance scene-dependent semantic relationships between non-graph AEs in end-to-end training. These scene-dependent event-relational graphs (ERGs) contain not only the activation of AEs (i.e. nodes in the graph) in the audio clip, but also their relations (represented as edges) that are relevant to ASC tasks. The graph in ERGs is a single graph. Thus, spatial-temporal graph neural networks [28], [29], which aim to capture spatial and temporal dependencies of graph sequences or multigraphs, are unsuitable for modeling ERGs. ERGs are fed to a gated graph convolutional network (Gated GCN) [30] for ASC. Experiments show that graph representations learned by ERGL, from only several explicit audio event

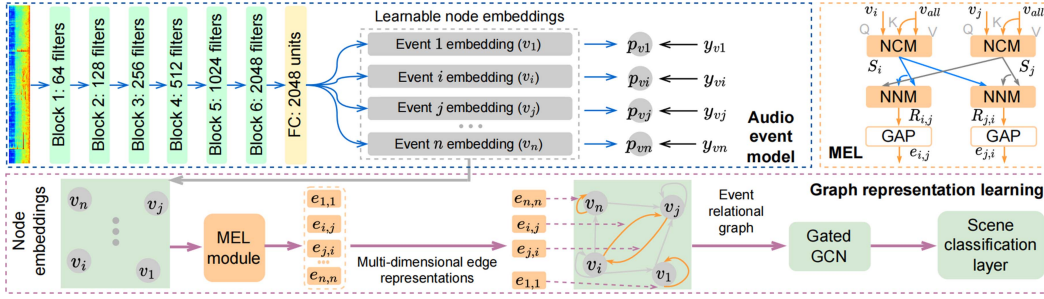


Fig. 1. Framework of the proposed scene-dependent event-relational graph learning (ERGL) for ASC.

semantic embeddings, can facilitate the discrimination between different ASs.

## II. AUDIO EVENT-RELATIONAL GRAPH LEARNING

This section discusses the proposed approach for learning the scene-related event-relational graph (ERG) representation from non-graph audio clips in an end-to-end manner. First, we learn a set of AEs embeddings, where each embedding ( $v_i$ ) contains the  $i$ -th audio event-related information, and is treated as the  $i$ -th node in the ERG. Then, we learn a pair of multi-dimensional edge features ( $e_{i,j}$ ,  $e_{j,i}$ ) to describe the relations between each pair of nodes ( $v_i$ ,  $v_j$ ). Thus, the obtained graph explicitly describes the occurrence of a set of AEs and their relations relevant to the given scene. Finally, the obtained ERG with  $n$  nodes and  $n \times n$  edges is fed into Gated GCN for ASC.

### A. Audio Event Node Embedding Learning

We first propose a model derived from PANNs [31] for node embedding generation, where each describes a specific AE. As shown in Fig. 1, the spectrogram of the audio clip is fed into a set of convolutional blocks. Each block contains two convolutional layers with kernels of size  $3 \times 3$ , a batch normalization [32], and a ReLU function [33]. Then, a fully-connected (FC) layer with 2048 units generates joint representations for all AEs. Unlike PANNs, we employ  $n$  independent FC layers with 64 units to learn  $n$  embeddings separately, where each embedding describes a unique pre-defined audio event.

During training, each event embedding  $v_i$  is fed into the following event classification layer to predict the corresponding event probability  $p_{vi}$ , where mean squared error (MSE) loss is used to measure the distance between the prediction  $p_{vi}$  and the label  $y_{vi}$  (i.e.,  $\mathcal{L}_{\text{event}} = \text{MSE}(p_{vi}, y_{vi})$ ). To train the audio event node embedding generation module, we employ PANNs, which contains 527 classes of AEs, to generate pseudo labels of AEs. This produces a 527-dimensional soft pseudo label  $y = [y_{v1}, y_{v2}, \dots, y_{v527}]$  for each audio clip, describing the occurrence probabilities of 527 classes of AEs. Since real-world acoustic scene datasets rarely have all 527 classes of AEs, i.e., the number of occurred AEs would be much smaller than 527, we rank all AEs by accumulating their probabilities in all training data, and use a set of top-ranked (Top  $n$ ) AEs with the highest overall probability describing each scene. As a result, each graph contains  $n$  nodes and  $n \times n$  edges.

### B. Audio Event-Relational Edge Feature Learning

Once all audio event (node) embeddings are obtained, we propose a multi-dimensional edge feature learning (MEL) module to learn scene-related relations between each pair of AEs. Here,

our hypothesis is that the co-occurrence patterns of all event pairs may include key clues for ASC tasks. The proposed MEL module in Fig. 1 consists of two sub-modules, the node-context relation modeling (NCM) and the node-node relation modeling (NNM). NCM first learns the ASC-task-specific relation cues between each node (event) and the global context (scene), generating a scene-aware representation to represent each node. Then, NNM models the semantic relations between nodes in each node pair, to generate the final multi-dimensional edge feature describing the AE-based scene-aware relations between each pair of nodes.

**NCM:** For each node  $v_i$ , NCM conducts cross-attention [34] between it and the global contextual representation  $v_{\text{all}}$  consisting of the mean of all node features, where node  $v_i$  is used as the query, and  $v_{\text{all}}$  is employed as the key and value.

$$\text{NCM}(\mathbf{Q}, \mathbf{K}) = \Phi(\mathbf{Q}\mathbf{W}_q(\mathbf{K}\mathbf{W}_k)^T / \sqrt{d_k})\mathbf{K}\mathbf{W}_v \quad (1)$$

where  $\Phi$  is the softmax function,  $\mathbf{W}_{\{q,k,v\}}$  are learnable weights, and  $d_k$  is a factor equal to the number of channels in  $\mathbf{K}$ . As a result, the obtained representations  $\mathcal{S}_i$  ( $i = 1, 2, \dots, n$ ) encode scene-aware cues for each audio event.

$$\mathcal{S}_i = \text{NCM}(v_i, v_{\text{all}}), \quad \mathcal{S}_j = \text{NCM}(v_j, v_{\text{all}}) \quad (2)$$

**NNM:** After extracting all scene-aware event node features, NNM module then models the semantic relationship between nodes by capturing multi-dimensional (m-d) edge features. In particular, NNM consists of cross-attention [34] and global average pooling (GAP) [19] layer, which takes a pair of scene-aware event features ( $\mathcal{S}_i, \mathcal{S}_j$ ) as input, and a pair of m-d edge features ( $e_{i,j}, e_{j,i}$ ) as output. In detail, NNM first conducts:

$$\mathcal{R}_{i,j} = \text{NNM}(\mathcal{S}_j, \mathcal{S}_i), \quad \mathcal{R}_{j,i} = \text{NNM}(\mathcal{S}_i, \mathcal{S}_j) \quad (3)$$

where the edge feature  $\mathcal{R}_{i,j}$  encodes  $\mathcal{S}_j$ -related cues in  $\mathcal{S}_i$ , and correspondingly,  $\mathcal{R}_{j,i}$  encodes  $\mathcal{S}_i$ -related cues in  $\mathcal{S}_j$ . Next, edge features  $\mathcal{R}_{i,j}$  and  $\mathcal{R}_{j,i}$  are fed into the GAP layer to obtain the multi-dimensional edge feature vectors  $e_{i,j}$  and  $e_{j,i}$ .

$$e_{i,j} = \text{GAP}(\mathcal{R}_{i,j}), \quad e_{j,i} = \text{GAP}(\mathcal{R}_{j,i}) \quad (4)$$

Consequently, the produced  $e_{i,j}$  and  $e_{j,i}$  capture multiple ASC-task-specific cues related to both event nodes  $v_i$  and  $v_j$ .

### C. Scene-Aware Event-Relational Graph

Once the ERG (denoted as  $G^0$ ) that contains  $n$  node embeddings  $v = \{v_1, v_2, \dots, v_n\}$  and  $n \times n$  multi-dimensional directed edge representations  $e = \{e_{1,1}, \dots, e_{i,j}, \dots, e_{n,n}\}$  is obtained, we feed  $G^0$  to the Gated GCN [30], [35] for ASC.

Since the model contains  $U$  GCN layers, its output is  $G^U = (v^U, e^U)$ , which is a graph with the same topology as  $G^0$ . The  $i$ -th node represents the activation state of the  $i$ -th event in the scene. The latent node features in  $G^U$  are concatenated as the scene representation and input to the final scene classification

TABLE I  
ACC OF ERGL AT DIFFERENT  $n$  VALUES ON THE VALIDATION SET

#	Top $n$	Acc (%)	#	Top $n$	Acc (%)
1	10	94.03 $\pm$ 2.41	5	100	94.22 $\pm$ 2.91
2	25	<b>95.92</b> $\pm$ 2.29	6	200	93.05 $\pm$ 2.55
3	50	94.65 $\pm$ 2.75	7	300	92.46 $\pm$ 3.89
4	75	94.47 $\pm$ 1.89	8	400	92.29 $\pm$ 2.07

layer. Cross entropy (CE) [36] is used as the loss function in ASC between the prediction  $p_s$  and the scene true label  $y_s$ ,  $\mathcal{L}_{\text{scene}} = CE(p_s, y_s)$ . Hence, the final loss of ERGL is  $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{event}} + \lambda_2 \mathcal{L}_{\text{scene}}$ , where  $\lambda_i$  ( $i = 1, 2$ ) default to 1 in this letter. The effect of  $U$ , which defaults to 2, on the model performance will be explored in the experiments later.

### III. EXPERIMENTS AND RESULTS

#### A. Dataset, Baseline, Experimental Setup, and Metric

This letter uses TUT Urban Acoustic Scenes 2018 dataset (UAS) [37] with 8640 10-second clips. The UAS contains 10 classes of real-life acoustic scenes, 24 hours in total. However, UAS does not provide labels for audio events. To obtain the event labels used in Section III-B, pre-trained model PANNs are used to annotate audio clips with 527 classes of AEs pseudo labels. To compare with other methods on the same test set, we follow the setup of [37], where the test, training, and validation sets contain 2518, 5509, and 613 samples, respectively.

In addition to a typical CNN-based approach [37] for ASC, this letter also employs attention-based [15], spatio-temporal attention [20], multiple-instance learning [18], and scene-event joint learning [22], [23], [24], [38] as baselines for comparison.

Following [31], the log mel spectrogram with 64 bins is used as the acoustic feature, which is extracted by the Short-Time Fourier Transform with a Hamming window of size 1024 and a hop size of 320 samples. A batch size of 64 and AdamW optimizer [39] with a learning rate of 1e-4 are used to minimize the loss. The systems are trained for 400 epochs. The accuracy (Acc) [1] is used as the performance metric.

#### B. Results and Analysis

The pseudo labels composed of probability outputs by PANNs [31] are used as supervision information for training the audio event model, as shown in Fig. 1. The accuracy of pseudo labels that do not involve human verification cannot be evaluated. Since our goal is ASC, we will mainly show ASC results, rather than the accuracy of pseudo labels on AEs.

*The number  $n$  of audio events:* We first evaluate the impact of the choice of  $n$ , i.e., the number of top AEs used in ERGL, on the performance of ERGL. The Acc of ERGL does not increase monotonically as  $n$  increases, as shown in Table I. The reason may be that as the number of events  $n$  increases, the number of nodes in the graph increases linearly, but the number of edges in the graph grows in the order of  $n^2$ , which sharply increases the burden for learning the multi-dimensional edge features with the MEL module. The increased number of parameters does not provide more useful information to the model, but may compromise its performance.

The ERGL works best when  $n = 25$ , indicating that only using 25 classes of AEs can describe the 10 classes of scenes in the dataset. In preprocessing the distribution of AEs in each scene, we also found that the 25 classes of AEs automatically

TABLE II  
ACC OF ERGL AT DIFFERENT  $U$  LAYERS ON THE VALIDATION SET

#	$U$	Acc (%)	#	$U$	Acc (%)
1	1	94.42 $\pm$ 1.97	4	4	95.09 $\pm$ 2.74
2	2	<b>95.92</b> $\pm$ 2.29	5	5	94.79 $\pm$ 2.43
3	3	95.66 $\pm$ 2.58	6	6	94.29 $\pm$ 1.92

TABLE III  
ABLATION STUDY OF AE RELATION-RELATED MEL IN ERGL

MEL	NCM	✗	✓	✗	✓
	NNM	✗	✗	✓	✓
Test set	Acc (%)	73.35 $\pm$ 1.98	74.42 $\pm$ 1.99	75.69 $\pm$ 1.54	<b>78.08</b> $\pm$ 2.06

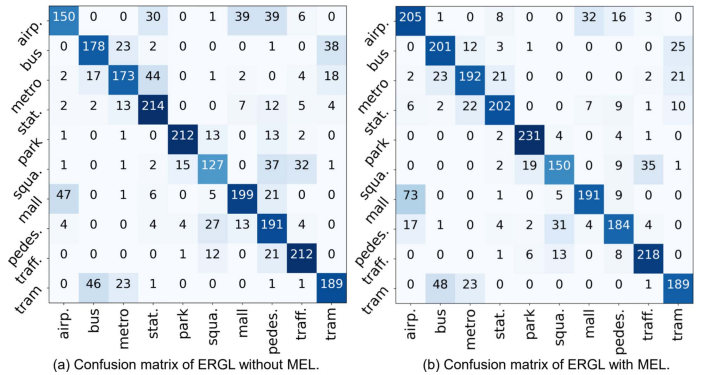


Fig. 2. Confusion matrix of ERGL w/o and w/ AE-based relational edges on the test set. (X-axis: Predicted label; Y-axis: True label).

selected by the model cover most of the dominant AEs in each scene. Therefore, we set  $n = 25$  in the following experiments.

*The number ( $U$ ) of GCN layers:* Table II explores the ERGL performance under different numbers of GCN layers. The results in Table II illustrate that increasing  $U$  does not lead to better results. The reason for this may be that the 2-layer Gated GCN already achieves a good balance between model performance and computational efficiency on the graph consisting of semantic embeddings of 25 classes of AEs, and also that adding extra layers would make the model deeper and harder to train. Subsequent experiments will set  $U$  as 2.

*Ablation study of AE-based relations in ERGL:* The MEL module in ERGL aids in capturing multi-dimensional AE-based semantic relations between nodes. In MEL, NCM learns the relation between the node and global contextual representation to represent the node, while NNM uses cross-attention to capture the semantic relations between nodes. To investigate how well NNM captures semantic relations, Table III presents the ablation study to compare the performance of ERGL with (w/) and without (w/o) AE-based relations. Table III shows that ERGL w/ NNM outperforms ERGL w/ NCM. For graph representation learning in ERGL, NNM, which aims to capture semantic relations between nodes, is more valuable than NCM, which focuses on learning relations between the node and global contextual representation. This shows that NNM based on cross-attention capturing AE-based semantic relations is effective. Fig. 2 shows confusion matrices of ERGL w/o and w/ AE-based relations to explore where the AE-based relational approach works and where it does not.

Fig. 2 illustrates that AE-based relational edges effectively help ERGL improve its accuracy in 6 scenes: “airp. (airport), bus, metro, park, squa. (public square), traff. (street traffic)”.



## REFERENCES

- [1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Berlin, Germany: Springer, 2018.
- [2] B. Defréville, F. Pachet, C. Rosin, and P. Roy, "Automatic recognition of urban sound sources," in *Proc. Audio Eng. Soc. Conv. 120*, 2006, pp. 1–9.
- [3] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Trans. Multimedia*, vol. 17, pp. 1733–1746, 2015.
- [4] L. Pham et al., "Lightweight deep neural networks for acoustic scene classification and an effective visualization for presenting sound scene contexts," *Appl. Acoust.*, vol. 211, pp. 109–489, 2023.
- [5] P. Foster, S. Sigtia, S. Krstulovic, J. Barker, and M. D. Plumbley, "Chime-home: A dataset for sound source recognition in a domestic environment," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2015, pp. 1–5.
- [6] Y. Hou et al., "Joint prediction of audio event and annoyance rating in an urban soundscape by hierarchical graph representation learning," in *Proc. INTERSPEECH*, 2023, pp. 68–72.
- [7] G. Grossi, R. Lanzarotti, P. Napolitano, N. Noceti, and F. Odone, "Positive technology for elderly well-being: A review," *Pattern Recognit. Lett.*, vol. 137, pp. 61–70, 2020.
- [8] G. Y. Kim, S.-S. Shin, J. Y. Kim, and H.-G. Kim, "Sound event detection and haptic vibration based home monitoring assistant system for the deaf and hard-of-hearing," in *Proc. Workshop Multimedia Accessible Hum. Comput. Interface*, 2018, pp. 1–7.
- [9] V. Carrasco et al., "Application of deep learning to enforce environmental noise regulation in an urban setting," *Sustainability*, vol. 15, no. 4, 2023, Art. no. 3528.
- [10] A. Mesaros et al., "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 379–393, Feb. 2018.
- [11] Y. Hou, B. Kang, and D. Botteldooren, "Audio-visual scene classification via contrastive event-object alignment and semantic-based fusion," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2022, pp. 1–6.
- [12] S. Wang, T. Heittola, A. Mesaros, and T. Virtanen, "Audio-visual scene classification: Analysis of DCASE 2021 challenge submissions," in *Proc. Detection Classification Acoustic Scenes Events*, 2021, pp. 45–49.
- [13] D. Barchiesi, D. Giannoulis, D. Stowell, and M. D. Plumbley, "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Process. Mag.*, vol. 32, no. 3, pp. 16–34, May 2015.
- [14] P. Jin, Z. Si, H. Wan, and X. Xiong, "Emotion classification algorithm for audiovisual scenes based on low-frequency signals," *Appl. Sci.*, vol. 13, no. 12, 2023, Art. no. 7122.
- [15] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, and B. Schuller, "Attention-based convolutional neural networks for acoustic scene classification," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2018, pp. 39–43.
- [16] A. M. Basbug and M. Sert, "Acoustic scene classification using spatial pyramid pooling with convolutional neural networks," in *Proc. IEEE Int. Conf. Semantic Comput.*, 2019, pp. 128–131.
- [17] Z. Li et al., "Multi-level attention model with deep scattering spectrum for acoustic scene classification," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2019, pp. 396–401.
- [18] W. G. Choi, J. H. Chang, J. M. Yang, and H. G. Moon, "Instance-level loss based multiple-instance learning framework for acoustic scene classification," 2022, *arXiv:2203.08439*.
- [19] Y. Hou, Q. Kong, J. Wang, and S. Li, "Polyphonic audio tagging with sequentially labelled data using crnn with learnable gated linear units," in *Proc. Detection Classification Acoustic Scenes Events Workshop*, 2018, pp. 78–82.
- [20] H. Phan et al., "Spatio-temporal attention pooling for audio scene classification," in *Proc. INTERSPEECH*, 2019, pp. 3845–3849.
- [21] L. Zhou, Y. Zhou, X. Qi, J. Hu, T. L. Lam, and Y. Xu, "Attentional graph convolutional network for structure-aware audiovisual scene classification," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–15, 2023.
- [22] H. L. Bear, I. Nolasco, and E. Benetos, "Towards joint sound scene and polyphonic sound event recognition," in *Proc. INTERSPEECH*, 2019, pp. 1236–1240.
- [23] N. Tonami, K. Imoto, R. Yamanishi, and Y. Yamashita, "Joint analysis of sound events and acoustic scenes using multitask learning," *IEICE Trans. Inf. Syst.*, vol. 104, no. 2, pp. 294–301, 2021.
- [24] Y. Hou, B. Kang, W. Van Hauwermeiren, and D. Botteldooren, "Relation-guided acoustic scene classification aided with event embeddings," in *Proc. Int. Joint Conf. Neural Netw.*, 2022, pp. 1–8.
- [25] D. Oldoni, B. De Coensel, A. Bockstael, M. Boes, B. De Baets, and D. Botteldooren, "The acoustic summary as a tool for representing urban sound environments," *Landscape Urban Plan.*, vol. 144, pp. 34–48, 2015.
- [26] S. Song, Z. Shao, S. Jaiswal, L. Shen, M. Valstar, and H. Gunes, "Learning graph representation of person-specific cognitive processes from audio-visual behaviours for automatic personality recognition," 2021, *arXiv:2110.13570*.
- [27] C. Luo, S. Song, W. Xie, L. Shen, and H. Gunes, "Learning multi-dimensional edge feature-based AU relation graph for facial action unit recognition," in *Proc. Int. Joint Conf. Artif. Intell.*, 2022, pp. 1239–1246.
- [28] G. Huo, Y. Zhang, B. Wang, J. Gao, Y. Hu, and B. Yin, "Hierarchical spatio-temporal graph convolutional networks and transformer network for traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 4, pp. 3855–3867, Apr. 2023.
- [29] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
- [30] Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow, "Gated graph sequence neural networks," in *Proc. Int. Conf. Learn. Representations*, 2016, pp. 1–20.
- [31] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.
- [32] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 7705–7716.
- [33] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *Ann. Statist.*, vol. 48, no. 4, pp. 1875–1897, 2020.
- [34] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [35] D. Zhang, H. Lan, Z. Ma, Z. Yang, X. Wu, and X. Huang, "Spatial-temporal gated graph convolutional network: A new deep learning framework for long-term traffic speed forecasting," *J. Intell. Fuzzy Syst.*, vol. 44, no. 6, pp. 10437–10450, 2023.
- [36] H. Phan, T. N. T. Nguyen, P. Koch, and A. Mertins, "Polyphonic audio event detection: Multi-label or multi-class multi-task classification problem?," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 8877–8881.
- [37] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proc. Detection Classification Acoustic Scenes Events Workshop*, 2018, pp. 9–13.
- [38] T. Komatsu, K. Imoto, and M. Togami, "Scene-dependent acoustic event detection with scene conditioning and fake-scene-conditioned loss," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 646–650.
- [39] L. Ilya and H. Frank, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–18.
- [40] Y. Hou, F. K. Soong, J. Luan, and S. Li, "Transfer learning for improving singing-voice detection in polyphonic instrumental music," in *Proc. INTERSPEECH*, 2020, pp. 1236–1240.
- [41] J. F. Gemmeke et al., "AudioSet: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 776–780.
- [42] T. Nguyen and F. Pernkopf, "Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters," in *Proc. Detection Classification Acoust. Scenes Events Workshop*, 2018, pp. 34–38.