This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2023.3319233

Audio Event-Relational Graph Representation Learning for Acoustic Scene Classification

Yuanbo Hou, *Student Member, IEEE*, Siyang Song, Chuang Yu, Wenwu Wang, *Senior Member, IEEE*, 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 paper 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. Visualizations of graph representations learned by ERGL are available here (https://github.com/Yuanbo2020/ERGL).**

***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].

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 method with mel spectrograms as input 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][24]. Relation-guided ASC [25] 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 utilized 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 the variations in relations among the occurring AEs in these scenes [26]. This paper 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 [27][28], 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 [29][30], which aim to capture spatial and temporal dependencies of graph sequences or multigraphs, are unsuitable for modelling ERGs. ERGs are fed to a gated graph convolutional network (Gated GCN) [31][32] for ASC. Experiments show that graph representations learned by ERGL, from only several explicit audio event semantic embeddings, can facilitate the discrimination between different ASs. The rest of the paper is organized as follows. Section II introduces ERGL. Section III describes experimental setup, and analyzes results. Section IV draws conclusions.

II. Audio Event-Relational Graph Learning

This section discusses the proposed approach for learning the scene-related event-relational graph (ERG) representation

Y. Hou and D. Botteldooren are with the Waves Research Group, Ghent University, Belgium (e-mail: {Yuanbo.Hou, Dick.Botteldooren}@UGent.be).

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

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

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2023.3319233

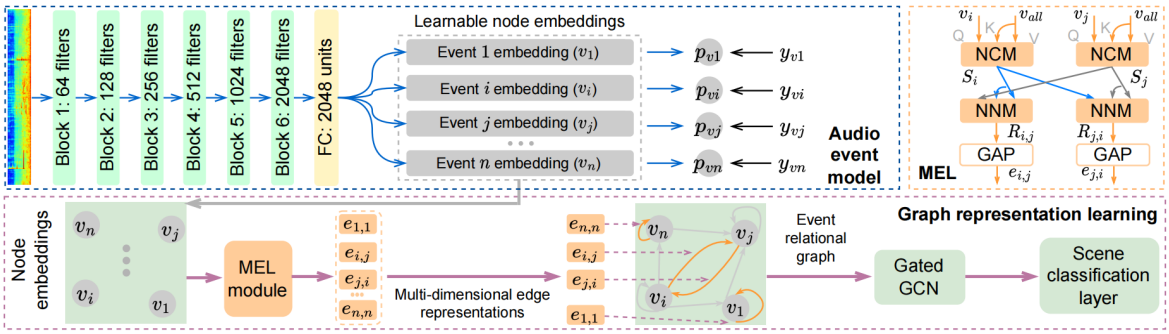


Fig. 1: The framework of the proposed scene-dependent **e**vent-**r**elational **g**raph **l**earning (ERGL) for ASC.

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 [33] for node embedding generation, where each describes a specific AE. PANNs offer excellent performance and generalization ability in AE recognition [33]. 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 × 3, a batch normalization [34], and a ReLU activation function [35]. 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., $\mathcal{L}_{\text{event}} = MSE(p_{vi}, y_{vi})$). To train the audio event node embedding generation module, we employ PANNs, pre-trained on AudioSet [36], which contains 527 classes of AEs, to generate pseudo labels of AEs. This produces a 527-dimensional soft pseudo label $y = [y_{v1}, y_{v2}, ..., y_{v527}]$ for each audio clip, describing the occurrence probabilities of 527 classes of pre-defined 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 occurrence probabilities in all training data, and only 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 **m**ulti-dimensional **e**dge feature **l**earning (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 modelling (NCM) and the node-node relation modelling (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 [37] between it and the global contextual representation v_{all} consisting of the mean of all node features, where node v_i is used as the query, and v_{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 Φ 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 $\mathcal{S}_i$ ($i = 1, 2, \cdots, n$) 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 [37] 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 $v = \{v_1, v_2, \cdots, v_n\}$ and $n \times n$ multi-dimensional directed edge representations $e = \{e_{1,1}, \cdots, e_{i,j}, \cdots, e_{n,n}\}$ is obtained, we feed G^0 to the Gated GCN [31][32] 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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2023.3319233

layer. Cross entropy (CE) [38] is used as the loss function in ASC between the scene prediction p_s and the scene true label y_s, $\mathcal{L}_{\text{scene}} = CE(p_s, y_s)$. Hence, the final loss of the proposed ERGL is $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{event}} + \lambda_2 \mathcal{L}_{\text{scene}}$, where λ_i $(i = 1, 2)$ are the loss weight in training, and λ_i default to 1 in this paper. The effect of U, which defaults to 2, on the model performance will be explored in the experiments discussed later.

III. Experiments and Results

A. Dataset, baseline, experimental setup, and metric

This paper uses TUT Urban Acoustic Scenes 2018 dataset (UAS) [39] 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 Sec. 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 [39], where the test, training, and validation sets contain 2518, 5509, and 613 samples, respectively.

In addition to a typical CNN-based approach [39] for ASC, this paper also employs attention-based [15], spatio-temporal attention [20], multiple-instance learning [18], and scene-event joint learning [22][23][24][25] as baselines for comparison.

Following [33], 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. Dropout and normalization are used in training to prevent over-fitting of the model. A batch size of 64 and AdamW optimizer [40] with a learning rate of 1e-4 are used to minimize the loss. The systems are trained on a GPU card (Tesla V100) for 400 epochs. The average accuracy (*Acc*) [1] is used as the performance metric.

B. Results and analysis

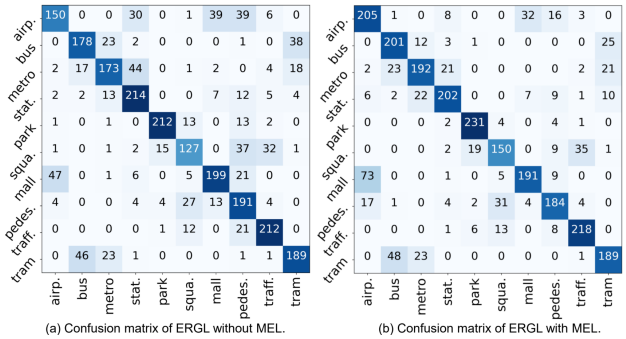
The pseudo labels composed of probability outputs by PANNs [33] are used as supervision information for training the audio event model, as shown in Fig. 1. The accuracy of pseudo labels that do not involve human verification cannot be evaluated. Since our goal is ASC, we will mainly show ASC results, rather than the accuracy of pseudo labels on AEs.

TABLE I: *Acc* of ERGL at different n values on the validation set.

#	Top n	*Acc* (%)	#	Top n	*Acc* (%)
1	10	94.03 ± 2.41	6	150	93.76 ± 2.76
2	25	**95.92** ± 2.29	7	200	93.05 ± 2.55
3	50	94.65 ± 2.75	8	250	93.10 ± 1.97
4	75	94.47 ± 1.89	9	300	92.46 ± 3.89
5	100	94.22 ± 2.91	10	400	92.29 ± 2.07

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



(a) Confusion matrix of ERGL without MEL. (b) Confusion matrix of ERGL with MEL.

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.)

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.

TABLE II: *Acc* of ERGL at different U layers on the validation set.

#	U	*Acc* (%)	#	U	*Acc* (%)
1	1	94.42 ± 1.97	5	5	94.79 ± 2.43
2	2	**95.92** ± 2.29	6	6	94.29 ± 1.92
3	3	95.66 ± 2.58	7	7	94.18 ± 1.63
4	4	95.09 ± 2.74	8	8	94.11 ± 1.85

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: Ablation study of AE relation-related MEL in ERGL.

MEL	NCM	✗	✓	✗	✓
	NNM	✗	✗	✓	✓
Test set *Acc* (%)		73.35±1.98	74.42±1.99	75.69±1.54	**78.08**±2.06

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)". The accuracy of *airp.* is improved the most, mainly because

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2023.3319233

the misclassified samples between *stat.* (metro station), *pedes.* (street pedestrian), and *airp.* is reduced from 30 and 39 to 8 and 16, respectively. In contrast, introducing MEL increases the misclassified samples of *stat.* and *metro*. Even for humans, it is challenging to distinguish these similar scenes relying on audio only. In short, introducing AE-based relational edges can effectively improve the performance of ERGL in 6 scenes, increasing its *Acc* from 73.35% to 78.08% in Table III.

Comparison with non-ensemble ASC methods[1]**.** Table IV shows the results of CNN [39] and other methods on the same test set. PANNs [33] are explored in two modes. In the fixed mode [41], the parameters of PANNs are not updated in training, and the scenes are classified using the AEs' knowledge learned from AudioSet. The result of fixed-mode PANNs is comparable to Baseline, indicating that PANNs with AEs knowledge have a certain discriminative ability for scenes. In contrast, ERGL using just audio event embeddings significantly improves ASC accuracy even though these event embeddings are learned from pseudo labels without verification. Overall, the proposed end-to-end EGRL without data augmentations offers competitive results. This illustrates that ERGL can effectively discriminate different scenes by relying only on several scene-aware events semantic embeddings.

TABLE IV: Comparison of non-ensemble systems on the test set.

System	Model structure	*Acc (%)*
PANNs [33] (Fixed mode)	VGG-like CNN	56.9
Baseline [39]	CNN	59.7
NNF-CNNEns [42]	CNN and nearest neighbor filters	69.3
Spatio-temporal Attention [20]	CRNN	72.5
Attention-based CNN [15]	Attention-Based CNN	72.6
PANNs (Fine-tuning mode)	VGG-like CNN	73.8
Instance-based ASC [18]	CNN	73.9
Wavelet-based spectrum [17]	CRNN	76.6
Proposed ERGL	CNN and Graph Learning	**78.1**

Comparison with other scene-event analysis methods. The proposed ERGL infers target scenes based on the ERG in the scene, which is scene-event joint analysis. Table V compares ERGL with other scene-event joint analysis methods. The method based on the same latent space for scene and event classification [22] performs the worst. The reason may be that real-life scenes and AEs differ at the semantic level and the feature space. Papers [23][24] use shared base and separated high-level features to identify scenes and AEs. RGASC [25] exploits the scene-event relationship to guide the model to achieve mutually beneficial scene-event classification. The ERGL relies only on semantic embeddings of AEs to achieve one-way event-to-scene inference, and recognizes scenes based on the corresponding explicit semantic ERG. Notably, ERGL, which only needs 25 classes of AEs' information, outperforms RGASC with 527 classes of AEs' information. Overall, the proposed ERGL achieves promising results, demonstrating the feasibility of ASC based on the ERG.

Analysis of confusion. To further explore the reasons for the misclassification between scenes for graph representation-based classification. Fig. 3 presents the structure of graph representations of audio clips from different scenes. In Fig. 3

[1]The results are based on DCASE2018 T1A website. Since the proposed ERGL only uses one end-to-end model with one type of acoustic feature without data augmentation, but the top 3 methods in T1A are ensembles of multiple models with multiple features, we omit these results in Table IV.

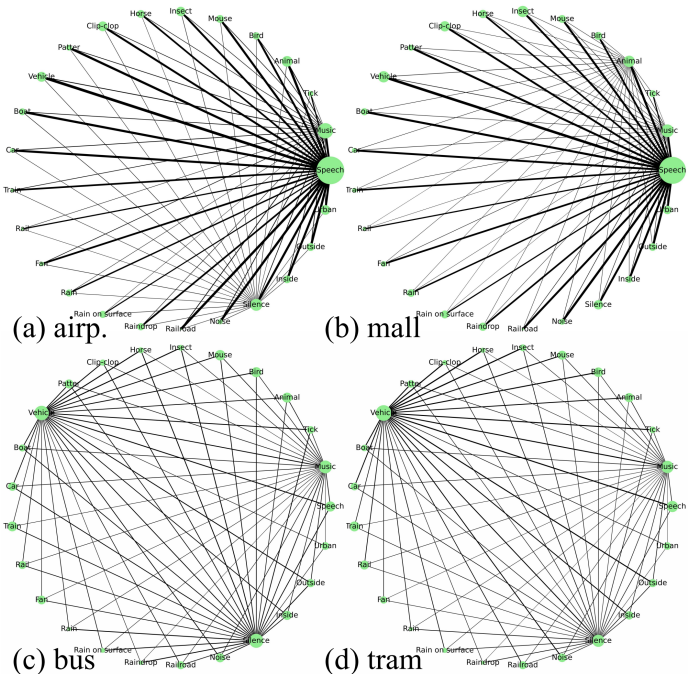


Fig. 3: Graph structures of test audio samples from different scenes. Green dots represent 25 classes of AEs used in this paper. A larger dot denotes a higher probability of the event. A thicker line denotes a larger edge value between nodes in the graph representation. (*For visualization, a node is considered inactive if its probability is less than 0.1. Edges between two inactive nodes are not displayed. Multi-dimensional features of the edge are represented by their mean value.*)

(a) and (b), the dominant AEs in *airp.* and *mall* scenes are *speech* and *music*, so the connections in the graph are mainly gathered around *speech* and *music*. This may be the reason that ERGL confuses the *airp.* and *mall* scenes in Fig. 2. In addition to these similarities, the third focused audio event in Fig. 3 (a) is *silence*, while that in Fig. 3 (b) is *animal* sounds, which reflects the differences between the two scenes. In Fig. 3 (c) and (d), the dominant AEs in *bus* and *tram* scenes are *vehicle*, *music*, *speech*, *train*, *car* and *silence*. And the AEs with dominant connections are the same: *vehicle*, *music* and *silence*. With similar dominant events and graph structures, the model tends to be confused by these similar scenes.

TABLE V: *Acc* of scene-event joint analysis methods on test set.

#	Method	*Acc* (%)
1	Scene and event jointly classification [22]	52.35
2	MTL-based event and scene analysis [23]	61.69
3	Conditional scene and event recognition [24]	66.39
4	Relation-guided ASC (RGASC) [25]	77.35
5	The proposed ERGL	**78.08**

IV. CONCLUSION

To perform ASC and simultaneously clearly answer which cues are used in classifying, we propose a scene-dependent audio event-relational graph representation learning method for ASC, which represents acoustic scenes by a set of scene-aware nodes with explicit AEs semantic embeddings, and specifically produces scene-task-relevant multi-dimensional edge features to describe AE-based semantic relations between nodes. Experimental results show that ERGL achieves competitive ASC performance by learning ERGs, which are constructed on semantically meaningful embeddings of only a limited number of AEs. Future work will try to build a dynamic graph-based model and test ERGL on more datasets.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Signal Processing Letters. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/LSP.2023.3319233

REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*, Springer, 2018.

[2] B. Defréville, F. Pachet, C. Rosin, and P. Roy, "Automatic recognition of urban sound sources," in *Audio Engineering Society Convention 120*. Audio Engineering Society, 2006.

[3] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[4] L. Pham, D. Ngo, D. Salovic, A. Jalali, A. Schindler, P. X. Nguyen, K. Tran, and H. C. Vu, "Lightweight deep neural networks for acoustic scene classification and an effective visualization for presenting sound scene contexts," *Applied Acoustics*, 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 *Proceedings of IEEE WASPAA*, 2015, pp. 1–5.

[6] Y. Hou, S. Song, C. Luo, A. Mitchell, Q. Ren, W. Xie, J. Kang, W. Wang, and D. Botteldooren, "Joint Prediction of Audio Event and Annoyance Rating in an Urban Soundscape by Hierarchical Graph Representation Learning," in *Proceedings of INTERSPEECH*, 2023, pp. 68–72.

[7] G. Grossi, R. Lanzarotti, P. Napoletano, N. Noceti, and F. Odone, "Positive technology for elderly well-being: A review," *Pattern Recognition Letters*, 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 *Proceedings of Workshop on Multimedia for Accessible Human Computer Interface*, 2018, pp. 1–7.

[9] V. Carrasco, J. P. Arenas, P. Huijse, D. Espejo, V. Vargas, et al., "Application of Deep Learning to Enforce Environmental Noise Regulation in an Urban Setting," *Sustainability*, vol. 15, no. 4, pp. 3528, 2023.

[10] A. Mesaros, T. Heittola, E. Benetos, P. Foster, M. Lagrange, T. Virtanen, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2017.

[11] Y. Hou, B. Kang, and D. Botteldooren, "Audio-visual scene classification via contrastive event-object alignment and semantic-based fusion," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 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 *Proceedings of Detection and Classication of Acoustic Scenes and 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 Processing Magazine*, vol. 32, no. 3, pp. 16–34, 2015.

[14] P. Jin, Z. Si, H. Wan, and X. Xiong, "Emotion Classification Algorithm for Audiovisual Scenes Based on Low-Frequency Signals," *Applied Sciences*, vol. 13, no. 12, pp. 7122, 2023.

[15] Z. Ren, Q. Kong, K. Qian, M. D. Plumbley, and B. Schuller, "Attention-based convolutional neural networks for acoustic scene classification," in *Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) 2018 Workshop*, 2018, pp. 39–43.

[16] A. M. Basbug and M. Sert, "Acoustic scene classification using spatial pyramid pooling with convolutional neural networks," in *Proceedings of IEEE International Conference on Semantic Computing (ICSC)*, 2019, pp. 128–131.

[17] Z. Li, Y. Hou, X. Xie, S. Li, L. Zhang, S. Du, and W. Liu, "Multi-level attention model with deep scattering spectrum for acoustic scene classification," in *Proceedings of IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, 2019, pp. 396–401.

[18] WG Choi, JH Chang, JM Yang, and HG Moon, "Instance-level loss based multiple-instance learning framework for acoustic scene classification," *arXiv preprint arXiv:2203.08439*, 2022.

[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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, November 2018, pp. 78–82.

[20] H. Phan, O. Y Chén, L. Pham, P. Koch, M. De V., I. McLoughlin, and A. Mertins, "Spatio-temporal attention pooling for audio scene classification," in *Proceedings of 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 Transactions on Instrumentation and Measurement*, 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 *Proceedings of 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 Transactions on Information and Systems*, vol. 104, no. 2, pp. 294–301, 2021.

[24] T. Komatsu, K. Imoto, and M. Togami, "Scene-dependent acoustic event detection with scene conditioning and fake-scene-conditioned loss," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 646–650.

[25] Y. Hou, B. Kang, W. Van Hauwermeiren, and D. Botteldooren, "Relation-guided acoustic scene classification aided with event embeddings," in *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.

[26] 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 and Urban Planning*, vol. 144, pp. 34–48, 2015.

[27] 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," *arXiv preprint arXiv:2110.13570*, 2021.

[28] 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 *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, 2022, pp. 1239–1246.

[29] 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 Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 3855–3867, 2023.

[30] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on neural networks and learning systems*, vol. 32, no. 1, pp. 4–24, 2020.

[31] 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," *Journal of Intelligent & Fuzzy Systems*, vol. 44, no. 6, pp. 10437–10450, 2023.

[32] Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow, "Gated graph sequence neural networks," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2016.

[33] 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 Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[34] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," in *Proceedings of International Conference on Neural Information Processing Systems*, 2018, pp. 7705–7716.

[35] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *The Annals of Statistics*, vol. 48, no. 4, pp. 1875–1897, 2020.

[36] J. F. Gemmeke, D. PW. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[38] H. Phan, T. Nguyen, Ngoc T., P. Koch, and A. Mertins, "Polyphonic audio event detection: multi-label or multi-class multi-task classification problem?," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8877–8881.

[39] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2018, pp. 9–13.

[40] L. Ilya and H. Frank, "Decoupled weight decay regularization," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

[41] Y. Hou, F. K. Soong, J. Luan, and S. Li, "Transfer learning for improving singing-voice detection in polyphonic instrumental music," in *Proceedings of INTERSPEECH*, 2020, pp. 1236–1240.

[42] T. Nguyen and F. Pernkopf, "Acoustic scene classification using a convolutional neural network ensemble and nearest neighbor filters," in *Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2018, pp. 34–38.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/