Relation-guided acoustic scene classification aided with event embeddings

Yuanbo Hou
WAVES
Ghent University
Gent, Belgium
Yuanbo.Hou@UGent.be

Bo Kang
IDLAB
Ghent University
Gent, Belgium
Bo.Kang@UGent.be

Wout Van Hauwermeiren
WAVES
Ghent University
Gent, Belgium
Wout.VanHauwermeiren@UGent.be

Dick Botteldooren
WAVES
Ghent University
Gent, Belgium
Dick.Botteldooren@UGent.be

***Abstract*—In real life, acoustic scenes and audio events are naturally correlated. Humans instinctively rely on fine-grained audio events as well as the overall sound characteristics to distinguish diverse acoustic scenes. Yet, most previous approaches treat acoustic scene classification (ASC) and audio event classification (AEC) as two independent tasks. A few studies on scene and event joint classification either use synthetic audio datasets that hardly match the real world, or simply use the multi-task framework to perform two tasks at the same time. Neither of these two ways makes full use of the implicit and inherent relation between fine-grained events and coarse-grained scenes. To this end, this paper proposes a relation-guided ASC (RGASC) model to further exploit and coordinate the scene-event relation for the mutual benefit of scene and event recognition. The TUT Urban Acoustic Scenes 2018 dataset (TUT2018) is annotated with pseudo labels of events by a simple and efficient audio-related pre-trained model PANN, which is one of the state-of-the-art AEC models. Then, a prior scene-event relation matrix is defined as the average probability of the presence of each event type in each scene class. Finally, the two-tower RGASC model is jointly trained on the real-life dataset TUT2018 for both scene and event classification. The following results are achieved. 1) RGASC effectively coordinates the true information of coarse-grained scenes and the pseudo information of fine-grained events. 2) The event embeddings learned from pseudo labels under the guidance of prior scene-event relations help reduce the confusion between similar acoustic scenes. 3) Compared with other (non-ensemble) methods, RGASC improves the scene classification accuracy on the real-life dataset.**

***Index Terms*—Acoustic scene classification, audio event classification, pseudo label, collaboratively classify**

I. Introduction

Acoustic scene classification (ASC) tags audio recordings using the predefined semantic labels that characterize the environment and situation in which it was recorded. Audio event classification (AEC) is dedicated to multi-label classification on audio clips and aims to identify the presence of target audio events. ASC and AEC can be used in a wide variety of applications such as robot hearing [1], audio forensics [2], emergency detection [3] and road surveillance [4].

Prior studies related to IEEE AASP Challenges in Detection and Classification of Acoustic Scenes and Events (DCASE) [5]–[7] commonly handle ASC and AEC as two separate tasks and tune models for each task individually. However, real-world audio streams include both acoustic scenes and events and they are inherently correlated. For example, in the acoustic scene *metro station*, audio events of *bell ringing* and *engine starting* are likely to occur. Such fine-grained events are the fundamental building blocks of polyphonic acoustic scenes. Therefore, a joint scene and event recognition method based on an artificially synthesized dataset is proposed in [8], expecting to train a shared acoustic feature encoder for scenes and events. However, the artificial dataset in [8] does not accurately catch complexities between real-world acoustic scenes and events. Then in [9], robust representations for environmental audio scenes and events are learned by generative model-driven representations and have proved to be effective in audio-related tasks. Another class of studies for joint analysis of scene and event refers to multi-task learning (MTL) [10]. Several convolutional layers are shared in a multi-task model as they [11] expect to learn and utilize shared low-level representations and separated high-level representations of scenes and events. However, the one-hot hard labels of scenes used in [11] cannot model the extent to which audio events and acoustic scenes are related. To alleviate this issue, the output of a trained scene model is used as the teacher in [12] to guide the learning of the scene branch, which works as the student, in the joint scene-event classification model based on the teacher-student learning [13]. To learn the knowledge of events under scene conditions, a scene-event joint analysis model based on scene-conditioned loss is proposed [14]. Overall, in previous scene and event joint analysis works, papers [8] [9] [11] do not explore the implicit scene-event relation, paper [12] exploits the one-way scene-to-scene relation. Although paper [14] uses the one-way scene-to-event relation by conditional loss, that relation is derived from the artificial dataset in [8], and is difficult to match the complex and intricate scene-event relation in the real world.

In contrast to prior works, this paper is not only interested in obtaining a shared representation encoder with scenes and events knowledge, but also in how the real-world implicit and inherent scene-event relation can be used to guide the model to bidirectionally fuse the information of coarse-grained scenes and fine-grained events to reduce confusion of similar scenes, even if the event information is derived from unverified pseudo labels (a proxy to unavailable ground-truth labels) [15].

To the best of our knowledge, there are no publicly available large real-life datasets that contain both acoustic scene and

event labels. Hence, a large real-life acoustic scene dataset, TUT Urban Acoustic Scenes 2018[1], with diverse audio events is used in this paper [7]. In order to obtain labels for events in the real-life acoustic scene dataset, a simple and efficient pre-trained audio-related model PANN [16] is used to tag each audio clip with pseudo labels of 527 classes of audio events. Relying on true labels of scenes and pseudo labels of events, a scene-event relation matrix is derived to model the implicit relation between real-life scenes and events. Then, with the aid of pseudo labels and the prior knowledge of joint scene-event relation matrix, a relation-guided ASC and AEC two-tower model is proposed to mutually estimate the knowledge of scenes and corresponding events and explore the possibility of collaboratively classifying scenes and events.

This paper is organized as follows. Section 2 introduces the method. Section 3 describes the dataset, experimental setup, results, and analysis. Finally, Section 4 gives conclusions.

II. Method

In this section, the scene-event relation matrix is presented and is applied to the relation-guided two-tower convolutional neural networks (CNN) for the acoustic scene classification (ASC) and audio event classification (AEC) tasks.

A. Prior scene-event relation matrix

Coarse-grained acoustic scenes are highly correlated with fine-grained audio events, for example, *car passing* and *horn screeching* are more likely to occur in the *street* scene than *cheerful music*. In contrast, *joyful music* accompanied by *people walking* is more common in the *shopping mall* scene. The relation between real-life scenes and events is not simply 1 for presence or 0 for absence, but rather a likelihood expressed by probability. Inspired by such an intuitive observation, this paper attempts to build the scene-event relation matrix on the real-world acoustic scene dataset, instead of the simple connections in [14] represented by 0 and 1 on synthetic datasets. Since there are no public large real-life datasets that contain both acoustic scenes and events labels, an audio-related model PANN [16] is used in this paper to tag audio clips with pseudo labels of events. PANN [16] is trained and performs well on Audioset [17], which contains 527 classes of polyphonic audio events in daily life.

Given the probability of 527 classes of events for the i-th audio clip (a_i) of the j-th scene S_j is $P(S_j, a_i) \in \mathbb{R}^{1\times 527}$,

$$P(S_j, a_i) = [p_{e_1}, p_{e_2}, p_{e_3}, ..., p_{e_{527}}] \quad (1)$$

where $p_{en} \in [0, 1], n \in [1, 527]$, and p_{e_n} is the probability of the occurrence of the n-th event in the audio clip. The $p_{en} \in [0, 1]$ implies that in the AEC task, different audio events are generally considered to be independent of each other. Then, the average probability of audio events in the acoustic scene S_j can be notated as $P(S_j)$,

$$P(S_j) = 1/I_j \sum_{i=1}^{I_j} P(S_j, a_i) \quad (2)$$

[1]Dataset available: https://zenodo.org/record/1228142#.YfJ-Qv7MJnI

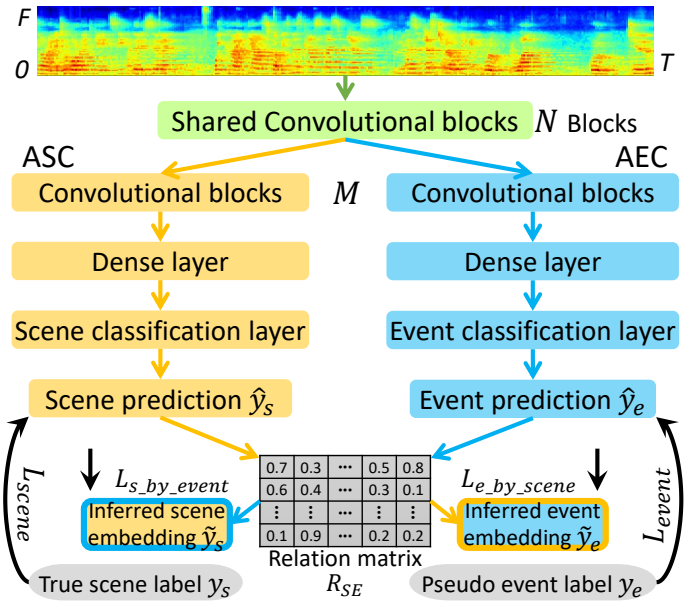


Fig. 1: The architecture of the relation-guided acoustic scene classification (RGASC) model.

where I_j is the total number of audio clips in the scene S_j. Then, the prior scene-event relation matrix R_{SE} is composed of $P(S_j)$ rows for all scenes: $R_{SE} \in \mathbb{R}^{K\times 527}$,

$$R_{SE} = [P(S_1), P(S_2), P(S_3), ..., P(S_K)]^T \quad (3)$$

where K is the number of acoustic scenes in the dataset. Next, R_{SE} from the training set will be introduced into the two-tower model of ASC and AEC. The role of R_{SE} is to guide the model to coordinate and utilize the implicit relation between coarse-grained scenes and fine-grained events during the training phase.

B. Relation-guided collaborative two-tower model

The core of this part is how to embed the existing fixed prior relation matrix R_{SE} relating coarse-grained information from true labels of scenes and fine-grained information from pseudo labels of events into the learning process of the model. To exploit the fixed prior knowledge of R_{SE}, a relation-guided two-tower model is proposed in Fig. 1.

The raw waveform is first converted to time-frequency (T-F) representations using log mel spectrograms [18]. Then, the first N convolutional blocks of the two-tower model are used to extract shared basic acoustic features of scene and event as inspired by [8], [9] and [11]. Compared with high-level acoustic features, basic local acoustic features learned by models are more transferable and applicable [19], which is beneficial for model generalization [20]. Then, the remaining M convolutional blocks are applied to the ASC tower and the AEC tower to capture local patterns that are beneficial to scenes and events, respectively. There are a total of 6 convolutional blocks in the proposed model. Therefore, given the first N convolution blocks are shared, the remaining $M = 6-N$ convolutional blocks will be used to learn the task-

oriented representations of each tower. The optimal ratio of N and M will be further explored in the experimental section.

Referring to CNNs in VGG [21], each convolutional block contains 2 convolutional layers with the kernel size of (3 × 3). Batch normalization [22] and ReLU activation functions [23] are used to accelerate and stabilize the training. Next, a linear dense layer is applied to the high-level representations, followed by a scene classification dense layer with softmax activation function and an event classification dense layer with sigmoid activation function, respectively. More details and code, please see the homepage[2].

ASC is a task dedicated to single-label multi-class classification, so the cross entropy loss [7] is used as the loss function in the ASC tower between the scene prediction $\hat{y}_s \in \mathbb{R}^K$ and the scene true label $y_s \in \mathbb{R}^K$,

$$L_{scene} = -\sum_{j=1}^{K} y_{s_j} \log(\hat{y}_{s_j}) \tag{4}$$

AEC aims to perform multi-label classification on audio clips to detect multiple targets simultaneously. Therefore, the binary cross-entropy [16] loss is used in the AEC tower between the event prediction $\hat{y}_e \in \mathbb{R}^{527}$ and the pseudo label $y_e \in \mathbb{R}^{527}$,

$$L_{event} = -\sum_{i=1}^{527} y_{e_i} \log(\hat{y}_{e_i}) + (1 - y_{e_i}) \log(1 - \hat{y}_{e_i}) \tag{5}$$

To guide the model to explore relations between scenes and events based on the prior knowledge of R_{SE} from the training set. The scene prediction $\hat{y}_s$ is mapped to the latent event space via R_{SE} to obtain the corresponding event information $\tilde{y}_e$ inferred from the scene prediction, $\tilde{y}_e = \hat{y}_s \cdot R_{SE}$. The R_{SE} is derived from the probability of events in each scene, hence the embedding vector $\tilde{y}_e$ uses the prior information of events in the scene. To measure the distance between the inferred event vector $\tilde{y}_e$ and the actual event prediction $\hat{y}_e$ in the latent space, the $\hat{y}_e$ from the AEC tower is used as the reference in the mean squared error (MSE) loss,

$$L_{\text{e\_by\_scene}} = 1/527 \sum_{i=1}^{527} (\hat{y}_{e_i} - \tilde{y}_{e_i})^2 \tag{6}$$

The embedding vector $\tilde{y}_e$ is not a probability distribution, so the regression loss MSE is used to minimize the relation-guided $L_{\text{e\_by\_scene}}$ loss of inferred event by scene information. This expects the inferred event vector $\tilde{y}_e$ to be close to the actual event vector $\hat{y}_e$ in the latent representation space. Furthermore, the study [24] on the entropy-based loss and MSE shows that MSE loss can better correct the error between the estimated value and the target.

Similarly, the inner product $\tilde{y}_s = \hat{y}_e \cdot R_{SE}^T$ defines the relation-guided embedding vector for scenes. The relation-guided embedding vector $\tilde{y}_s$ from event prediction indicates the possibility of different scenes, and higher similarity means the corresponding scene is more likely to occur. Similar to $\tilde{y}_e$, $\tilde{y}_s$ is not a probability distribution, so MSE is adopted to minimize the loss of inferred scene by event information to expect the inferred scene vector $\tilde{y}_s$ to be close to the actual scene vector $\hat{y}_s$ in the latent representation space.

$$L_{\text{s\_by\_event}} = 1/K \sum_{i=1}^{K} (\hat{y}_{s_i} - \tilde{y}_{s_i})^2 \tag{7}$$

[2]Homepage: https://github.com/Yuanbo2020/RGASC

The final loss function of the two-tower models is given by the weighted sum of the separate loss functions:

$$L = \lambda_1 L_{\text{scene}} + \lambda_2 L_{\text{s\_by\_event}} + \lambda_3 L_{\text{event}} + \lambda_4 L_{\text{e\_by\_scene}} \tag{8}$$

where λ_i is the scale factor of each loss function. λ_i defaults to 1. In the experimental section, various configurations of λ_i are explored. The ASC tower will benefit from L_{scene} and $L_{\text{e\_by\_scene}}$. The loss $L_{\text{e\_by\_scene}}$ is fed to the ASC tower, expecting to obtain a more coordinated scene prediction with the event prediction of AEC tower. Likewise, the AEC tower will benefit from L_{event} and $L_{\text{s\_by\_event}}$, the loss $L_{\text{s\_by\_event}}$ benefits the collaborative learning of AEC tower.

III. Experiments and results

A. Dataset and Experimental Setup

The dataset used in this paper is the TUT Urban Acoustic Scenes 2018 development dataset (TUT2018) [7] with 8640 10-seconds clips totaling 24 hours and contains 10 classes of acoustic scenes from real life. For each acoustic scene, there are 864 examples. These audio recordings were recorded in 6 different European locations. This real-life dataset does not contain labels for events. Therefore, to obtain the event labels, a pure CNN-based pre-trained model PANN[3] [16] is used to tag each audio clip with a pseudo label indicating the probability of 527 classes of audio events.

The log mel-bank energy with 64 banks [18] is used as the acoustic feature in this paper. This is extracted by the Short-Time Fourier Transform (STFT) with a Hamming window length of 46 *ms* and a window overlap of $1/3$ [25]. Dropout [26] and normalization are used in the training to prevent over-fitting of the model. Adam optimizer [27] with a default initial learning rate of 0.001 minimizes the loss function. The default training batch size is 64. To facilitate the comparison of experimental results with other systems, the training/testing split of the TUT2018 dataset follows the default split of the DCASE 2018 Task 1 Subtask A[4]. The model is trained on a single graphical card (Tesla V100-SXM2-32GB) for a fixed amount of 100 epochs. The average accuracy (Acc) [25] is used as the metric in this paper. A higher Acc indicates a better performance to distinguish different scenes.

B. Results and Analysis

This section analyzes the performance of the proposed method based on the following **R**esearch **Q**uestions (**RQ**):

• **RQ1**: How many convolutional blocks should be shared?

As discussed in Section II-B, the first question to be explored is what proportion of the joint scene-event representations should be shared? From another perspective, the ratio between shared and individual blocks determines whether the model is dedicated to learning shared knowledge or is more inclined to explore task-dependent knowledge. There is both competition and mutual influence between the shared and separated blocks in the RGSAC model.

[3]We used the CNN14 in PANN in this paper. The pre-trained model (the model named Cnn14_16k_mAP=0.438.pth) of CNN14 is available here: https://zenodo.org/record/3987831

[4]http://dcase.community/challenge2018/task-acoustic-scene-classification

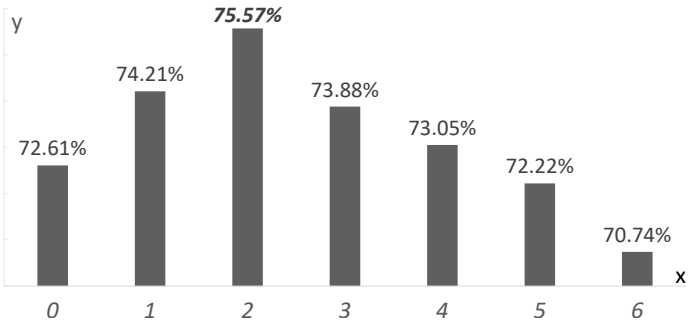


Fig. 2: The effect of different numbers of shared blocks on the proposed model on the test set of real-life dataset. X-axis is the number of shared blocks, and y-axis is the accuracy of scene classification.

There are a total of 6 convolution blocks in the proposed RGASC model. Therefore, when the number of shared blocks is 0 in Fig. 2, the learning of scene representations and event representations in the two-tower RGASC model will be independent, and the model will not be able to learn the joint scene-event representation. When the number of shared blocks is 6, the learning of scene and event representations will be completely overlapping, then the model can only rely on the subsequent dense layer of each tower to learn the individual task-oriented representations.

As shown in Fig. 2, increasing the number of shared blocks does not consistently improve the classification accuracy of the model. The performance of the model peaks when the number of shared blocks is 2. Then, further increase in the number of shared blocks will degrade the performance of the model. That means the optimal structure of the two-tower model is the structure when the number of shared blocks is 2. Subsequent experiments will be conducted on this model structure.

• **RQ2**: How do different weights λ_i for the four losses influence the performance of the model?

During the training phase, different values of λ_i represent the difference in importance of scene information (L_{scene}), scene information inferred from the event prediction ($L_{\text{s\_by\_scene}}$), event information (L_{event}), and event information inferred from the scene prediction ($L_{\text{e\_b\_event}}$), respectively. Except for the scene information, the other three types of information are derived from pseudo labels. Since only the labels of scenes are available in the real-life dataset TUT2018, where the accuracy of the pseudo labels of events tagged by the pre-trained model PANN [16] cannot be evaluated due to the absence of reference event labels, and the goal of this paper is to improve the accuracy of scene classification, so the experiments will focus on the results of ASC.

In this section, the importance of different types of information is explored that corresponds to different λ for model learning. First, an ablation study has been conducted to compare the role of different types of information in the proposed RGASC. Table I lists the result of enabling and disabling certain parts of RGASC. For # 1 in Table I, only scene information is exploited, which is a pure ASC (puASC) model without the aid of additional information. # 2 predicts scenes by $\tilde{y}_s = \hat{y}_e \cdot R_{SE}^T$, which calculates the similarity between each row of the relation matrix R_{SE}^T and event information $\hat{y}_e$ provided by pseudo labels to derive the possibility of each scene. The R_{SE}^T with fixed prior knowledge can be viewed as a table. That is, # 2 in Table I can be regarded as the result of a lookup table that relies on the event information from pseudo labels, so its result is poor. # 3 uses the true information of scenes and pseudo information of events to obtain a shared joint scene-event representations extractor, which is similar to prior works [8] [9] [11]. # 4 relies only on the prior relation matrix to derive each other's outputs without learning the knowledge of scenes from the true labels and the knowledge of events from the pseudo labels. Without the support of accurate task-dependent representations of scenes and events, # 4, which only learns the implicit and intricate scene-event relation through the prior relation matrix, actually cannot learn the knowledge of scene classification, resulting in its inability to classify effectively, so its performance is poor. Compared with # 3, # 5 based on $L_{\text{e\_by\_scene}}$ expects to obtain more accurate and coordinated event-related scene prediction and enhance the discrimination ability of the scene branch, which in turn brings a better classification result. In contrast to # 5, # 6 attempts to obtain more accurate scene-related event information. # 5 outperforms # 6, which indicates that increasing the weight of scene information is more beneficial to the ASC task. The ablation study in Table I shows that the more the model pays attention to the scene-related information, the better its performance.

Second, fine-grained control of the weight of each loss function is explored. The fusion of different semantic information in Table II tries to adjust the weights of the other three kinds of information from pseudo labels to maximize their benefit. In other words, the right amount of noise-filled pseudo information needs to be introduced to help recognize

TABLE I: The ablation study of the proposed model.

#	L_{scene}	$L_{\text{s\_by\_event}}$	L_{event}	$L_{\text{e\_by\_scene}}$	*Acc (%)*
1	✔	✗	✗	✗	71.46
2	✗	✗	✔	✗	36.90
3	✔	✗	✔	✗	72.94
4	✗	✔	✗	✔	10.07
5	✔	✗	✔	✔	**75.80**
6	✔	✔	✔	✗	74.36

TABLE II: The effect of different λ_i values on the ASC task.

#	λ_1	λ_2	λ_3	λ_4	*Acc (%)*
1	1	0	1	0.01	75.92
2	1	0	0.5	0.01	76.27
3	1	0.1	1	0.1	75.05
4	1	0.1	0.1	0.1	75.62
5	1	0.1	0.5	0.1	75.79
6	1	0.01	0.5	0.01	***77.35***
7	1	0.01	0.01	0.01	76.48
8	1	0.001	0.01	0.001	76.71

TABLE III: Comparison of classification results of different systems on the development set of TUT2018.

System	Model structure	*Acc (%)*
PANN [16] (Fixed mode)	VGG-like CNN	56.9
Baseline [7]	CNN	59.7
CNN_Surrey [25]	CNN	68.0
NNF_CNNEns [28]	CNN and nearest neighbor filters	69.3
Attention [29]	CRNN with Self-attention	70.8
ABCNN [30]	Attention-Based CNN	72.6
PANN (Fine-tuning mode)	VGG-like CNN	73.8
MLTF [31]	CNN and SVM	75.3
Wavelet-based DSS [32]	CRNN	76.6
Proposed RGASC	VGG-like CNN	**77.4**

similar scenes. Finally, giving maximum weight to L_{scene} and secondary weight to L_{event}, while absorbing the scene-event relation information ($L_{\text{s\_by\_event}}$ and $L_{\text{e\_by\_scene}}$) with weaker weights, makes the best result of # 6 in Table II, which gradually merges the coarse-grained true information of scene and the fine-grained pseudo information of event based on scene-event relation matrix.

• **RQ3**: Does the proposed RGASC system in this paper perform better than other systems?

The published results of the DCASE2018 Task 1 Subtask A challenge are compared in this section in Table III. Only the non-ensemble methods are taken into consideration[5]. In addition to the well-performing convolutional neural networks (CNN) [7] [25], convolutional recurrent neural network (CRNN) [29] with self-attention works well, where self-attention [33] is used to model the relationship between different positions of sequences output by CRNN. Next, to achieve more optimal classification, an attention pooling layer is used in CNN to reduce the feature dimension [30]. On the other hand, in addition to the exploration of feature dimensions, the paper [31] proposes a CNN-based multiple layer temporal feature (MLTF) to try to capture the dynamic temporal information of audio signals efficiently. Furthermore, wavelet-based Deep Scattering Spectra (DSS) [32] is used to exploit higher-order information of acoustic features in the scene classification based on a CRNN model. For a comprehensive comparison, the pre-trained model PANN [16] used in this paper is also added to the comparison, and the performance of PANN is explored in two modes referring to the transfer learning [19]. In fixed mode, the parameters of PANN will not be updated during training, it will use the prior knowledge of 527 classes of events learned from Audioset to classify scenes. In fine-tuning mode, PANN will learn and absorb scene information based on the existing knowledge to update parameters.

[5]The non-ensemble results are obtained for comparison from the DCASE2018 Task 1 Subtask A (T1A) website. The proposed RGASC only uses a single model with one type of acoustic feature and does not involve any data augmentation methods, while the top 3 methods in T1A are mostly an ensemble of multiple models with multiple features, so the top 3 results are omitted in Table III. In detail, the Top-1 in T1A is an ensemble of 2 deep CNN trained with 11 types of acoustic features. Next, the Top-2 uses depth-wise separable CNN trained with 3 multi-scale acoustic features. Finally, the Top-3 is an ensemble of 6 big and deep models trained with 4 types of acoustic features. Therefore, the ensemble was performed on up to 24 models in total.

TABLE IV: Comparison of ASC results with other scene-event joint analysis methods on the development set of TUT2018.

#	Method	*Acc(%)*
1	Joint scene and event recognition [8]	52.35
2	Joint event and scene analysis using MTL [11]	61.69
3	Conditional scene and event recognition [14]	66.39
4	RGASC	**77.35**

Results in Table III show that the event knowledge has a certain ability to distinguish scenes. The score of fixed-mode PANN is close to Baseline, while the fine-tuned PANN gets a better result than the CRNN with self-attention. This indicates that even simple pure CNN without recurrent layers and diverse attention mechanisms can achieve promising results with the help of a large dataset (e.g. Audioset totals 5.8 thousand hours). Compared to other submissions in Table III, the proposed RGASC for recognizing coarse-grained acoustic scenes aided by relation-guided fine-grained event information is effective, even if the fine-grained event information is derived from pseudo labels without any verification.

Table IV also compares the RGASC with some existing methods for joint scene-event analysis. Among them, the first listed method [8] performs scene and event classification based on the same joint embedding space and scores the worst. This is easy to understand because real-life coarse-grained scenes and fine-grained events contain their own different characteristics and attributes. Then, the second-worst model [11] based on MTL [10] attempts to exploit both shared joint and separate individual representations of scenes and events. The third method [14] jointly analyses scenes and events based on the one-way scene-to-event conditional loss. The performance of [14] is better than that of [11], which indicates that the scene-conditioned loss plays the expected role. Overall, the proposed RGASC scores best out of the discussed joint analysis models of scenes and events.

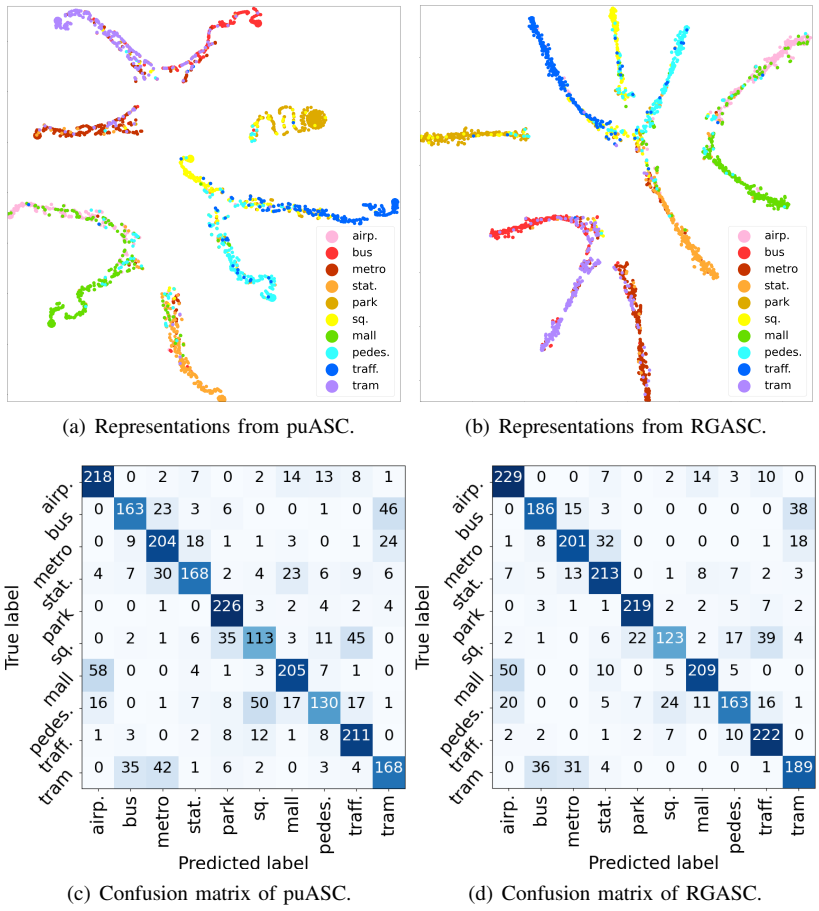
• **RQ4**: Does the introduction of event information from pseudo labels improve the recognition of acoustic scenes?

Table V specifically shows the role of pseudo-label information. The accuracy of every scene class is compared for puASC (# 1 in Table I) and the proposed RGASC (# 6 in Table II). RGSAC effectively improves the classification accuracy except for the two scenes of *park* and travelling by an underground metron (*metro*). In particular, the classification accuracy of the metro station (*stat.*) and pedestrian street (*pedes.*) is improved by 17.38% and 13.36%, respectively.

To gain deeper insights, Fig. 3 intuitively visualizes the gain of relation-guided pseudo-label information using t-SNE

TABLE V: Classification accuracy (%) on test set of the puASC with only true labels of scenes and the proposed RGASC aided by pseudo labels of events.

scene	*airp.*	*bus*	*metro*	*stat.*	*park*	*sq.*	*mall*	*pedes.*	*traff.*	*tram*
puASC	*82.26*	*67.35*	*78.16*	*64.86*	*93.38*	*52.31*	*73.47*	*52.63*	*85.77*	*64.36*
RGASC	*86.42*	*76.86*	*77.01*	***82.24***	*90.50*	*56.94*	*74.91*	***65.99***	*90.24*	*72.41*



(a) Representations from puASC. (b) Representations from RGASC.

(c) Confusion matrix of puASC. (d) Confusion matrix of RGASC.

Fig. 3: Visualization of high-level representations of acoustic scenes from models of puASC and RGASC using t-SNE [34] and the corresponding confusion matrices on the test set of the development set of TUT2018.

[34]. There are 9 sub-clusters in Fig. 3 (a) in the 10-class classification task. Many samples from similar scenes, like *bus* and *tram*, public square (*sq.*) and *park*, street traffic (*traff.*) and *park*, are mixed. Even for the human auditory system, relying on audio alone to distinguish these similar scenes is challenging [35]. The 10 classes of scenes are clearly shown in Fig. 3 (b) of RGASC. Even the *sq.* that are covered by other scenes in Fig. 3 (a) are delineated as a separate sub-cluster. The distinction between different scenes represented by different color sub-clusters is more obvious and the confusion is thus reduced, achieving a better classification result. This indicates that the relation-guided information fusion between the fine-grained event pseudo labels and the coarse-grained scene true labels works. In addition, in the bottom row of Fig. 3 the corresponding confusion matrices are presented.

The pseudo-labels of events used in this paper cannot ensure their accuracy due to the lack of ground-truth reference labels. However, the experimental results show that the imprecise pseudo-label information introduced by the relation matrix does boost the accuracy of scene classification. Pseudo labels without manual verification may prove their benefits, because pseudo labels can still depict the possibility of events in different scenes to some extent [36]. Therefore, the AEC tower in Fig. 1 trained based on pseudo labels can learn inaccurate but still effective event representations and distribution information, and transform this information into cues that can be applied to identify different scenes under the guidance of the relation matrix proposed in this paper, so as to enhance the recognition ability of the scene classification model. From the perspective of teacher-student learning [37], the pre-trained model PANN [16] used to output event pseudo labels in this paper can be regarded as the teacher model, and the AEC tower aiming to extract event information based on pseudo labels can be regarded as the student model. The teacher model outputs its rich knowledge of events into pseudo labels and transfers it to the student model [38]. Although the student model may not perform as well as the teacher model, the student model still has some discernment about events [39]. In this paper, the fine-grained event information learned by the student model will be used as the reference to correct the event information inferred by the ASC tower to enhance the scene branch's discrimination of diverse events within the scene, and further identify the differences within different scenes to boost the discrimination of the ASC tower for similar events. And then, the accuracy of scene classification is improved.

IV. CONCLUSION

Inspired by natural relations between real-life varied acoustic scenes and diverse events, this paper proposes to use the scene-event relation to guide the model to collaboratively classify scenes and events. The proposed relation-guided ASC (RGASC) framework effectively coordinates the coarse-grained true information of scenes and fine-grained pseudo information of events. Experiments show that the introduction of pseudo-label information performs well under the guidance of fixed prior relation matrix R_{SE}, and RGASC shows promising performance in differentiating similar polyphonic scenes.

Future work will enable the model to autonomously learn the R_{SE} during the training phase of model, and test it on more diverse datasets.

V. ACKNOWLEDGEMENTS

This research received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme.

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] H. Malik, "Acoustic environment identification and its applications to audio forensics," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 11, pp. 1827–1837, 2013.

[3] E. Principi, S. Squartini, R. Bonfigli, G. Ferroni, and F. Piazza, "An integrated system for voice command recognition and emergency detection based on audio signals," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5668–5683, 2015.

[4] N. Almaadeed, M. Asim, S. Al-Maadeed, A. Bouridane, and A. Beghdadi, "Automatic detection and classification of audio events for road surveillance applications," *Sensors*, vol. 18, no. 6, pp. 1858, 2018.

[5] 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, 2018.

[6] A. Mesaros, A. Diment, B. Elizalde, T. Heittola, E. Vincent, B. Raj, and T. Virtanen, "Sound event detection in DCASE 2017 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 6, pp. 992–1006, 2019.

[7] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Proceedings of DCASE Workshop*, 2018, pp. 9–13.

[8] 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.

[9] S. Chandrakala and S.L. Jayalakshmi, "Generative model driven representation learning in a hybrid framework for environmental audio scene and sound event recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 3–14, 2019.

[10] J. Jung, H. Shim, J. Kim, and H. Yu, "Dcasenet: An integrated pretrained deep neural network for detecting and classifying acoustic scenes and events," in *ICASSP*, 2021, pp. 621–625.

[11] N. Tonami, K. Imoto, R. Yamanishi, et al., "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.

[12] K. Imoto, N. Tonami, Y. Koizumi, M. Yasuda, R. Yamanishi, and Y. Yamashita, "Sound event detection by multitask learning of sound events and scenes with soft scene labels," in *ICASSP*, 2020, pp. 621–625.

[13] R. Shi, R. W. M. Ng, and P. Swietojanski, "Teacher-student training for acoustic event detection using audioset," in *ICASSP*, 2019.

[14] T. Komatsu, K. Imoto, and M. Togami, "Scene-dependent acoustic event detection with scene conditioning and fake-scene-conditioned loss," in *ICASSP*, 2020, pp. 646–650.

[15] Y. Kong, X. Wang, Y. Cheng, Y. Chen, and C. L. P. Chen, "Graph domain adversarial network with dual-weighted pseudo-label loss for hyperspectral image classification," *IEEE Geosci. Remote. Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[16] 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.

[17] J. F. Gemmeke, D. PW. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017, pp. 776–780.

[18] A. Bala, A. Kumar, and N. Birla, "Voice command recognition system based on MFCC and DTW," *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7335–7342, 2010.

[19] 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.

[20] J. Larsen and L. K. Hansen, "Generalization performance of regularized neural network models," in *Proceedings of IEEE Workshop on Neural Networks for Signal Processing*, 1994, pp. 42–51.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[22] J. Bjorck, C. Gomes, B. Selman, and K. Q. Weinberger, "Understanding batch normalization," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 7705–7716.

[23] 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.

[24] T. Kim, J. Oh, N. Kim, S. Cho, and S. Yun, "Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation," in *Proceedings of IJCAI*, 2021, pp. 2628–2635.

[25] Q. Kong, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "DCASE 2018 challenge surrey cross-task convolutional neural network baseline," in *Proceedings of the DCASE 2018 Workshop*, 2018, pp. 217–221.

[26] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

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

[29] J. Wang and S. Li, "Self-attention mechanism based system for dcase2018 challenge task1 and task4," *Proceedings of DCASE Challenge*, pp. 1–5, 2018.

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

[31] L. Zhang and J. Han, "Acoustic scene classification using multi-layered temporal pooling based on deep convolutional neural network," Tech. Rep., DCASE Challenge, 2018.

[32] Z. Li, L. Zhang, S. Du, and W. Liu, "Acoustic scene classification based on binaural deep scattering spectra with CNN and LSTM," Tech. Rep., DCASE Challenge, 2018.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[34] L. V. D. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[35] Andrew J. K. and Kerry M. W., "Listening in complex acoustic scenes," *Current Opinion in Physiology*, vol. 18, pp. 63–72, 2020.

[36] W. Hu, C. Chen, F. Ye, Z. Zheng, and Y. Du, "Learning deep discriminative representations with pseudo supervision for image clustering," *Information Sciences*, vol. 568, pp. 199–215, 2021.

[37] J. Bae, D. Yeo, J. Yim, N. Kim, C. Pyo, and J. Kim, "Densely distilled flow-based knowledge transfer in teacher-student framework for image classification," *IEEE Transactions on Image Processing*, vol. 29, pp. 5698–5710, 2020.

[38] H. Pham, Z. Dai, Q. Xie, and Q. V. Le, "Meta pseudo labels," in *Proceedings of the CVPR*, 2021, pp. 11557–11568.

[39] M. Noroozi, A. Vinjimoor, P. Favaro, and H. Pirsiavash, "Boosting self-supervised learning via knowledge transfer," in *Proceedings of the CVPR*, 2018, pp. 9359–9367.