



# Audio-Visual Scene Classification Based on Multi-modal Graph Fusion

Han Lei, Ning Chen\*

East China University of Science and Technology, Shanghai, China

y30200944@mail.ecust.edu.cn, chenning\_750210@163.com

## Abstract

Audio-Visual Scene Classification (AVSC) task tries to achieve scene classification through joint analysis of the audio and video modalities. Most of the existing AVSC models are based on feature-level or decision-level fusion. The possible problems are: i) Due to the distribution difference of the corresponding features in different modalities is large, the direct concatenation of them in the feature-level fusion may not result in good performance. ii) The decision-level fusion cannot take full advantage of the common as well as complementary properties between the features and corresponding similarities of different modalities. To solve these problems, Graph Convolutional Network (GCN)-based multi-modal fusion algorithm is proposed for AVSC task. First, the Deep Neural Network (DNN) is trained to extract essential feature from each modality. Then, the Sample-to-Sample Cross Similarity Graph (SSCSG) is constructed based on each modality features. Finally, the DynaMic GCN (DM-GCN) and the ATtention GCN (AT-GCN) are introduced respectively to realize both feature-level and similarity-level fusion to ensure the classification accuracy. Experimental results on TAU Audio-Visual Urban Scenes 2021 development dataset demonstrate that the proposed scheme, called AVSC-MGCN achieves higher classification accuracy and lower computational complexity than state-of-the-art schemes.

**Index Terms:** Audio-visual scene classification, Graph convolutional network, Similarity fusion

## 1. Introduction

Acoustic Scene Classification (ASC) tries to realize scene classification based on the analysis of recorded audio signal [1, 2, 3, 4, 5], which can be effectively applied to audio monitoring [6], smart wearable devices [7], and robot navigation, etc, has great research significance.

Since audio signal and video signal have complementary properties, the joint learning of them can enhance the performance in various tasks, such as source separation [8], audio-visual alignment for lip-reading [9] and environment recognition for mobile robot by previewing audio [10]. It is the same case in scene classification. In DCASE2021 challenge, Audio-Visual Scene Classification (AVSC) task was introduced for the first time. The performance of the AVSC model is highly dependent on the fusion strategy. In addition, the computational complexity of the AVSC model is another important factor that should be considered. Most of the existing AVSC models are based on feature-level fusion [11, 12, 13], or decision-level fusion [14], or the combination of them [15]. However, since the features extracted from different modalities have diverse characteristics and thus their distributions may be quite different, the direct concatenation of them may not achieve best fusion result. In addition, since the decision-level fusion without consider the correlations between feature sets, cannot take full advantage of

the common as well as complementary properties of features, it may not achieve satisfactory fusion results. Another shortcoming of feature-level and decision-level fusion is that they cannot full utilize the topological patterns of the audio sample or video sample communities to realize feature propagation.

To solve these problems, a new multi-modal graph fusion model is proposed for AVSC, named AVSC-MGCN. The Graph Convolutional Network (GCN) [16] is introduced in the AVSC task to fuse the features extracted from different modalities and the Sample-to-Sample Cross Similarity Graphs (SSCSGs) constructed based on different modality features. The merits of the GCN-based fusion strategy are as follows: i) It can realize the feature fusion and similarity fusion at the same time. ii) It can implement optimization on the multi-modal features. iii) It can take full advantage of the common as well as complementary properties of the topological patterns of the similarity graphs constructed based on different modality features. iv) It can be easily extended to fuse more modalities. v) The computational complexity of it is lower than the conventional fusion combination methods. Experimental results on TAU Audio-Visual Urban Scenes 2021 development dataset [13] demonstrated that no matter which GCN architecture, DynaMic GCN (DM-GCN) or ATtention GCN (AT-GCN) is adopted, the obtained AVSC scheme achieves higher classification accuracy and lower computational complexity than state-of-the-art AVSC schemes [13][15].

## 2. Proposed scheme

The new AVSC-MGCN model proposed in this paper is shown in Fig.1(a). It consists of three parts: multi-modal embeddings extraction, SSCSG construction based on K-Nearest Neighbor ( $k$ NN), and multi-modal graph fusion module. The following is a detailed explanation to these parts.

### 2.1. Multi-modal embeddings extraction

In this experiment, the Deep Neural Networks (DNNs) proposed by [15] are used to extract multi-modal embeddings. In the audio module, the 128-dimensional vector extracted from the last convolution layer through global average pooling is used as audio embedding. In the video module, we average the output features of Bi-GRU along the frame axis to get 64-dimensional visual embedding. The embeddings of all the samples in the audio modality is recorded as  $\mathbf{X}_a = \{\mathbf{x}_{a,i} | i = 1, \dots, N\}$ , where  $N$  is the total number of samples in the dataset and the video modality is recorded as  $\mathbf{X}_v$ .

### 2.2. SSCSG construction based on $k$ NN

In this paper, the SSCSG structure  $G_a = (\mathbf{V}_a, \mathbf{E}_a, \mathbf{A}_a)$  in the audio modality and  $G_v = (\mathbf{V}_v, \mathbf{E}_v, \mathbf{A}_v)$  in the video modality are constructed, where each sample in the dataset is regarded as a node, and all audio nodes constitute the node set  $\mathbf{V}_a$ . The audio edge set  $\mathbf{E}_a$  determines whether to form neighbors ac-

\*Corresponding author

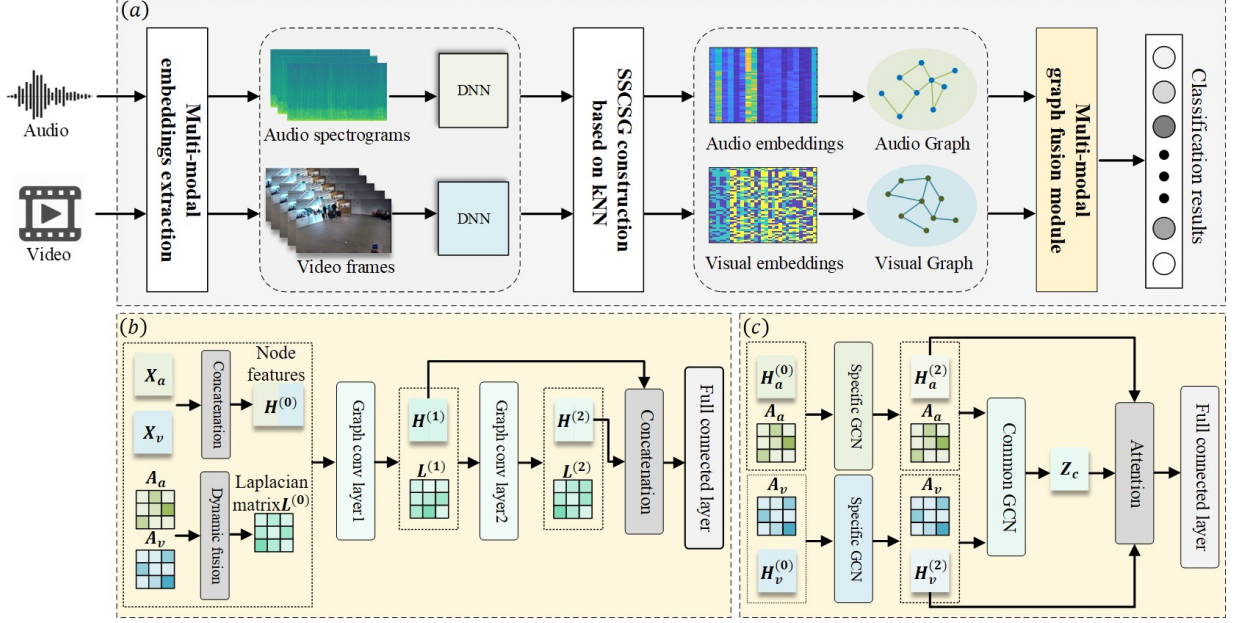


Figure 1: Block diagram of (a) AVSC-MGCN model and two graph fusion modules: (b) DM-GCN and (c) AT-GCN.

coding to the similarity degree of embeddings among samples, and the weight of each edge is represented by adjacency matrix  $\mathbf{A}_a = \{A_a(i, j) | i, j = 1, \dots, N\} \in \mathbb{R}^{N \times N}$ . The audio adjacency matrix  $A_a(i, j)$  indicates the similarity between the  $i$ -th sample and the  $j$ -th sample, see Eq.(1):

$$A_a(i, j) = \exp\left(-\frac{d(\mathbf{x}_{a,i}, \mathbf{x}_{a,j})}{\mu_a}\right) \quad (1)$$

where  $\mathbf{x}_{a,i}$  and  $\mathbf{x}_{a,j}$  are audio embeddings,  $d(\cdot, \cdot)$  is the Euclidean distance function, and  $\mu_a$  is the mean Euclidean distance of all audio embeddings.

Meanwhile, to mimic the sparsity of the cross-similarity graph and reduce the interference of noise,  $k$ NN is adopted to replace global similarity with local similarity. Only the  $k$ -top edges with high similarity of each sample are retained, so as to maximize the use of similarity between the same class, better optimize learning embeddings, and reduce the intra-class distance.  $k$  is set to 60 in our experiments.

### 2.3. Multi-modal graph fusion module

In this paper, we propose two graph fusion modules: DM-GCN and AT-GCN. These fusion modules perform nonlinear fusion from two angles of graph: the adjacency matrix and the node feature.

#### 2.3.1. DM-GCN

As shown in Fig1(b), this paper takes the dataset samples as the nodes of the graph, concatenates the embeddings from different modalities as the initial value of node features, and then fuses adjacency matrices from multiple modalities for similarity fusion. Considering that multiple modalities have different feature sets and contain different information. Some graph structures may contain misleading information, can easy to introduce noise, so using the same and static weight for each modality's adjacency matrix is not a good choice. Inspired by [17], the dynamic weighting method is adopted to construct specific struc-

tural information for each graph, assign a trainable weight to the adjacency matrix of each modality, and performs dynamic weighted Laplacian learning. The fused Laplacian matrix  $\mathbf{L}^{(t)}$  in the  $t$ -th layer is shown in Eq.(2):

$$\mathbf{L}^{(t)} = \theta_a^{(t)} \tilde{\mathbf{D}}_a^{-\frac{1}{2}} \tilde{\mathbf{A}}_a \tilde{\mathbf{D}}_a^{-\frac{1}{2}} + \theta_v^{(t)} \tilde{\mathbf{D}}_v^{-\frac{1}{2}} \tilde{\mathbf{A}}_v \tilde{\mathbf{D}}_v^{-\frac{1}{2}} \quad (2)$$

where  $\theta_a^{(t)} + \theta_v^{(t)} = 1$ ,  $\theta_a^{(t)}$  and  $\theta_v^{(t)}$  represents the trainable Laplacian weights in the  $t$ -th layer of audio and video respectively. Meanwhile, using the renormalization trick, defining  $\tilde{\mathbf{A}}_a = \mathbf{A}_a + \mathbf{I}_N$  and the diagonal matrix  $\tilde{D}_a(i, i) = \sum_j \tilde{A}_a(i, j)$ . Then we input the fused Laplacian matrix and the concatenated embeddings into GCN, it is formulated as:

$$\mathbf{H}^{(t+1)} = \sigma(\mathbf{L}^{(t)} \mathbf{H}^{(t)} \mathbf{W}^{(t)}) \quad (3)$$

where  $\mathbf{H}^{(t)}$  represents the embeddings of the  $t$ -th layer, and the initial value  $\mathbf{H}^{(0)}$  is the concatenation of  $\mathbf{X}_a$  and  $\mathbf{X}_v$ ,  $\mathbf{W}^{(t)}$  is the trainable weight matrix of the  $t$ -th layer,  $\sigma(\cdot)$  is the activation function. Generally,  $ReLU(\cdot) = \max(0, \cdot)$  is selected.

In order to make full use of the GCN embeddings from different layers, we concatenate the output embeddings of the first and second layer, so as to obtain the more scales and more stable feature, and then input it to the final Full Connected (FC) layer with the softmax activation function.

#### 2.3.2. AT-GCN

As shown in Fig1(c), its core idea is to aggregate and propagate specific and common node features in the two modalities, and then use attention mechanism to learn the adaptive importance weight of the above features.

First, we input  $\mathbf{A}_a$ ,  $\mathbf{X}_a$  and  $\mathbf{A}_v$ ,  $\mathbf{X}_v$  into two independent GCNs, respectively. Two specific graph convolution modules are used to extract Specific Graph Convolution (SpGC) features (see Eq.(4), taking audio modality as an example).

$$\mathbf{H}_a^{(t+1)} = \sigma(\tilde{\mathbf{D}}_a^{-\frac{1}{2}} \tilde{\mathbf{A}}_a \tilde{\mathbf{D}}_a^{-\frac{1}{2}} \mathbf{H}_a^{(t)} \mathbf{W}_a^{(t)}) \quad (4)$$

where,  $\mathbf{H}_a^{(t)}$  represents the audio SpGC feature of the  $t$ -th layer, the initial value  $\mathbf{H}_a^{(0)} = \mathbf{X}_a$ , and  $\mathbf{W}_a^{(t)}$  represents the specific trainable weight matrix of the audio modality of the  $t$ -th layer.

Considering the common characteristics between the two modalities, we design a common graph convolution module to extract Common Graph Convolution (CoGC) features using a parameter sharing strategy (see Eq.(5)) :

$$\mathbf{Z}_a^{(t+1)} = \sigma(\tilde{\mathbf{D}}_a^{-\frac{1}{2}} \tilde{\mathbf{A}}_a \tilde{\mathbf{D}}_a^{-\frac{1}{2}} \mathbf{Z}_a^{(t)} \mathbf{W}_c^{(t)}) \quad (5)$$

where  $\mathbf{Z}_a^{(t)}$  represents the audio CoGC feature of the  $t$ -th layer, the initial value  $\mathbf{Z}_a^{(0)}$  is the audio SpGC feature  $\mathbf{H}_a^{(T_s)}$ ,  $T_s$  is the number of SpGC layers,  $\mathbf{W}_c^{(t)}$  is the trainable weight matrix shared by the two modalities of the  $t$ -th layer. So we can obtain the CoGC features of the two modalities  $\mathbf{Z}_a^{(T_c)}$  and  $\mathbf{Z}_v^{(T_c)}$ .  $T_c$  is the number of CoGC layers. Finally, the common feature of the multi-modal can be expressed as follows:

$$\mathbf{Z}_c = \frac{\mathbf{Z}_a^{(T_c)} + \mathbf{Z}_v^{(T_c)}}{2} \quad (6)$$

After extracting the SpGC features  $\mathbf{H}_a^{(T_s)}$ ,  $\mathbf{H}_v^{(T_s)}$  in the audio and video modalities and the fused CoGC feature  $\mathbf{Z}_c$ , attention mechanism is used to automatically learn the importance weights of the above three features, and then perform the weighted summation method to extract the scene information with the strongest correlation. Finally, the feature with the most relevant information can be classified through the last FC layer with the softmax activation function.

### 3. Experiments

#### 3.1. Dataset and metrics

The experiments are conducted on the development dataset of TAU Audio-Visual Urban Scenes 2021 [13]. This dataset contains synchronized audio and video recordings from 12 European cities, and they were recorded in 10 different scenes. The dataset contains 34 hours of data with a training/test split of 7:3. The length of each sample is 10-second.

AVSC task uses two evaluation metrics: classification accuracy (Acc) and log multi-class cross-entropy loss (Log-loss). The value of classification accuracy is the ratio of the number of correctly classified samples to the total number of samples in the test subset. Log-loss is the log value of the average cross-entropy loss of all samples in the test subset, the smaller the value, the better.

#### 3.2. Experiment setup

For the acoustic signal, the Auditory Toolbox [18] is adopted to extract gammatone features, resample the audio sample from 48KHz to 44.1KHz with 64 frequency bands, 40ms window size, and 50% overlap. For the video signal, it is subsampled to the frame rate of 5 frames per second. Each frame is a color image of 224×224 pixels. To consist with the experimental setup in DCASE2021 task 1b, 1 second segment is randomly extracted from each sample in the training subset of TAU Audio-Visual Urban Scenes 2021 development dataset to obtain the training dataset of 8646 samples. The testing samples in the above dataset are split into segments of 1 second long without overlap to obtain the testing dataset, which is composed of 36450 samples. The audio and visual embeddings of total 45096 samples in both training set and testing set are adopted as the node embeddings of GCN. The number of units in the first and

second layer of GCN are set as 256 and 128, respectively. The dropout [19] rate for the first layer of GCN is 0.5. The Adam [20] with learning rate of 0.0005, weight decay of 0.0001, and maximum epoch of 500, is adopted as the optimizer. In addition, the mixup data augmentation [21] is adopted during audio and visual embedding extraction. The experimental results of [15] are obtained by us under the same experimental environment as ours. All the experiments are conducted on one GeForce RTX 2080Ti GPU.

Table 1: *The effectiveness of GCN in embedding optimization.*

Model	Audio-only		Visual-only	
	Log-loss	Acc	Log-loss	Acc
without GCN	0.962	68.4%	0.720	86.2%
with GCN	0.929	72.1%	0.592	88.1%

#### 3.3. Experimental results

##### 3.3.1. The effectiveness of GCN in embedding optimization

The superiority of GCN over traditional deep learning architecture is that it can take advantage of the topological characteristics of the SSCSG to optimize the node embeddings. To investigate the effectiveness of GCN in embedding optimization, the audio (or visual) embeddings extracted by the DNN module in Fig.1 are utilized to initialize the node embeddings of the GCN, which is composed of two layers of 128 units, and the SSCSG constructed based on DNN-extracted embeddings is fed to the GCN. In addition, the output embeddings of the first and second layers of GCN are concatenated to obtain the optimized embeddings and then passed through a FC layer with the softmax activation function. The visualization of the embeddings of testing samples obtained before and after GCN optimization for audio and video inputs are compared in Fig.2 by t-SNE [22]. It can be seen that, for both audio and video samples, after GCN-based optimization, the intra-class distances are reduced and the inter-class distances are enhanced at the same time. So the GCN-based optimization can enhance the distinctiveness of embedding effectively. The corresponding Log-loss and classification accuracy obtained by the models with or without GCN-based optimization are compared in Table 1. It is obvious that the introduction of GCN-based optimization helps to reduce the Log-loss and enhance the classification accuracy at the same time. So, it is verified that the introduction of GCN-based embedding optimization helps to enhance the scene classification performance.

Table 2: *Performance comparison on the development dataset.*

Model	Audio-Visual	
	Log-loss	Acc
[13]	0.658	77.0%
[15]	0.688	88.3%
Mean-GCN	0.358	90.4%
AT-GCN(ours)	0.304	90.7%
DM-GCN(ours)	0.329	91.3%

##### 3.3.2. The effectiveness of GCN-based graph fusion

To verify the superiority of GCN-based fusion over conventional fusion schemes in AVSC task, the official baseline of

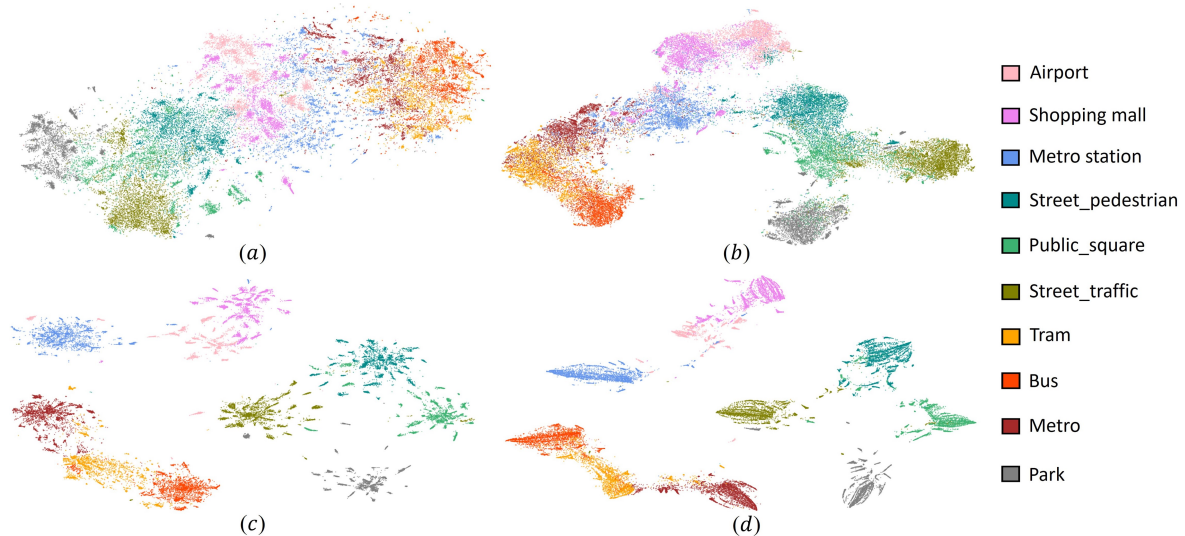


Figure 2: Visualization of embeddings by *t*-SNE. (a) and (b) are output embeddings before and after GCN optimization in audio modality, (c) and (d) are output embeddings before and after GCN optimization in video modality.

Table 3: Complexity comparison of fusion modules.

Model	Model Complexity	
	Parameters	Time
[15]	272k	89 hours
Mean-GCN	86k	5 hours
AT-GCN(ours)	182k	5 hours
DM-GCN(ours)	86k	5 hours

DCASE [13], which adopts concatenation-based feature fusion, and the scheme in [15], which combines bidirectional GRU-based feature fusion and decision-level fusion, are adopted as the baselines. To achieve fair performance comparison, the multi-modal features adopted by the proposed model are the same as those of [15]. In this paper, DM-GCN and AT-GCN modules are used for graph fusion respectively. Meanwhile, in order to study the advantages of the dynamic weighted adjacency matrix method based on DM-GCN, Mean-GCN is adopted. The difference between DM-GCN and Mean-GCN is that the latter uses the average adjacency matrix  $\mathbf{A} = \frac{1}{2}(\mathbf{A}_a + \mathbf{A}_v)$  as the input of GCN.

- *Classification accuracy and Log-loss comparison*

The performances of the graph fusion models based on GCN and those of the baseline [13][15] are compared in Table 2 in terms of Log-loss and classification accuracy. It can be seen that: i) All three GCN-based models outperform the baselines [13][15] in terms of Log-loss and classification accuracy. ii) AT-GCN performs better than DM-GCN in Log-loss, but lower than DM-GCN in classification accuracy. iii) DM-GCN performs better than Mean-GCN in terms of both Log-loss and classification accuracy, which proves that the dynamic weighting method can better integrate the similarity information of multi-modal. iv) Through comparing the results shown in Table 1 and those shown in Table 2, we can see that the GCN-based graph fusion models outperform the single GCN-based schemes for audio or video. The possible reason is that the GCN-based graph fusion can take full advantage of the common as well as complemen-

tary properties of multi-modal features.

- *Computational complexity comparison*

In this experiment, the computational complexity and the size of the parameters of each fusion schemes are compared (see Table 3). It can be seen that: i) The parameters and training time of the GCN-based graph fusion are much lower than the baseline [15]. ii) Among the GCN-based fusion models, the Mean-GCN and DM-GCN performs better than AT-GCN. iii) If both the classification accuracy and computational complexity are considered, DM-GCN is the best choice because it achieves the highest classification accuracy and lowest computational complexity and its Log-loss is a little bit higher than AT-GCN. In general, the GCN-based graph fusion scheme achieves higher efficiency than the baseline.

## 4. Discussion and conclusion

Experimental results shown in Table 2 demonstrate that AVSC-MGCN outperforms the baselines [13][15] in terms of Log-loss and Acc. The possible reasons are: on the one hand, since the adjacent matrix is constructed based on the feature similarities among samples, it's easier for the samples belonging to the same class to become neighbors, and the GCN can utilize the obtained graph topology to optimize the features further to reduce the intra-class distances and increase the inter-class distances as well. On the other hand, the graph-based fusion can achieve feature fusion and adjacent matrix fusion at the same time to take full advantage of the complementarity between different modalities. In addition, as shown in Table 3, the size of the parameters of the proposed model is much lower than that of [15]. As a result, the computational complexity of the proposed model is much lower too. The most important thing is that the proposed model can be flexible extended to fuse any number of multi-modal features to enhance the performance of AVSC task or those of other classification tasks.

## 5. Acknowledgements

This work was supported by the National Natural Science Foundation of China [grant number 61771196].

## 6. References

- [1] 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.
- [2] H. Hu, C.-H. H. Yang, X. Xia, X. Bai, X. Tang, Y. Wang, S. Niu, L. Chai, J. Li, H. Zhu *et al.*, "A two-stage approach to device-robust acoustic scene classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 845–849.
- [3] L. Pham, I. McLoughlin, H. Phan, and R. Palaniappan, "A robust framework for acoustic scene classification." in *INTERSPEECH*, 2019, pp. 3634–3638.
- [4] L. Pham, I. McLoughlin, H. Phan, R. Palaniappan, and A. Mertins, "Deep feature embedding and hierarchical classification for audio scene classification," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.
- [5] J.-w. Jung, H.-j. Shim, J.-h. Kim, and H.-J. Yu, "Dcasenet: An integrated pretrained deep neural network for detecting and classifying acoustic scenes and events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 621–625.
- [6] R. Radhakrishnan, A. Divakaran, and P. Smaragdis, "Audio analysis for surveillance applications," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005*. IEEE, 2005, pp. 158–161.
- [7] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: A survey," *IEEE communications surveys & tutorials*, vol. 16, no. 1, pp. 414–454, 2013.
- [8] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586.
- [9] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE, 2017, pp. 3444–3453.
- [10] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Content analysis for acoustic environment classification in mobile robots." in *AAAI Fall Symposium: Aurally Informed Performance*, 2006, pp. 16–21.
- [11] M. Wang, C. Chen, Y. Xie, H. Chen, Y. Liu, and P. Zhang, "Audio-visual scene classification using transfer learning and hybrid fusion strategy," DCASE2021 Challenge, Tech. Rep., 2021.
- [12] Q. Wang, S. Zheng, Y. Li, Y. Wang, Y. Wu, H. Hu, C.-H. H. Yang, S. M. Siniscalchi, Y. Wang, J. Du *et al.*, "A model ensemble approach for audio-visual scene classification," DCASE2021 Challenge, Tech. Rep., 2021.
- [13] S. Wang, A. Mesaros, T. Heittola, and T. Virtanen, "A curated dataset of urban scenes for audio-visual scene analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 626–630.
- [14] L. Pham, A. Schindler, M. Schütz, J. Lampert, S. Schlarb, and R. King, "Deep learning frameworks applied for audio-visual scene classification," *arXiv preprint arXiv:2106.06840*, 2021.
- [15] J. Naranjo-Alcazar, S. Perez-Castanos, M. Cobos, F. J. Ferri, and P. Zuccarello, "Task 1b dcase 2021: Audio-visual scene classification with squeeze-excitation convolutional recurrent neural networks," DCASE2021 Challenge, Tech. Rep., 2021.
- [16] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations (ICLR)*, 2017.
- [17] S. Li, W.-T. Li, and W. Wang, "Co-gcn for multi-view semi-supervised learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4691–4698.
- [18] M. Slaney, "Auditory toolbox," *Interval Research Corporation, Tech. Rep.*, vol. 10, no. 1998, p. 1194, 1998.
- [19] 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.
- [20] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [21] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.
- [22] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research (JMLR)*, vol. 9, no. 11, pp. 2579–2605, 2008.