# A GLOBAL-LOCAL ATTENTION FRAMEWORK FOR WEAKLY LABELLED AUDIO TAGGING

*Helin Wang*[1], *Yuexian Zou*[1,2,*], *Wenwu Wang*[3]

[1]ADSPLAB, School of ECE, Peking University, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
[3]Center for Vision, Speech and Signal Processing, University of Surrey, UK

## ABSTRACT

Weakly labelled audio tagging aims to predict the classes of sound events within an audio clip, where the onset and offset times of the sound events are not provided. Previous works have used the multiple instance learning (MIL) framework, and exploited the information of the whole audio clip by MIL pooling functions. However, the detailed information of sound events such as their durations may not be considered under this framework. To address this issue, we propose a novel two-stream framework for audio tagging by exploiting the global and local information of sound events. The global stream aims to analyze the whole audio clip in order to capture the local clips that need to be attended using a class-wise selection module. These clips are then fed to the local stream to exploit the detailed information for a better decision. Experimental results on the AudioSet show that our proposed method can significantly improve the performance of audio tagging under different baseline network architectures.

***Index Terms***— Audio tagging, weak labels, two-stream framework, class-wise attentional clips

## 1. INTRODUCTION

Audio Tagging is a technique for predicting the presence or absence of sound events within an audio clip [1]. The Detection and Classification of Acoustic Scenes and Events (DCASE) challenges [2, 3] provide strongly labelled datasets for audio tagging, where the onset and offset time of sound events are annotated. However, such annotation is time-consuming and hard to obtain, and these audio tagging datasets are relatively small. Recently, weakly labelled audio tagging has attracted increasing interest in the audio signal processing community [4, 5], where the datasets (*e.g.* AudioSet [6]) only annotate the types of sound events present in each audio clip but do not provide any timestamp information of their onset and offset.

As the duration of sound events can be very different and the overlaps of sound events often occur in an audio clip, audio tagging with weakly labelled data is a challenging problem. A popular approach for this problem is based on multiple instance learning (MIL) [7, 8]. In MIL, the input sequence is treated as a bag and split into the set of instances, where multiple instances in the same bag share the same labels. There are two main MIL strategies, *i.e.* instance-level approach [9] and embedding-level approach [10, 11]. The embedding-level approach integrates the instance-level feature representations into a bag-level contextual representation and then directly carries out bag-level classification, which shows better performance than the instance-level approach [12]. The methods for aggregating the information from the instances play an important part in the MIL frameworks. The default choices are global max pooling (GMP) [13] and global average pooling (GAP) [14], but they are often less flexible for adapting to practical applications and may lose detailed information relevant to acoustic events. For example, GMP cannot capture the information of the long-duration event *keyboard typing* well, while the short-duration event *mouse click* may be ignored by GAP. More recently, attention mechanisms have been employed to detect the occurence of sound events [15–18] and have achieved promising results. However, these methods attempt to make a decision on the whole audio clips and are limited in capturing the detailed information of the acoustic events.

To address this issue, in this paper, we propose a novel global-local attention (GL-AT) framework for weakly labelled audio tagging where the global and local information are successively modelled in the audio clip. Our method is inspired by the behaviors of human annotators for an audio dataset [6]. At first, the annotators may glimpse over an audio clip roughly, and determine some possible categories and their temporal regions. Then, these possible regions guide them to make refined decisions on specific categories following a region-by-region inspection. Thus, in a similar fashion, we can solve the audio tagging with a two-stream framework, including a global stream and a local stream. More specifically, the global stream takes an audio clip as the input
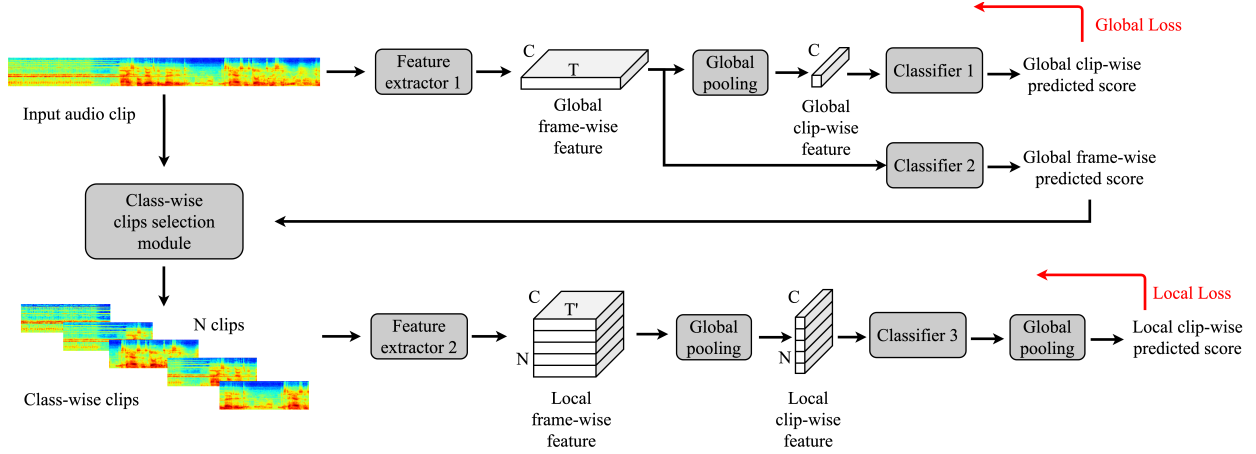
**Fig. 1**. Overall architecture of our two-stream framework for weakly labelled audio tagging (GL-AT).

to a deep neural network and learns global representations supervised by the weak labels. Several class-wise attentional sub-clips are then selected according to the global representations, and fed to another neural network to learn the local representations. The final class distributions are obtained by aggregating the predicted global class distributions and local class distributions. The optimization of the local stream is influenced by the global stream because the sub-clips are selected according to the prediction results of the global stream. In addition, the local stream improves the optimization of the global stream where sub-clip selection and classification are both performed.

The contributions of this paper can be summarized into three aspects. Firstly, we present a two-stream global-local attention framework that can efficiently recognize sound events within an audio clip with weakly labelled data. Secondly, we propose an effective class-wise sub-clip selection module which can dynamically generate several attentional sub-clips with low complexity and high diversity. Thirdly, experimental results show that our proposed framework can significantly improve the performance of AudioSet tagging, and can be used in different baselines.

## 2. PROPOSED METHOD

In this section, we firstly present the two-stream framework for weakly labelled audio tagging (*i.e.* GL-AT). Then, a class-wise clips selection module is presented, which bridges the gap between the global and local streams.

### 2.1. GL-AT

**Global Stream.** The overall framework of our proposed framework is shown in Fig. 1. Given an input audio clip $\boldsymbol{A} \in \mathbb{R}^{T_0 \times S}$, where $T_0$ and $S$ are the duration (*e.g.* 10s) and the sampling rate, respectively. Let's denote its corresponding

label as $\boldsymbol{y} \in \mathbb{R}^L$, where $y^i = \{0, 1\}$ denotes whether label $i$ appears or not and $L$ denotes the number of labels.

The feature extractor $\mathcal{F}(\cdot)$ is firstly applied, which can be either convolutional neural networks (CNNs) [19–23] or convolutional recurrent neural networks (CRNNs) [9]. We assume that $\boldsymbol{M} = \mathcal{F}(\boldsymbol{A}; \theta_{\mathcal{F}})$ is the global frame-wise feature after the feature extractor, where $\theta_{\mathcal{F}}$ denotes the parameters of the feature extractor and $\boldsymbol{M} \in \mathbb{R}^{T \times C}$. Here, $T$ and $C$ denote the number of output frames and the dimension of the features for each frame, respectively. Then a global pooling function $\mathcal{P}(\cdot)$ is applied to obtain the global clip-wise feature $\boldsymbol{M'} \in \mathbb{R}^{1 \times C}$. Following [22], both maximum and average operations are used for global pooling. In order to get the prediction score $\hat{\boldsymbol{y}} \in \mathbb{R}^L$, a classifier $\mathcal{C}(\cdot)$ containing two fully-connected layers is applied [22].

$$\hat{\boldsymbol{y}} = \mathcal{C}(\boldsymbol{M'}; \theta_{\mathcal{C}}) \tag{1}$$

where $\theta_{\mathcal{C}}$ denotes the parameters of the classifier. We then use a sigmoid function $\sigma(\cdot)$ to turn $\hat{\boldsymbol{y}}$ into a range $[0, 1]$, and obtain the global clip-wise prediction score $\hat{\boldsymbol{y}}_{\boldsymbol{g}} \in \mathbb{R}^L$.

$$\hat{\boldsymbol{y}}_{\boldsymbol{g}} = \frac{1}{1 + exp(-\hat{\boldsymbol{y}})} \tag{2}$$

**Local Stream.** Let $\{\boldsymbol{A}_1, \boldsymbol{A}_2, \cdots, \boldsymbol{A}_N\}$ be a set of $N$ local clips selected from the input audio clip $\boldsymbol{A}$. These local clips have the same duration (but are shorter than $\boldsymbol{A}$), and are fed to another feature extractor which has the same structure as the global stream. Then the local prediction scores $\{\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \cdots, \hat{\boldsymbol{y}}_N\}$ are obtained using (1) and (2). Finally, these local predicted scores are aggregated by the global pooling function:

$$\hat{\boldsymbol{y}}_{\boldsymbol{l}} = \mathcal{P}(\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \cdots, \hat{\boldsymbol{y}}_N) \tag{3}$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{l}} \in \mathbb{R}^L$ is the local clip-wise prediction score. Note that this two-stream framework can be trained end-to-end and

transferred easily to different feature extractor networks. During the training stage, these two streams are jointly trained. At the inference stage, we fuse the predictions from the global stream ($\hat{y}_g$) and the local stream ($\hat{y}_l$) with the global pooling function to generate the final prediction score of the audio.

**Two-stream Learning.** Given a training dataset $\left\{ \boldsymbol{A}^i, \boldsymbol{y}_i \right\}_{i=1}^{D}$, where $D$ denotes the number of training examples. $\boldsymbol{A}^i$ is the $i$-th audio clip and $\boldsymbol{y}_i$ represents its corresponding labels. The overall loss function of our two-stream learning is formulated as the sum of two streams,

$$\mathcal{L} = \mathcal{L}_g + \mathcal{L}_l \tag{4}$$

where $\mathcal{L}_g$ and $\mathcal{L}_l$ represent the global and the local loss, respectively. Specifically, the binary cross entropy loss is applied for both streams,

$$\mathcal{L}_g = \sum_{i=1}^{D} \sum_{j=1}^{L} y_i^j \log\left( \hat{y}_{gi}^j \right) + \left( 1 - y_i^j \right) \log\left( 1 - \hat{y}_{gi}^j \right) \tag{5}$$

$$\mathcal{L}_l = \sum_{i=1}^{D} \sum_{j=1}^{L} y_i^j \log\left( \hat{y}_{li}^j \right) + \left( 1 - y_i^j \right) \log\left( 1 - \hat{y}_{li}^j \right) \tag{6}$$

where $\hat{y}_{gi}^j$ and $\hat{y}_{li}^j$ are the prediction scores of the $j$-th category of the $i$-th audio clip from the global stream and local stream, respectively. The Adam [24] is employed as the optimizer.

## 2.2. From Global to Local

Potential audio clips are not available in weak labels, and in this paper, we propose a simple but efficient method to dynamically generate candidate audio clips following two basic principles. On the one hand, the diversity of candidate clips should be as high as possible to cover all possible sound events within an audio clip. On the other hand, the number of candidate clips should be as small as possible to reduce computational complexity and storage space.

In order to generate the candidate clips, we firstly calculate the class-wise activation. A classifier is directly applied to the global frame-wise feature $\boldsymbol{M}$, and the global frame-wise predicted score $\boldsymbol{S}_g \in \mathbb{R}^{T \times L}$ can be obtained using (1) and (2), where $T$ denotes the number of frames within an audio clip. The class-wise activation of the $i$-th category is denoted as $S_g^i \in \mathbb{R}^T$, which indicates the importance of the frames leading to the classification of an audio clip to class $i$.

Such class-wise activation $\boldsymbol{S}_g$ is discriminative among different categories, and we can employ the activation to localize the potential candidate clips. However, there are often a large number of categories (*e.g.* 527 categories in AudioSet), and only a small number of categories (less than 10) appear in an audio clip. If the activation of all the categories is used, the generated clips are too many to be computationally efficient. To address this issue, we sort the predicted score $\boldsymbol{S}_g$ in a descending order and select the top $N$ class-wise attentional

activation (denoted as $\left\{ S_g^i \right\}_{i=1}^{N}$). $N$ is a hyperparameter and the choice of $N$ is discussed in our experiments.

The value of $S_g^i(t)$ represents the probability that the sub-clip belongs to the $i$-th category at timestamp $t$. In order to localize the clips of interest with low computational complexity, we employ a temporal window of size $\tau$ to select the candidate clips. For each $S_g^i(t)$, the frame with the maximum activation (denoted as $m$) is set as the medium frame, and the range of the candidate clip is set as $[m - \tau/2, m + \tau/2]$. If the maximum boundary of the candidate clip is over the duration of the audio clip, the range is re-weighted to $[T_0 - \tau, T_0]$, where $T_0$ denotes the duration of the audio clip. In addition, if the minimum boundary is less than $0$, we re-weight the range to $[0, \tau]$, where $\tau$ is a hyperparameter and discussed in our experiments.

## 3. EXPERIMENTS

In this section, we report the experimental results and comparisons that demonstrate the effectiveness of the proposed method. In addition, ablation studies are carried out to show the contribution of the crucial components.

### 3.1. Experimental Setups

**Dataset.** Audioset [6] is used in our experiments, which is a large-scale dataset with over 2 million 10-second audio clips from YouTube videos, with a total of 527 categories. The same dataset divisions and pre-processing approaches (*e.g.* re-sampling and data-balancing) are applied as [22].
**Metrics.** Mean average precision (mAP), mean area under the curve (mAUC) and d-prime are used as our evaluation metrics, which are the most commonly used metrics for audio tagging. These metrics are calculated on individual categories and then averaged across all categories.
**Implementation Details.** We compare the proposed method with the recent state-of-the-arts including TALNet [9], CNN10 [22], ResNet38 [22] and AT-SCA [25]. Specifically, all these models are applied as the global stream of our framework, and the local stream uses the same feature extractor and classifier as in the global stream. Thus, we obtain the results of the models with our two-stream learning framework, which are TALNet⊕GL-AT, CNN10⊕GL-AT, ResNet38⊕GL-AT and AT-SCA⊕GL-AT. Batch size is set as 32 and all networks are trained with 800k iterations in total. Unless otherwise stated, we set the hyperparameters $N$ as 5 and $\tau$ as 3s in our experiments.

### 3.2. Comparsions with the State-of-the-Arts

Table 1 demonstrates the performance of our proposed method (GT-AL[1]) and other state-of-the-art methods on

---

**Table 1**. Accuracy comparisons of our method and state-of-the-arts on the AudioSet.

| Method | mAP | mAUC | d-prime |
|---|---|---|---|
| TAL Net (2019) [9] | 0.362 | 0.965 | 2.554 |
| TAL Net* | 0.368 | 0.967 | 2.600 |
| **TALNet⊕GL-AT (ours)** | **0.401** | **0.970** | **2.659** |
| CNN10 (2019) [22] | 0.380 | 0.971 | 2.678 |
| CNN10* | 0.382 | 0.969 | 2.664 |
| **CNN10⊕GL-AT (ours)** | **0.408** | **0.974** | **2.742** |
| ResNet38 (2019) [22] | 0.434 | 0.974 | 2.737 |
| ResNet38* | 0.429 | 0.974 | 2.713 |
| **ResNet38⊕GL-AT (ours)** | **0.438** | **0.975** | **2.774** |
| AT-SCA (2020) [25] | 0.390 | 0.970 | 2.652 |
| AT-SCA* | 0.392 | 0.969 | 2.658 |
| **AT-SCA⊕GL-AT (ours)** | **0.413** | **0.971** | **2.677** |

*The listed results of TAL Net, CNN10, ResNet38 and AT-SCA are reproduced, and all the experimental setups are the same as the original papers [9, 22, 25].

**Table 2**. Ablative study of two streams in CNN10⊕GL-AT

| Method | Global | Local | mAP | mAUC |
|---|---|---|---|---|
| CNN10 | √ | | 0.382 | 0.969 |
| | √ | | 0.389 | 0.970 |
| CNN10⊕GL-AT | | √ | 0.400 | 0.972 |
| | √ | √ | 0.408 | 0.974 |

the Audioset. The results indicate that the proposed GT-AL can significantly improve the performance of all the compared methods, which confirms the effectiveness of our two-stream learning framework. Among them, TALNet [9] is a CRNN-based model, and CNN10 [22] and ResNet38 [22] are CNN-based models. AT-SCA [25] applies the spatial and channel-wise attention to detect sound events by learning what and where to attend in the signal. However, all these methods attempted to make a decision once and aggregated the global information by the MIL pooling functions. The aggregation often cannot make good use of the detailed information because of the variability of different sound events. By comparison, our method captures the discriminative local information from the global information and then checks attentionally the local information for a better decision, which provides a more efficient way to exploit the weak labels.

### 3.3. Ablation Study

In order to explore the effectiveness of the two streams, we jointly train the global and local streams in GL-AT, and during the inference stage, the influence of each stream is demonstrated in Table 2. Thanks to the joint training strategy, GL-AT outperforms the baseline method (CNN10) with only the
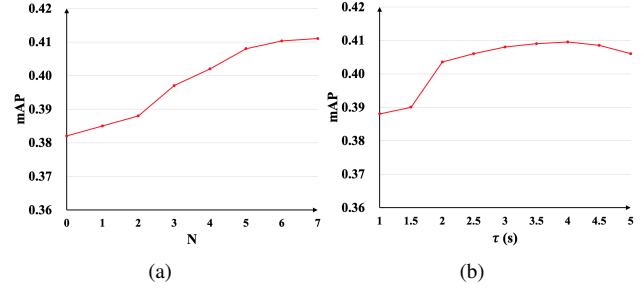


(a)  (b)

**Fig. 2**. Accuracy comparisons of CNN10⊕GL-AT with different values of $N$ and $\tau$ (metric: mAP)

global stream. This is because the class-wise attentional clips selection module connects the two streams, and the optimization of the global stream is influenced by the local stream, which leads to better robustness. In addition, we can see that using the local stream alone performs better than only using the global stream, for the reason that the local stream is able to focus on the detailed information of the audio. Nonetheless, the global stream plays an important role in guiding the learning of the local stream, and employing both global and local streams achieves the best results in our experiments.

Furthermore, we exploit the influence of the number of the local clips ($N$) and the duration of the local clips ($\tau$). As shown is Fig. 2(a), the mAP performance shows an upward trend with the gradual increase in $N$. This means that it is useful to improve the audio tagging performance with more local clips. However, more clips will increase the computational cost because all the selected clips are fed to the networks. We can see that the performance tends to be stable when $N$ is set to $\{5, 6, 7\}$, and we set $N$ to $5$ for a balance of accuracy and complexity. In addition, we test different $\tau$ values, and report the results in Fig. 2(b). As the values of $\tau$ increase, the accuracy is boosted and then drops, which achieves high accuracy when $\tau$ is $[3, 4]$ (similar to [26]). We argue that the length (*i.e.* about $3s$) is the duration of most sound events. If the duration of the clips is too short (*e.g.* $1s$), the complete information of a sound event cannot be captured. While on the other hand, if the duration is too long (*e.g.* over $4s$), the detailed information will be compromised. The extreme is that when the duration is $10s$ (the whole length of the audio clip), the local stream works as the global stream, which does not capture the detailed information intended.

## 4. CONCLUSIONS

We have presented a two-stream framework to take advantage of the global and local information of audio, which resembles the multi-task learning principle. Experimental results on the Audioset show that our method can boost the performance of different state-of-the-art methods in audio tagging.

# 5. REFERENCES

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

[2] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP Challenge," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[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] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018.

[5] N. Turpault, R. Serizel, A. P. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

[6] J. F. Gemmeke, D. P. 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[7] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[8] S.-Y. Tseng, J. Li, Y. Wang, F. Metze, J. Szurley, and S. Das, "Multiple instance deep learning for weakly supervised small-footprint audio event detection," in *Proc. Interspeech*, 2018, pp. 3279–3283.

[9] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 31–35.

[10] C.-C. Kao, M. Sun, W. Wang, and C. Wang, "A comparison of pooling methods on LSTM models for rare acoustic event classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 316–320.

[11] L. Lin, X. Wang, H. Liu, and Y. Qian, "Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1466–1478, 2020.

[12] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern Recognition*, vol. 74, pp. 15–24, 2018.

[13] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[14] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[15] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, and M. D. Plumbley, "Weakly labelled audioset tagging with attention neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1791–1802, 2019.

[16] C. Yu, K. S. Barsim, Q. Kong, and B. Yang, "Multi-level attention model for weakly supervised audio classification," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018.

[17] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *35th International Conference on Machine Learning (ICML)*. International Machine Learning Society (IMLS), 2018, pp. 3376–3391.

[18] H. Wang, Y. Zou, D. Chong, and W. Wang, "Environmental sound classification with parallel temporal-spectral attention," in *Proc. Interspeech*, 2020, pp. 821–825.

[19] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[20] K. Choi, G. Fazekas, and M. Sandler, "Automatic tagging using deep convolutional neural networks," in *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*, 2016, pp. 805–811.

[21] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 121–125.

[22] 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.

[23] L. Ford, H. Tang, F. Grondin, and J. Glass, "A deep residual network for large-scale acoustic scene analysis," in *Proc. Interspeech*, 2019, pp. 2568–2572.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR)*, 2015.

[25] S. Hong, Y. Zou, W. Wang, and M. Cao, "Weakly labelled audio tagging via convolutional networks with spatial and channel-wise attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 296–300.

[26] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 1041–1044.