

Multi-Scale Time-Frequency Attention for Acoustic Event Detection

Jingyang Zhang¹, Wenhao Ding¹, Jintao Kang^{2,3}, Liang He¹

¹Department of Electronic Engineering, Tsinghua University, China

²Institute of Forensic Science, Ministry of Public Security, China

³Key Laboratory of Intelligent Speech Technology, Ministry of Public Security, China

jz288@duke.edu, dingwenhao95@gmail.com, kangjintao@cifs.gov.cn, heliang@mail.tsinghua.edu

Abstract

Most attention-based methods only concentrate along the time axis, which is insufficient for Acoustic Event Detection (AED). Meanwhile, previous methods for AED rarely considered that target events possess distinct temporal and frequential scales. In this work, we propose a *Multi-Scale Time-Frequency Attention* (MTFA) module for AED. MTFA gathers information at multiple resolutions to generate a time-frequency attention mask which tells the model where to focus along both time and frequency axis. With MTFA, the model could capture the characteristics of target events with different scales. We demonstrate the proposed method on Task 2 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 Challenge. Our method achieves competitive results on both development dataset and evaluation dataset.

Index Terms: multi-scale learning, time-frequency attention, acoustic event detection, DCASE2017 Challenge

1. Introduction

Acoustic Event Detection (AED) defines a task which requires identification of whether some certain events occur in an audio clip, and their timestamps should be given if they do exist. AED has drawn great attention among researchers since it is crucial for a system especially when visual cues are not sufficient or not available. For example, an autonomous car should be able to detect sirens to give way for an ambulance approaching from behind [1]. Household robots [2] and road surveillance systems [3] can also benefit from AED.

To facilitate the research of AED, Task 2 of Detection and Classification of Acoustic Scenes and Events (DCASE) 2017 Challenge [4] provides datasets and a well-designed testbench. The task asks participants to locate the onset of three classes of rare events (*babycry*, *glassbreak*, *gunshot*) within audio clips.

One straightforward approach to this is making frame-level predictions, which was adopted by many participants including 1st and 2nd place [5, 6] of DCASE2017 Task2. Kao *et al.* [7], on the contrary, applied *Region Proposal Network* [8] to perform event-level detection and directly generate the onset/offset of target events. However, this method needs pre-training and suffers from over-fitting due to a large number of parameters. Wang *et al.* [9] further incorporated attention mechanism to derive utterance-level predictions from frame-level output, attempting to leverage attention to improve frame-wise detection.

Following the majority of previous works, we produce frame-level predictions in this work. As attention mechanism has achieved excellent performance in various tasks like emotion recognition [10], conversational analysis [11], and speech recognition [12], we assume it can also benefit AED as it helps

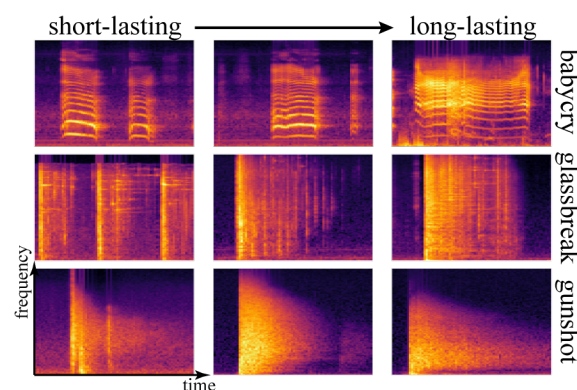


Figure 1: *Mel-spectrogram* of the events in DCASE2017 Task2.

the model focus on when target events take place. However, [9] and most attention-based methods operate only along the time axis, which is insufficient for AED because events present different frequential characteristics. It can be seen in Figure 1 that while *babycry* and *glassbreak* cover a wider range in the frequency domain, *gunshot* consists of high-frequency components only at the beginning and soon low-frequency energies become dominant. This variation makes it necessary to tell the model where to focus along the frequency axis.

Besides, previous methods did not take into account that target events possess different scales. As shown in Figure 1, the duration or speed varies among not only different types of events, but also different samples within each class. Meanwhile, frequential scales might also change. For example, different types of guns present contrasting characteristics. Thus, if the model perform on a single resolution, it might miss useful information to capture the distinct properties of each event.

From the above two perspectives, inspired by *Residual Attention* [13], we propose a *Multi-Scale Time-Frequency Attention* (MTFA) module for AED. MTFA fuses multi-scale information to generate a 2D time-frequency attention heatmap and produces attention-aware representations. The generated attention weights instruct the model to focus on not only important frames along the time axis but also import components along the frequency axis. To sum up, MTFA is designed to extract powerful features by leveraging multi-scale learning and attention mechanism. To the best of our knowledge, the proposed method outperforms previous single-model methods on development dataset of DCASE2017 Task2 and achieves state-of-the-art on evaluation dataset by reducing the error rate to 0.09 from previously best 0.13. The result provides clear evidence that MTFA has a strong ability to obtain discriminative representations.

Liang He is the corresponding author.

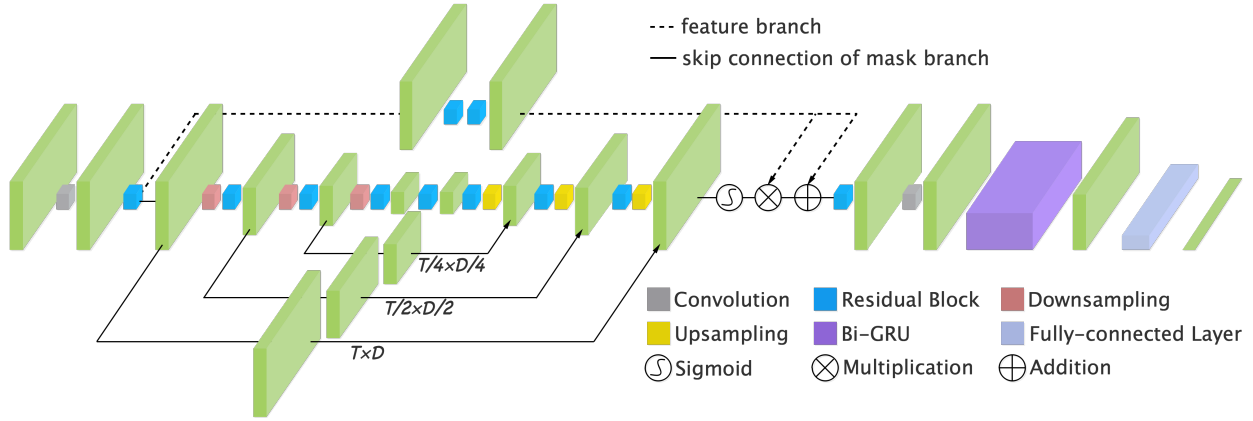


Figure 2: **Global structure of the proposed method.** Each green block represents a feature map or the raw input. MTFA consists of a **feature branch** for feature extraction and a **mask branch** for attention generation. The output of two branches is combined in a residual manner to produce multi-scale attention-aware features. A bi-directional GRU is applied to capture temporal dependencies between frames on top of MTFA. Finally, fully-connected layers will give frame-wise predictions.

2. Method Description

We present the overall model structure in Figure 2. The network mainly consists of three modules: a *Multi-Scale Time-Frequency Attention* module, a Recurrent Neural Network, and a classifier. The MTFA extracts multi-scale attention-aware features with a focus on important frames and frequential components. The RNN further captures temporal information. The classifier produces frame-level predictions which indicate whether the frame corresponds to the event.

2.1. Input Acoustic Features

Similar to previous works [5, 6, 7, 9, 14], we use log-scale filter banks to obtain Mel-spectrogram as the raw input. Specifically, we first apply a 40ms sliding window with a 20ms shift over the audio signal and then calculate 128-dimensional log filter bank energies for each frame. Finally, we will get a 2D time-frequency representation with a size of (1501×128) for each 30-second sound clip.

2.2. Multi-Scale Time-Frequency Attention

The intention of MTFA are two folds: (1) gathering information at multiple resolutions to model different scale characteristics presented by target events and (2) producing both temporal and frequential attention weights to focus on not only important frames but also discriminative frequency components. Note that these two ideas are not separate or independant.

First, we believe the fusion of multi-scale information is beneficial for AED. As reported by [14], the average duration of the events are: 2.25s for *babycry*, 1.16s for *glassbreak*, and 1.32s for *gunshot*. This is quite intuitive as the latter two events mainly consist of short-time impulses, and *babycry* generally lasts for a longer time. If we observe on one single time resolution, it might be insufficient to model either the long-lasting *babycry* or the relatively transient *glassbreak* and *gunshot*. [9] already extracted multi-resolution features along the time axis to improve the robustness against this time variation. However, frequency variation also exists in AED, as demonstrated by Figure 1. Due to the limitation of human sounds, low-frequency components are the major part of *babycry*, but *glassbreak* has nearly every frequency ingredient. *gunshot* differs a lot from

them as it starts with high-frequency impulses and vanishes with low-frequency energies. The change of gun types may also results in distinct properties. Therefore, in this work, we design MTFA to extract representations from multiple scales in both the time and frequency domain.

Second, to help the model focus more on the timestamp of target events and meanwhile pick out discriminative frequency components, we propose to generate both temporal and frequential attention. Specifically, by utilizing convolution as the basic operator, our method takes in a spectrogram and generates a 2D time-frequency attention heatmap, which contains element-wise attention weights. In other words, MTFA leverages the 2D nature of convolution and time-frequency features, thus enabling itself to take care of temporal and frequential attention simultaneously. Besides, the multi-channel nature of convolution provides the model with the capacity of collecting richer information to generate attention.

Specifically, our MTFA module is inspired by *Residual Attention* [13] which was proposed for image classification. The proposed MTFA includes a *feature branch* for feature extraction and a *mask branch* for attention generation.

We choose ResNet [15] consisting of two residual blocks to construct the *feature branch*. The kernel size of convolution within all residual blocks is set to 3×3 because small kernel fuses less background with target events and is more sensitive to the onset. For the *mask branch*, we utilize a bottom-up top-down structure called Hourglass [16] which has shown its effectiveness in human pose estimation and image segmentation. Hourglass aims to fuse information from multiple resolutions, which aligns with our intention of multi-scale learning. While encoding coarse-grained information by downsampling feature maps to lower and lower resolutions along both time and frequency axis, Hourglass uses skip connection to maintain fine-grained information as well (the residual block within each skip connection is not shown in Figure 2 for brevity). In this work we use three Max-pooling layers with kernel size and stride of 2×2 for downsampling, resulting in an observation at four different scales: $(T \times D)$, $(T/2 \times D/2)$, $(T/4 \times D/4)$, $(T/8 \times D/8)$, where T and D correspond to the length of MTFA's input along time and frequency axis, respectively. Then, low-resolution feature maps are upsampled back to the original scale using Nearest-Neighbor interpolation, and then merged with high-

resolution representations which come through skip connections. Finally, the Sigmoid function is applied to generate an attention heatmap whose elements range within $[0, 1]$.

Suppose the input of MTFA has a size of $(C \times T \times D)$, where C denotes the number of feature channels within MTFA, and it is shared by the *feature branch* and the *mask branch*. Then, the output M of the *mask branch* has the same size as the representation F produced by the *feature branch*, which is also $(C \times T \times D)$. To get attention-aware features, one can simply perform element-wise multiplication between M and F , i.e., $M \otimes F$. However, since M ranges from 0 to 1, this naive production might degrade the values and make it difficult for the following propagation. Thus, following [13], we use *residual connection* [15] and set the output A of MTFA to be

$$A(x) = (1 + M(x)) \otimes F(x), \quad (1)$$

where x represents the input of MTFA module.

2.3. Recurrent Neural Network

We use a RNN to capture temporal dependencies between frames, which has been proved to work well when being combined with the CNN architecture. Here we use a bi-directional *Gated Recurrent Unit* (GRU) which has U hidden units and two layers, leading the size of the final high-level feature map to be $(T \times 2U)$. Fully-connected layers and sigmoid function are applied to output frame-wise predictions.

3. Experiments

3.1. Dataset

We evaluate the proposed method on Task 2 of DCASE2017 Challenge [4]. The dataset consists of 30-second sound clips, each of which has one isolated target event (either *babycry*, *glassbreak*, or *gunshot*) along with recordings of different acoustic scenes taken from TUT Acoustic Scenes 2016 [17] as the background. We use the provided synthesizer to generate mixtures with different event-to-background ratios and random onset time. For the train set of development dataset, we generate 5000 mixtures for each target event with event-to-background ratios of -6, 0, 6dB and event presence probability of 0.99 in order to have more positive samples. For the test set of development dataset, 500 mixtures for each event are generated following the default parameters for fair comparison. The training data for development and evaluation dataset is the same.

3.2. Experiment Setup

During the training phase, the model takes a chunk of the spectrogram as input, i.e., several consecutive frames, rather than the whole spectrogram to increase batch size and training speed. In our experiments, the temporal length T is set to 256 and the shift between each chunk is 128. During the inference phase, instead, we pass the spectrogram as a whole into the network. Since the input will be downsampled trice during propagation, we pad the spectrogram with the last frame to satisfy this length constraint. The number of channels within the MTFA module C is set to 64, and the number of GRU units U is also 64. We apply dropout [18] with a rate of 0.3 for *babycry* and *glassbreak*, and 0.4 for *gunshot* as we observe more severe overfitting with it.

The task could be regarded as a 0/1 classification since our method makes frame-wise predictions. Therefore, we use binary cross-entropy as the loss function. We use Adam [19] as

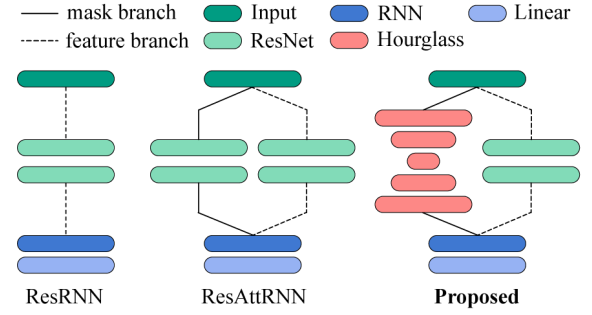


Figure 3: A simple overview of structures for *ablation study*.

the optimizer with a learning rate of 0.001. The training is terminated when the validation loss has stopped improving for 10 epochs, and all experiments finished training within 40 epochs according to our observation. The model with the lowest validation loss is selected for the test. We implement our models with PyTorch [20].

When the model is to be evaluated, the output occurrence probabilities are binarized with thresholds of 0.4, 0.2, and 0.4 for *babycry*, *glassbreak*, and *gunshot*, respectively. The binary predictions are further processed with a 540ms median filter, which follows the default setting of DCASE2017 baseline.

3.3. Evaluation Metrics

Event-based error rate and F1-score are two metrics used for DCASE2017 Task2. They are calculated using onset-only condition with a collar of 500ms. The evaluation is done automatically using `sed_eval` toolbox provided by the task organizers. Details can be found in [21].

3.4. Results

Ablation studies. Two more models are studied here to identify the contribution of the proposed MTFA. We construct the first one by removing the *mask branch* from MTFA and leaving only the *feature branch*, which is termed ResRNN. The second one is built by replacing the Hourglass in the *mask branch* with a ResNet, and we name it ResAttRNN. The general structures are shown in Figure 3. In a nutshell, while ResRNN does not perform any attention techniques, ResAttRNN generates temporal-frequential attention at only one single scale. Results are presented in the first three rows of Table 1. While ResRNN reaches comparable results to those of ResAttRNN on development dataset, its performance degrades a lot on evaluation dataset and is worse than ResAttRNN, indicating that temporal-frequential attention indeed helps to extract powerful features which could generalize better. Furthermore, compared with ResAttRNN, MTFA performs better on both development and evaluation dataset, confirming that multi-scale learning can further improve the performance.

Comparison with previous methods. We present results in Table 1. On development dataset, the proposed method achieves equally low error rate as R-CRNN [7] on *gunshot* and also gets desirable results on *babycry* and *glassbreak*. Note that 1D-CRNN [5] adopted ensemble strategy and fused the output of up to 5 models to get final predictions. However, our method could perform comparably well without ensemble and achieves the best average result compared with other single-model methods like [7, 9] and especially [14] which also generated temporal-frequential attention but in a different manner.

Table 1: Performance comparison between the proposed method and previous ones. Event-based error rate and F1-score are evaluated following the setting of DCASE2017. *** indicates the corresponding results are not presented in the papers.

Method	Development Dataset				Evaluation Dataset			
	babycry ER F1	glassbreak ER F1	gunshot ER F1	average ER F1	babycry ER F1	glassbreak ER F1	gunshot ER F1	average ER F1
ResRNN	0.09 95.8	0.03 98.4	0.17 91.0	0.10 95.1	0.22 89.3	0.07 96.4	0.24 87.4	0.18 91.0
ResAttRNN	0.13 93.4	0.02 98.8	0.18 90.4	0.11 94.2	0.22 88.9	0.04 97.8	0.17 91.2	0.14 92.6
Proposed	0.06 96.7	0.02 99.0	0.14 92.7	0.07 96.1	0.10 95.1	0.02 98.8	0.14 92.5	0.09 95.5
1D-CRNN [5]	0.05 97.6	0.01 99.6	0.16 91.6	0.07 96.3	0.15 92.2	0.05 97.6	0.19 89.6	0.13 93.1
CRNN [6]	*****	*****	*****	0.14 92.9	0.18 90.8	0.10 94.7	0.23 87.4	0.17 91.0
R-CRNN [7]	0.09 ***	0.04 ***	0.14 ***	0.09 95.5	*****	*****	*****	0.23 87.9
Multi-Scale RNN [9]	0.11 94.3	0.04 97.8	0.18 90.6	0.11 94.2	0.26 86.5	0.16 92.1	0.18 91.1	0.20 89.9
CRNN+Attention [14]	0.10 95.1	0.01 99.4	0.16 91.5	0.09 95.3	0.18 91.3	0.04 98.2	0.17 90.8	0.13 93.4

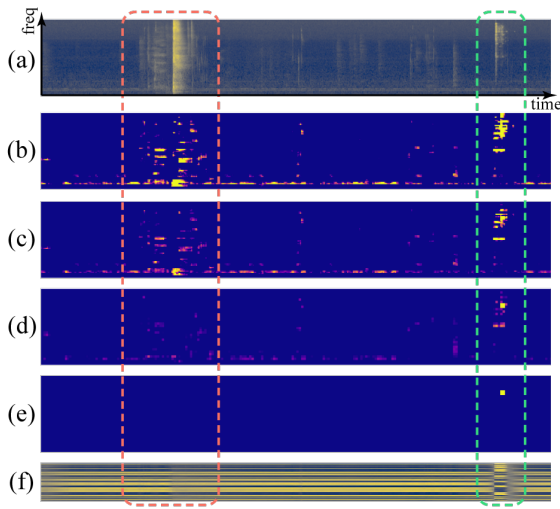


Figure 4: **Visualization of MTFA.** Taking glassbreak as an example, we visualize (a) the input spectrogram and (f) the output of RNN. From (b) to (e) are the outputs of MTFA from four scales with resolution from high to low. The target glassbreak is enclosed by the green box, while another sound which might interfere the prediction is enclosed by the pink box.

Note that for all methods, the performance on *gunshot* is the worst because the sound varies with different gun types and presents different temporal and frequential properties, which supports our analysis in the beginning.

On evaluation dataset, our method outperforms all previous methods on each target event. With the average error rate slightly increases from 0.07 on development dataset to 0.09 on evaluation dataset, it is confirmed that MTFA has a better generalization ability. To sum up, the above results demonstrate that with the help of multi-scale learning and attention mechanism, MTFA is more powerful an option than previous methods [5, 6, 7, 9, 14] to achieve discriminative features.

3.5. Visualization of MTFA

We visualize the output of the MTFA module in Figure 4 to provide a straightforward and intuitive understanding. A sound clip from evaluation dataset is taken as an example. As shown by the spectrogram (a), besides the target *glassbreak* near the end (en-

closed by the green box), the audio clip also comprises a man's 'cough' at the beginning (enclosed by the pink box). From (b) to (e), we present the outputs of the MTFA's *mask branch* extracted from four scales, where (b) and (e) has the highest and lowest resolution, respectively. It is obvious that low-resolution representations like (d) and (e) will drop details and only leave high-level information which only considers the location of the target event. Discarding details plays an important role here, since this move arms the model with the robustness against temporal and frequential variations, i.e., different scales possessed by events won't affect the high-level low-resolution results presented by (d) and (e). On the contrary, if analysis is performed using single resolution, like only with (b), the model might be more vulnerable to scale variations and produce worse performance, just like ResRNN and ResAttRNN in Table 1.

Meanwhile, despite that (b) assigns higher weights to background noise and the 'cough' sound beside *glassbreak*, (c), (d), and (e) gradually suppress these interferences more and more and only pay attention to the target event, making it quite easy for the following RNN to identify the target event, as clearly shown in (f). The visualization results provide another evidence that our MTFA is an effective approach for AED.

4. Conclusion

In this work, we propose a *Multi-Scale Time-Frequency Attention* module for acoustic event detection. Multi-scale learning is incorporated to capture the properties of events with different temporal and frequential scales, and time-frequency attention is applied to instruct the model to focus on not only important frames but also important frequency components. We evaluate the proposed method on DCASE2017 Task2. It turns out that MTFA is the best-performing single-model method on development dataset and achieves state-of-the-art on evaluation dataset, which proves its effectiveness on extracting discriminative multi-scale attention-aware features. We hope this work can provide insight for researchers on adopting multi-scale learning and temporal-frequential attention for acoustic-related tasks.

5. Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 61403224 and No. U1836219.

6. References

- [1] J. Schröder, S. Goetze, V. Grützmacher, and J. Anemüller, "Automatic acoustic siren detection in traffic noise by part-based models," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 493–497.
- [2] X. Wu, H. Gong, P. Chen, Z. Zhi, and Y. Xu, "Intelligent household surveillance robot," in *2008 IEEE International Conference on Robotics and Biomimetics*. IEEE, 2009, pp. 1734–1739.
- [3] P. Foggia, N. Petkov, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance of roads: A system for detecting anomalous sounds," *IEEE transactions on intelligent transportation systems*, vol. 17, no. 1, pp. 279–288, 2016.
- [4] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 85–92.
- [5] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," DCASE2017 Challenge, Tech. Rep., September 2017.
- [6] E. Cakir and T. Virtanen, "Convolutional recurrent neural networks for rare sound event detection," DCASE2017 Challenge, Tech. Rep., September 2017.
- [7] C.-C. Kao, W. Wang, M. Sun, and C. Wang, "R-cnn: Region-based convolutional recurrent neural network for audio event detection," in *Proc. Interspeech 2018*, 2018, pp. 1358–1362. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-2323>
- [8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.
- [9] W. Wang, C.-C. Kao, and C. Wang, "A simple model for detection of rare sound events," in *Proc. Interspeech 2018*, 2018, pp. 1344–1348. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-2338>
- [10] D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proc. Interspeech 2018*, 2018, pp. 152–156. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-1832>
- [11] C. Bothe, S. Magg, C. Weber, and S. Wermter, "Conversational analysis using utterance-level attention-based bidirectional recurrent neural networks," in *Proc. Interspeech 2018*, 2018, pp. 996–1000. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-2527>
- [12] X. Chang, Y. Qian, and D. Yu, "Monaural multi-talker speech recognition with attention mechanism and gated convolutional networks," in *Proc. Interspeech 2018*, 2018, pp. 1586–1590. [Online]. Available: <http://dx.doi.org/10.21437/Interspeech.2018-1547>
- [13] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.
- [14] Y.-H. Shen, K.-X. He, and W.-Q. Zhang, "Learning how to listen: A temporal-frequential attention model for sound event detection," *arXiv preprint arXiv:1810.11939*, 2018.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [16] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.
- [17] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1128–1132.
- [18] 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.
- [19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [20] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.
- [21] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016. [Online]. Available: <http://www.mdpi.com/2076-3417/6/6/162>