A Multi-feature Fusion Based Method For Urban Sound Tagging

Jisheng Bai, Chen Chen and Jianfeng Chen
School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
E-mail: {baijs,cc_chen524}@mail.nwpu.edu.cn, chenjf@nwpu.edu.cn

***Abstract*—Noise pollution is one of the serious issues for citizens. Mapping urban noise is essential to improve the quality of life for residents and construction for smart cities. Yet, most cities lack effective classification or tagging methods to monitor urban noise. To tackle this challenge, we propose a multi-feature fusion based method for urban sound tagging (UST). This method combines various features and Convolutional Neural Networks (CNNs) to predict whether noise of pollution is present in a 10-second recording. Log-Mel, harmonic, short-time Fourier transform (STFT) and Mel Frequency Cepstral Coefficents (MFCC) spectrograms are fed into different CNN architectures. And a fusion method is applied to make the final outputs. The proposed method is evaluated on the DCASE2019 task5 dataset and achieves a macro-AUPRC score of 0.68, outperforming the baseline system of 0.54.**

***Index Terms*—Noise pollution, Urban Sound Tagging, Multi-feature, Model fusion**

I. Introduction

It has been proved that noise pollution have effects on human life, economy and society. On the one hand, the exposure under noise can cause sleep disruption, heart disease and hearing loss, even learning and cognitive impairment in children [1]. On the other hand, although harmful levels of noise predominantly affect low-income and unemployed residents, these residents are the least likely to take the initiative of filing a complaint to the city officials. For reasons of comfort, public health and improving fairness, accountability, and transparency in public policies against noise pollution, to control and learn the distribution of noise is essential for government [2]. IEEE Audio and Acoustic Signal Processing (AASP) Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE) [3] is a series of challenges aimed on supporting the development of computational scene and event analysis methods. DCASE2019 task5 is a challenge evaluates systems for tagging short audio recordings with urban sound tags related to urban noise pollution. Meanwhile some of the most successful techniques in the challenge could inspire the development of an embedded solution for low-cost and scalable monitoring, analysis, and mitigation of urban noise.

Previous work on environmental sound classification relies on hand-craft features [4] and single feature type [5], and classifiers are usually based on Gaussian Mixtures Models (GMM) and Hidden Markov Models (HMM) [6] and deep CNN [7]. CNNs have been widely used in computer vision and have achieved state-of-the-art performance in several tasks such as image classification [8]. The filters can capture local patterns of the input feature maps, such as edges in lower layers and profiles of objects in higher layers. In sound detection and classification, the CNNs are successfully applied and achieve great results such as bird sound detection [9], acoustic scene classification [10] and domestic activities [11], [12]. CRNN has been proved to be state-of-art method of sound event detection [13], CapsNet [14] also achieve best results in sound event detection [15]. Inception-v3 [16] is applied in bird sound detection and improve detection performance [17]. Log-Mel spectrogram is a common feature and widely used in DCASE [18], [13].

In our system, firstly, log-Mel, harmonic, STFT, MFCC spectrograms are extracted as features. Then we experiment different features on VGG-like networks and analysis the influence between 8 coarse classes. After that, a gated activation [19] is further applied for sound event detection. Finally, we evaluate on evaluation data and fuse the results referred to the results of evaluation metrics between different classes.

This paper is organized as follows: the proposed method for UST will be introduced in Section II. Section III gives evaluation details such as datasets, evaluation metrics and settings. In Section IV, the evaluation results and discussion are shown and a conclusion is made in Section V.

II. METHOD

In this section, we illustrate the multi-feature fusion based method. The diagram of the proposed method is shown in Fig. 1.

A. Features

All recordings are resampled to target sample rate and converted to time-frequency spectrograms with STFT. This can be expressed as

$$\mathbf{STFT}(t,\omega) = \int_{-\infty}^{+\infty} s(\tau)g(\tau - t)e^{-j\omega\tau}d\tau \qquad (1)$$

where $g(\tau)$ is the window function, $s(\tau)$ is time-domain signal, t denotes the time index to obtain time localization by taking Fourier transform and ω is angular frequency.

Mel spectrum of each frame is computed by applying Mel filters, which is given by

$$MelSpec(m) = \sum_{k=f_{Mel}(m-1)}^{f_{Mel}(m+1)} H_m(k) * |X(k)|^2 \qquad (2)$$

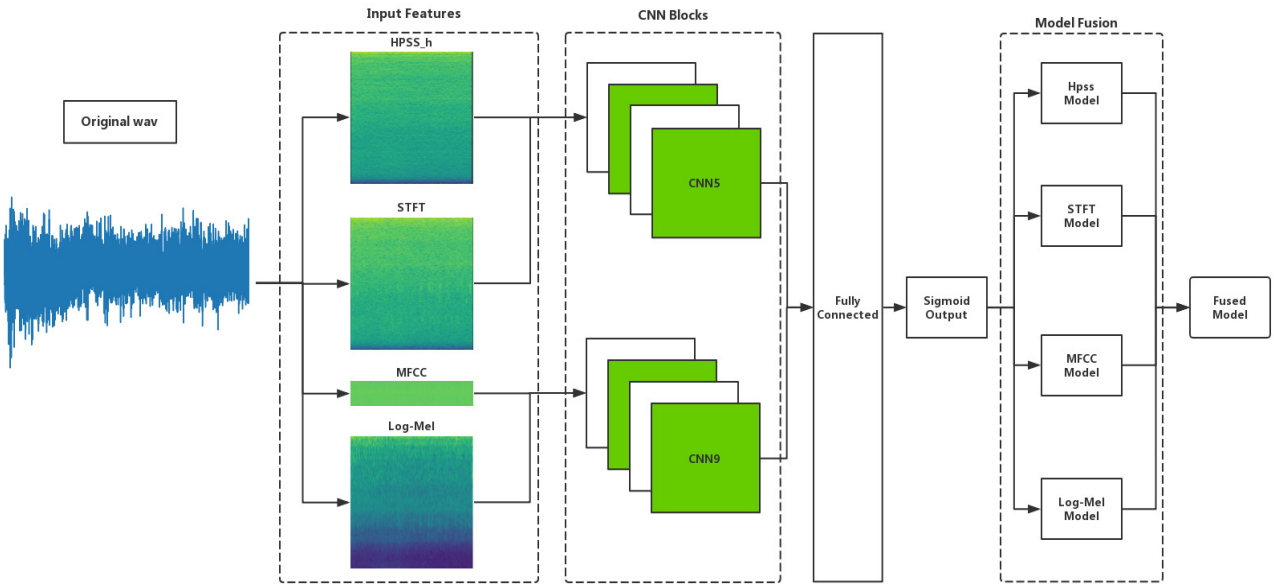

Fig. 1. The diagram of the proposed multi-feature fusion based method for Urban Sound Tagging

where $|X(k)|^2$ is the k^{th} power spectrum, $H_m(k)$ is the m^{th} Mel filter. The mapping from linear frequency to Mel frequency is shown in Eq. 3.

$$f_{Mel} = 2595 \cdot \lg\left(1 + \frac{f}{700}\right) \tag{3}$$

Finally, log-Mel spectrum is calculated by applying logarithm.

To generate MFCC, discrete cosine transform (DCT) is applied and the expression can be defined as follows

$$c_{MFCC}(i) = \sum_{m=1}^{M} (\lg MelSpec(m)) \cos\left[(m-0.5)\frac{i\pi}{M}\right] \tag{4}$$

where M is the number of Mel filter, $c_{MFCC}(i)$ is the i^{th} MFCC coefficient.

The harmonic percussive source separation (HPSS) [20] can split a signal $w(t)$ into harmonic part $h(t)$ and percussive part $p(t)$ and there are several approaches to separate. We can simplify the separation procedure as follows [21]

$$w(t) \overset{\text{HPSS}(l)}{\longrightarrow} h(t), p(t) \tag{5}$$

B. Network

To investigate different CNN architectures on the UST, VGG-like, CRNN, Inception-v3 and CapsNet are experimented based on log-Mel spectrogram with 64 Mel bands.

In VGG, the bigger convolutional kernels are replaced by smaller ones $(3 * 3)$ and they are stacked to increase the depth of CNN [8]. CRNN can take the advantages of CNN and RNN by stacking them. Inception-v3 factorizes bigger convolutions kernel into smaller ones and optimizes Inception module. CapsNet can catch space relationship by applying capsule and routing.

Several VGG-like networks are presented as main networks. The CNN5 and CNN9 architectures are similar to [22], it contains 4 convolitional layers or blocks and 1 dense layer, BN represents batch normalization , ReLu means leaky-ReLu activation. CRNN3 contains 3 convolutional layers and 2 recurrent layers, in CRNN9, 4 convolutional blocks are applied. Here we presented a new type of activation function named "gated", it is a deformation of the non-linear activation presented in [19] and can be expressed as

$$\mathbf{Z1} = \text{ReLu}(\mathbf{Y}) \tag{6}$$

$$\mathbf{Z2} = \text{Sigmoid}(\mathbf{Y}) \tag{7}$$

$$\mathbf{Z} = \mathbf{Z1} \otimes \mathbf{Z}2 \tag{8}$$

where $\mathbf{Y}$ is the output feature map, $Sigmoid$ is the activation function, then we multiply $\mathbf{Z1}$ and $\mathbf{Z2}$ to get the output of the gated activation $\mathbf{Z}$.

Both log-Mel and MFCC features are fed into CNN9, but STFT and harmonic spectrograms are fed into CNN5 because of the larger input parameters. The details of these network architectures are described in Table I.

C. Fusion

We use voting strategy as our fusion method. This strategy can be described as:

$$\mathbf{F}(n,c) = \sum \mathbf{F}_m(n,c) * \mathbf{I}(n,c) \tag{9}$$

TABLE I
FEATURE AND NETWORK ARCHITECTURE

	CNN5		CNN9	CNN9_gated
features	STFT	HPSS_h	log-Mel	MFCC
Conv1	3*3@64,BN,ReLu		(3*3@64,BN,ReLu)*2	(3*3@64,BN,Gated)*2
Pool1			2*2 average pooling	
Conv2	3*3@128,BN,ReLu		(3*3@128,BN,ReLu)*2	(3*3@128,BN,Gated)*2
Pool2			2*2 average pooling	
Conv3	3*3@256,BN,ReLu		(3*3@256,BN,ReLu)*2	(3*3@256,BN,Gated)*2
Pool3			2*2 average pooling	
Conv4	3*3@512,BN,ReLu		(3*3@512,BN,ReLu)*2	(3*3@512,BN,Gated)*2
Pool4			1*1 average pooling	
Dense			512	

where $\mathbf{F}_m(n,c)$ is the result matrix of the m^{th} model, $\mathbf{I}(n,c)$ is a matrix that its j^{th} column is set to 1 if the model achieve the best macro-AUPRC score among the models of one class, otherwise to 0, $*$ represents hadamard product.

III. EVALUATION

A. Development and evaluation datasets

We evaluate the proposed approach on the dataset of DCASE2019 task5 challenge. The development dataset contains a train split of 2351 recordings and a validate split of 443 recordings.

These recordings are from SONYC acoustic sensor network for urban noise pollution monitoring. The train and validate splits are disjoints and it make participants to develop computational systems for multilabel classification in a supervised manner. And validation subset can prevent overfitting during the training. The reference labels are coarse-level and fine-level taxonomies and each recording are listened at least three humans independently. The evaluation dataset contains 274 recordings and may be from validate split.

B. Evaluation Metrics

The UST challenge is a task of multilabel classification. The area under the precision-recall curve (AUPRC) is the classification metrics to evaluate. We vary τ between 0 and 1 and compute true positives (TP), false positives (FP), and false negatives (FN) for each coarse category. Then we can compute micro-averaged precision

$$P = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{10}$$

and recall

$$R = \mathrm{TP}/(\mathrm{TP}+\mathrm{TN}) \tag{11}$$

giving an equal importance to every sample. All values of τ in the interval $[0,1]$ will be computed to obtain different P and R. Finally, we can compute the area under the P-R curve. For each coarse category, the computations can be summarized as

$$\mathrm{TP} = \left(\sum_{k=1}^{K} t_k y_k\right) + t_0\left(1-\prod_{k=1}^{K} t_k y_k\right)\left(1-\prod_{k=0}^{K}(1-y_k)\right) \tag{12}$$

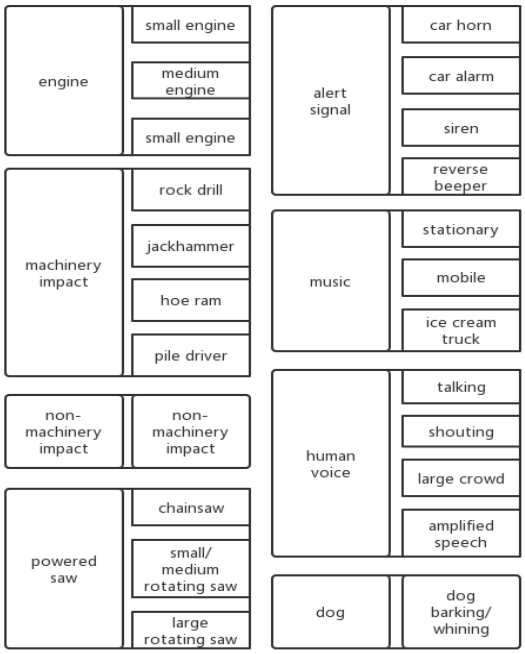

Fig. 2. Coarse-level taxonomy of urban sound tags in the DCASE Urban Sound Tagging task.

$$\mathrm{FP} = (1-t_0)\left(\left(\sum_{k=1}^{K}(1-t_k)\,y_k\right) + y_0\left(\prod_{k=1}^{K}(1-t_k)\right)\left(1-\prod_{k=1}^{K} y_k\right)\right) \tag{13}$$

$$\mathrm{FN} = \left(\sum_{k=1}^{K} t_k\,(1-y_k)\right) + t_0\left(\prod_{k=1}^{K}(1-t_k)\right)\left(\prod_{k=0}^{K}(1-y_k)\right) \tag{14}$$

where t_0 and y_0 respectively represent the presence of an incomplete tag in the ground truth and prediction, and t_k and y_k respectively represent the presence of fine tag k in the ground truth and prediction. For coarse-grained and fine-grained, micro-AUPRC and macro-AUPRC are both computed. We can compute $(P_1, R_1), (P_2, R_2), ..., (P_n, R_n)$ on every confusion matrix and then we get macro-P and macro-R by average them.

$$\text{macro} - P = \frac{1}{n}\sum_{i=1}^{n} P_i \tag{15}$$

$$\text{macro} - R = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{16}$$

If we average the corresponding elements, $\overline{TP}$, $\overline{FP}$, $\overline{FN}$ will be get firstly and then micro-P and micro-R are computed as following

$$\text{micro} - P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}} \tag{17}$$

$$\text{micro} - R = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \tag{18}$$

We mainly focus on coarse-level evaluation metrics for result analysis and model fusion. During experiments, f1score is also calculated as an extensional evaluation metrics, which is described as

$$F1 = \frac{2 \times P \times R}{P + R} \tag{19}$$

C. Settings

Main features are extracted with python librosa functions and audio recordings are preprocessed as follows.

Recordings are resampled to 32000 Hz and converted to time-frequency spectrogram with a Hanning window size of 1024 and hop length of 500 samples. Mel filters with different bands (64, 80 and 128) are applied and frequencies lower than 50 Hz and beyond 14000 Hz are removed. MFCC of 24 n-MFCC is calculated from log-Mel spectrogram.

Due to the limitation of GPU memory, recordings are resampled to 16000 Hz. Then STFT spectrograms with a Hanning window size of 1024 samples and a hop length of 664 samples, are extracted from recordings. The harmonic features can be converted from STFT. All spectrograms are converted to power spectrograms yielding a dynamic range of 80 dB.

Batch normalization [23] is applied to speed up and prevent overfitting during train steps. Leaky-ReLu or gated function are used as a non-linear activation after batch normalization respectively. Average pooling with size of 2*2 is applied to reduce the feature map. Then the frequency axis is averaged out and frame axis is maxed out after the last convolutional layer. For training, Tensorflow is implemented. Sigmoid cross entropy is utilized as loss function and AdamOptimizer as optimizer with a learning rate of 0.001. Training is done with batch size of 32 and we early stop the training if the macro-AUPRC does not improve in last 3 steps.

IV. EXPERIMENT RESULTS

A. CNN Architectures

We experiment CNN, CRNN, Inceptions-v3 and CapsNet, the results of different network architectures with Mel spectrograms of 64 bands are shown in Table II.

As we can see, the best CNN architecture for UST is CNN9, it achieve 0.06, 0.07 and 0.09 improvement on micro-AUPRC, micro-f1score and macro-AUPRC separately compared with baseline. CNN9 with gated function also achieves second best result. CRNN3 is far beyond CRNN9 and approximate to the result of Inception-v3. The worst result is CapsNet and it takes much more time to train.

B. Features Fusion

As it is described in Table III, the best macro-AUPRC of each coarse class is shown. Log-Mel performs well on 'machinery', 'non-machinery', 'alert' and 'human-voice'. As for 'music', obviously, harmonic components can be helpful for detecting music. MFCC is the most sensitive feature about 'dog'.For the 'engine' and 'powered-saw' classes, these additive noise seem to be differentiated by STFT. To fuse the results, the corresponding columns of $\mathbf{I}(n, c)$ in Eq. 9 are set to 1, for example, the column refers to 'music' of harmonic result matrix is set to 1, others to 0. The results of log-Mel, harmonic, STFT and MFCC models are fused to get the final result.

TABLE II
COARSE-LEVEL BEST PERFORMANCE

	Micro-AUPRC	Micro-f1score	Macro-AUPRC
Baseline	0.76	0.67	0.54
CNN9	0.82	0.74	0.63
CNN9_gated	0.81	0.72	0.62
CRNN3	0.72	0.66	0.51
Inception-v3	0.72	0.69	0.50
CRNN9	0.51	0.54	0.38
CapsNet	0.54	0.34	0.35

TABLE III
BEST MACRO-AUPRC SCORE OF EACH CLASS ON VALIDATE SPLIT

Coarse class	Feature	Macro-AUPRC
1_engine	STFT	0.85
2_machinery	log-Mel	0.54
3_non-machinery	log-Mel	0.62
4_powered-saw	STFT	0.80
5_alert	log-Mel	0.86
6_music	HPSS_h	0.47
7_human-voice	log-Mel	0.95
8_dog	MFCC	0.33

C. Discussion

By comparing the results, some conclusions can be made as follows:

In UST challenge, some architectures are supposed to obtain great performance, but they turn out to be worse, especially CRNN and CapsNet. Some reasons could be selecting the

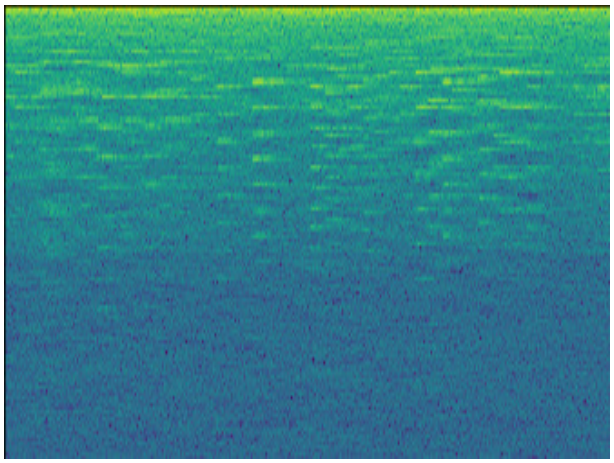

Fig. 3. Harmonic spectrogram of an example of music from dataset

appropriate hyper parameters of these architectures, or the hardness of training recurrent and capsule layers.

Different features should be generated for tagging different source of urban sound. STFT spectrogram can explore engine and power because it contains the original information. Harmonic spectrogram is discovered to recognize music better, because music contains harmonic waves apparently. Compared with other features, MFCC can future improve dog score from about 0.05 to 0.22. These may inspire us to classification different sound with specific features rather than one single type.

In our experiment, gated activation can further improve dog macro-AUPRC from 0.22 to 0.33 in comparison with leaky-ReLu activation in CNN9.

Final result is calculated by summarizing 4 different models, simple voting strategy can achieve a macro-AUPRC score of 0.68 and a 0.14 improvement compared with baseline system.

V. CONCLUSION

In this paper, we proposed a multi-feature fusion based method for the UST task. In our approach, four different features are generated as inputs of the networks. Then, VGG-like based CNN architectures are applied for urban sound tagging. Finally, we fused different results of the models according to the evaluation metrics of coarse classes. It is found that different features can be benefit for tagging different source of sound rather than one single type of feature. Our fusion method can achieve a macro-AUPRC score of 0.68, which is significantly better than DCASE task5 baseline system. For further work, different network architectures and hyper parameters selection will be studied, and advantages of tagging urban sound with different features will be researched as well.

ACKNOWLEDGMENT

The research work is supported by NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Information (U1609204)

REFERENCES

[1] S. A. Stansfeld and M. P. Matheson, "Noise pollution: non-auditory effects on health," *British medical bulletin*, vol. 68, no. 1, pp. 243–257, 2003.

[2] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, Feb 2019.

[3] 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 *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2013, pp. 1–4.

[4] T. Heittola, A. Mesaros, A. Eronen, and T. Virtanen, "Context-dependent sound event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 1, 2013.

[5] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 171–175.

[6] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini, "Automatic sound detection and recognition for noisy environment," in *2000 10th European Signal Processing Conference*. IEEE, 2000, pp. 1–4.

[7] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] M. Lasseck, "Acoustic bird detection with deep convolutional neural networks," DCASE2018 Challenge, Tech. Rep., September 2018.

[10] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Surrey-CVSSP system for DCASE2017 challenge task4," DCASE2017 Challenge, Tech. Rep., September 2017.

[11] D. Li and M. Wang, "Ciaic-moda system for dcase2018 challenge task5," DCASE2018 Challenge, Tech. Rep., September 2018.

[12] L. Vuegen, P. Karsmakers, B. Vanrumste *et al.*, "Weakly-supervised classification of domestic acoustic events for indoor monitoring applications," in *In proceedings of IEEE Conference on Biomedical and Health Informatics 2018*. IEEE, 2018.

[13] F. Vesperini, L. Gabrielli, E. Principi, and S. Squartini, "Polyphonic sound event detection by using capsule neural networks," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[14] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.

[15] Y. Liu, J. Tang, Y. Song, and L. Dai, "A capsule based approach for polyphonic sound event detection," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2018, pp. 1853–1857.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[17] M. Lasseck, "Audio-based bird species identification with deep convolutional neural networks," *Working Notes of CLEF*, vol. 2018, 2018.

[18] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. P. Shah, "Large-scale weakly labeled semi-supervised sound event detection in domestic environments," *arXiv preprint arXiv:1807.10501*, 2018.

[19] J. Bai, R. Wu, M. Wang, D. Li, D. Li, X. Han, Q. Wang, Q. Liu, B. Wang, and Z. Fu, "CIAIC-BAD system for DCASE2018 challenge task 3," DCASE2018 Challenge, Tech. Rep., September 2018.

[20] D. Fitzgerald, "Harmonic/percussive separation using median filtering," 2010.

[21] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 228–237, 2013.

[22] Q. Kong, Y. Cao, T. Iqbal, Y. Xu, W. Wang, and M. D. Plumbley, "Cross-task learning for audio tagging, sound event detection and spatial localization: Dcase 2019 baseline systems," *arXiv preprint arXiv:1904.03476*, 2019.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.