An Effective Perturbation based Semi-Supervised Learning Method for Sound Event Detection

Xu Zheng[1], *Yan Song*[1], *Jie Yan*[1], *Li-Rong Dai*[1], *Ian McLoughlin*[1,2], *Lin Liu*[3]

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.
[2]ICT Cluster, Singapore Institute of Technology, Singapore
[3]iFLYTEK Research, iFLYTEK CO., LTD, Hefei, China

{zx980216, yanjie17}@mail.ustc.edu.cn, {songy, ivm, lrdai}@ustc.edu.cn, linliu@iflytek.com

Abstract

Mean teacher based methods are increasingly achieving state-of-the-art performance for large-scale weakly labeled and unlabeled sound event detection (SED) tasks in recent DCASE challenges. By penalizing inconsistent predictions under different perturbations, mean teacher methods can exploit large-scale unlabeled data in a self-ensembling manner. In this paper, an effective perturbation based semi-supervised learning (SSL) method is proposed based on the mean teacher method. Specifically, a new independent component (IC) module is proposed to introduce perturbations for different convolutional layers, designed as a combination of batch normalization and dropblock operations. The proposed IC module can reduce correlation between neurons to improve performance. A global statistics pooling based attention module is further proposed to explicitly model inter-dependencies between the time-frequency domain and channels, using statistics information (*e.g.* mean, standard deviation, max) along different dimensions. This can provide an effective attention mechanism to adaptively re-calibrate the output feature map. Experimental results on Task 4 of the DCASE2018 challenge demonstrate the superiority of the proposed method, achieving about 39.8% F1-score, outperforming the previous winning system's 32.4% by a significant margin.
Index Terms: sound event detection, semi-supervised learning, independent component analysis, statistics pooling

1. Introduction

Sound event detection (SED) is the task of determining when and where target event categories occur in continuous audio. SED has attracted significant research attention due to its wide application in real-world systems, such as robotics [1], smart home devices [2], health care, and audio based indexing and retrieval [3, 4]. With the development of deep learning techniques, several mainstream deep neural networks (DNN), such as CNN, RNN and CRNN, have recently achieved state-of-the-art SED performance [4, 5, 6].

However, real-life SED is challenging, in part due to the lack of large-scale well annotated audio datasets, which are generally expensive and time-consuming to collect. Semi-supervised learning (SSL) SED methods that can exploit real data (which is either weakly labeled – without timestamp – or is unlabeled) to improve system performance, have thus drawn increasing research interest. Recent Detection and Classification of Acoustic Scenes and Events (DCASE) challenges have included a task for the evaluation of SSL based SED in domestic environments with weakly labeled audio data. There are several semi-supervised learning based methods in the literature,

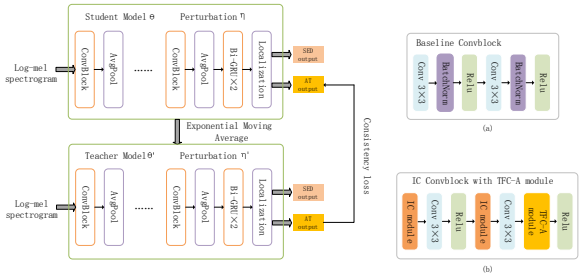

Figure 1: *(Left) framework of mean teacher based semi-supervised SED learning for large-scale weakly labeled data, and (a) the common Conv-Bn-Relu baseline convblock. (b) Our proposed Independent Component(IC) convblock with time-frequency-channel attention(TFC-A) module.*

including self-training [7], temporal ensembling (TE) [8], virtual adversary training (VAT) [9] and mean teacher (MT) [10]. In [7], the self-training based method was proposed to exploit the unlabeled data by generating pseudo-labels using models trained with small-size labeled data. In [8, 9, 10], perturbation based methods were proposed under the *smoothness* assumption, which indicates that two data points close to each other in feature space are likely to have the same label [11]. Among these methods, MT has shown promising SED performance in DCASE challenges, where the teacher acts as an ensemble of the students to generate the targets for SSL, and the consistency cost is employed as a regularization term. The key to effective MT is choosing suitable data and/or model perturbation to form a better teacher model from the student model and thus improve target quality. However, simply applying randomized data augmentation or dropout may not be optimal for introducing effective perturbation. In [12], a spec-augment technique was applied to improve data augmentation while in [13], different teacher and student models were exploited to perform SED and audio tagging(AT) respectively.

Inspired by independent component (IC) module in [14], in this paper, we propose an effective perturbation based semi-supervised learning (SSL) method based on mean teacher, as shown in Figure 1. This includes a new IC module, designed as a combination of batch normalization(BN) [15] and dropblock [16] operations. Its goal is two-fold: (1) Apply perturbations to the input of the internal convolutional layer to learn a better teacher and (2) Construct whitened input for convolutional filters in each intermediate convolutional layer. Com-

pared to dropout, the dropblock drops units in a contiguous region of a feature map, which can reduce the spatial correlation of the input. The IC module (BN+Dropblock) can approximate independent component analysis (ICA), which is traditionally implemented by two steps: the zero-phase component analysis (ZCA) to whiten the network activations, and rotation operations to get the final independent components [17]. That is, BN replaces the ZCA, while the dropblock reduces the dependencies within activations.

Furthermore, motivated by "squeeze-and-excitation" [18], which models inter-dependencies between the channels, a global statistics pooling based attention module is further proposed to explicitly model inter-dependencies between the time-frequency domain and channels using statistics information (*e.g.* mean, standard deviation, max) computed along different dimensions. This provides an effective attention mechanism which can adaptively re-calibrate the output feature map.

To evaluate the effectiveness of the proposed methods, extensive experiments have been conducted on DCASE2018 challenge task4 benchmarks. The experimental results show its superiority with 39.79% F1-score compared to 32.4% in the winning system.

2. Baseline method

In this section, we will firstly introduce the mean teacher based framework for SED [10], as shown in left of Figure 1. We adopt CRNN as the backbone architecture [19]. The CNN part is composed of 5 convolutional blocks, followed by two Bi-GRU layers to model long-term relationships and a localization module.

Since the task 4 of DCASE challenges focuses on weakly labeled data, which consists of AT and SED tasks. In mean teacher, two CRNNs (namely, teacher and student) with the same architecture are used. And the teacher model is updated by exponential moving average of the student model parameters. The consistency loss $L_{consist}$ is defined as the expected distance between the AT output of teacher model **T** (with weights θ' and perturbation $\eta\prime$) and student model **S** (with weights θ and perturbation η), which is

$$L_{consist} = MSE(\mathbf{S}_{\theta_{AT}}(\mathbf{x};\eta), \mathbf{T}_{\theta'_{AT}}(\mathbf{x};\eta\prime)) \qquad (1)$$

, where MSE is short for mean squared error.

In the baseline SED system, spec-augment [20] is applied to the input of CRNNs to perform input perturbation. This may be further improved by adding perturbation to the intermediate convolutional layer, such as dropout [21]. We will introduce our proposed method, in which the IC convolutional block(convblock) is used, instead of baseline convblock, to generate perturbation to intermediate convolutional layer. Furthermore, an attention mechanism based on global statistics information is proposed to improve the effectiveness of convblock output.

3. The perturbation based semi-supervised learning

As aforementioned, the proposed perturbation based SSL framework is obtained by replacing the traditional convblock (in Figure 1(a)) to Independent Component(IC) convblock with time-frequency-channel attention(TFC-A) module (in Figure 1(b)) The IC module is placed before the convolutional layer. It is worth noting that this is different from the operations in common practice, the BN layer is placed after the convolutional layer, followed by a non-linear activation. The TFC-A module, meanwhile, is proposed to explicitly model inter-dependencies between the time-frequency domain and channels. Details of the IC and attention modules will be described in the following subsections.

3.1. Independent Component(IC) module

Data or model perturbation plays an important role in the mean teacher SSL method. Generally, the perturbation is applied to the input spectrograms, via data augmentation techniques including Gaussian noise and spec-augment [20] . From this perspective, dropout [21] can be considered as a type of perturbation applied to the input of intermediate layer. However, it is shown that the dropout is less effective for convolutional layer than fully connected layer [16]. This may perhaps be caused by the fact that activation units in convolutional layer are spatially correlated, and the information of the dropped units can be partially recovered by the surrounding units. Furthermore, proposed by Li et al. [22], if dropout is placed before BN, it may lead to biased estimation of the mean and standard variation hyper-parameters in BN.

In this paper, an effective IC module, which consists of the BN and dropblock operator, is proposed. The BN is applied to normalize the input to distribution with zero mean and unit variance. And then in dropblock, which is a structured form of dropout, contiguous regions of a neurons are set to zero. By randomly dropping neurons in this way, the IC module can effectively introduce the perturbation to the input of convolutional layer. IC module in fact produces various student models, leading to a better teacher obtained by ensembling student models.

Besides, the IC module can also approximate the feature whitening operation, such as ZCA and ICA [17], which may facilitate the training procedure. Specially, ICA composes of two steps: 1) ZCA to de-correlate the input features, 2) Rotation to reduce the dependence. However, ICA is generally computational complex, especially for whitening the activations of a wide neural network.

For dropblock, there are two main hyper-parameters, $drop\_prob$ and $block\_size$. The $drop\_prob$ is defined same as dropout. The $block\_size$ defines the dropping area in activation map, and when $block\_size = 1$, dropblock resembles standard dropout.

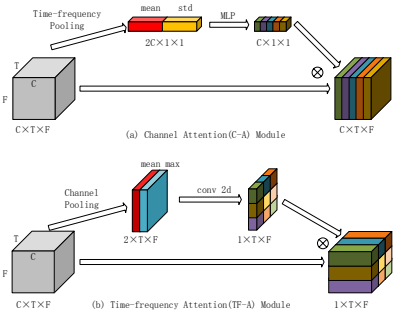

Figure 2: *Illustration of our proposed statistics pooling based attention modules, (a) channel attention (C-A) module, and (b) time-frequency attention (TF-A) module.*

3.2. Statistics pooling based attention modules

Given an intermediate feature map $\mathbf{U} \in \mathbb{R}^{C\times T\times F}$, the global statistics pooling based attention module consists of two parts: 1) Channel-attention module(C-A) with an 1D channel attention map $\mathbf{M}_C \in \mathbb{R}^C$, and 2) Time-frequency module(TF-A) with a 2D time-frequency attention map $\mathbf{M}_{TF} \in \mathbb{R}^{T\times F}$ where T, F and C are time, frequency and channel dimensions respectively.

Different statistics information from $\mathbf{U}$ (*e.g.* mean, standard deviation, max) is exploited for C-A and TF-A.

3.2.1. Channel Attention(C-A)

As shown in Figure 2(a), the statistics pooling is used to calculate the mean $\boldsymbol{\mu} \in \mathbb{R}^C$ and standard deviation $\boldsymbol{\sigma} \in \mathbb{R}^C$ over the time-frequency domain

$$\boldsymbol{\mu} = \frac{1}{T\times F}\sum_{i=1}^{T}\sum_{j=1}^{F}\mathbf{u}(i,j) \tag{2}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{1}{T\times F}\sum_{i=1}^{T}\sum_{j=1}^{F}\mathbf{u}^2(i,j) - \boldsymbol{\mu}^2} \tag{3}$$

The output of statistics pooling $\mathbf{z}$ is obtained by concatenating the mean and standard $\mathbf{z} = [\boldsymbol{\mu}; \boldsymbol{\sigma}]$. The channel attention map $\mathbf{M}_C$ is calculated via series of non-linear operations,

$$\mathbf{M}_C = \sigma(\mathbf{W}_2(\delta(\mathbf{W}_1\mathbf{z}))) \tag{4}$$

where $\mathbf{W}_1 \in \mathbb{R}^{2C\times\frac{C}{r}}$, $\mathbf{W}_2 \in \mathbb{R}^{\frac{C}{r}\times C}$ denote FC layers, $\sigma(\cdot)$ and $\delta(\cdot)$ denote sigmoid and ReLU respectively, r is the reduction rate.

3.2.2. Time-frequency Attention(TF-A)

As shown in Figure 2(b), in TF-A module, a 2D feature map $\mathbf{M}_{TF} \in \mathbb{R}^{T\times F}$ is used to provide the attention on time-frequency domain. It is obtained by first calculating the mean and max information over channels, followed by a convolutional layer and a sigmoid activation

$$\mathbf{M}_{TF} = \sigma(\mathbf{f}^{k\times k}[AvgPool(\mathbf{U}); MaxPool(\mathbf{U})]) \tag{5}$$

where $\mathbf{f}$ denotes a convolutional layer with kernel size $k \times k$.

3.2.3. Arrangement of Attention Modules

As aforementioned, the C-A and TF-A may provide complementary attentive information from their feature maps. Given C-A and TF-A modules, there are different arrangements of them, namely a sequential or parallel arrangement. For sequential arrangement of C-A and TF-A modules, we can have "C-A + TF-A" arrangement

$$\begin{aligned}\mathbf{U}' &= \mathbf{M}_C(\mathbf{U})\otimes\mathbf{U}\\ \mathbf{U}'' &= \mathbf{M}_{TF}(\mathbf{U}')\otimes\mathbf{U}'\end{aligned} \tag{6}$$

and reverse ordered sequential arrangement "TF-A + C-A"

$$\begin{aligned}\mathbf{U}' &= \mathbf{M}_{TF}(\mathbf{U})\otimes\mathbf{U}\\ \mathbf{U}'' &= \mathbf{M}_C(\mathbf{U}')\otimes\mathbf{U}'\end{aligned} \tag{7}$$

For parallel module is implemented as,

$$\mathbf{U}' = \mathbf{M}_{TF}(\mathbf{U})\otimes\mathbf{M}_C(\mathbf{U})\otimes\mathbf{U} \tag{8}$$

, where $\otimes$ denotes element-wise multiplication, and during multiplication, the attention map should be broadcasted. In experiments, we will evaluate different arrangement of C-A and TF-A modules.

4. Experiments Setup

4.1. Dataset

The experiments are conducted on the benchmark dataset from Task4 of the DCASE 2018 Challenge [23]. The dataset contains 1578 weakly-labeled training clips, 14412 unlabeled in-domain training clips, 39999 unlabeled out-of-domain training clips, 288 development clips and 880 evaluation clips. The average length of occurrence of each event class is presented in Table 1, indicating the very significant variance in duration between events. In our experiments, we utilize weakly-labeled clips with unlabeled in-domain clips as training set, and evaluate the performance on publicly available evaluation set.

Table 1: *Average length and median filter size of each class in the development dataset.*

Sound event label	**Average length(s)**	**Median filter size(s)**
Alarm_bell_ringing	1. 53	0.50
Blender	5. 35	1.75
Cat	0. 81	0.26
Dishes	0. 56	0.16
Dog	1. 03	0.34
Electric_shaver_toothbrush	7. 42	2.46
Frying	9. 34	3.10
Running_water	5. 61	1.85
Speech	1. 51	0.50
Vacuum_Cleaner	8. 66	2.86

4.2. Feature Extraction

The input features used in the proposed system are log-mel spectrograms, which are extracted from the audio signal resampled at 32 kHz. The spectrogram uses 64 Mel-scale filters and a window size of 32ms with 50% overlap between windows. As a result, each 10-second sound clip is transformed into a 2D time-frequency representation with a size of (640 × 64).

4.3. Experimental Settings

The neural networks are trained using the Adam optimizer [24], where the maximum learning rate is set to 0.001, and the total training epochs are set to 100. Specifically, there is a rampup for the learning rate over the first 20 epochs, and an adaptive median filter is used for backend processing. The filter size for each event category is selected according to Table 1.

For IC modules, we performed a set of experiments to determine that a sensible $block\_size$ is 5 for this configuration. In the TFC-A modules, similar empirical testing found a good reduction rate r is 8, and a reasonable kernel size for the convolutional layer in the TF-A module is 5×5.

Event based macro-F1 is used as the main metric For SED tasks. The experiment results are all evaluated by the sed_eval toolbox [25]. Onsets are evaluated with a collar tolerance of 200ms. Tolerance for offsets is computed per event as the maximum of 200ms or 20% of event length.

5. Results and Discussion

In experiments, we evaluate the performance of perturbation based SED systems including: 1) IC(dropblock): with IC module only to introduce perturbation on input of intermediate con-

Table 2: *Results of our proposed methods.*

System	Macro F1, %
Winner's system [26]	32.4
IC(dropblock)	39.30
TF-A + C-A	39.50
IC(dropblock)+TF-A + C-A	**39.79**

volutional layer, 2) TF-A + C-A: with the sequential arrangement of TF-A and C-A to introduce the attention mechanism for output, and 3) IC(dropblock + TF-A + C-A): the combination of both 1) and 2). As shown in Table 2, the proposed IC module, as well as TFC-A module can achieve F1-score over 39.00%. In addition, by combining the IC(dropblock) and "TF-A + C-A" modules, the F1-score can achieve 39.79%, which significantly outperforms the previously winning score of 32.4% [26]. The experiments results demonstrate the effectiveness of our proposed methods. A detailed ablation analysis will be conducted in the next subsections.

5.1. Evaluations of dropblock or dropout in IC module

We further evaluate the performance of dropout and dropblock in IC module with different $drop\_prob$. As shown in Table 3, the CRNN baseline in Figure 1(a) achieves an F1-score of 36.17%, with 3.77% gain over winner's system in DCASE2018 challenges. This may come from the superiority of the CRNN system as well as the adaptive median filter (in Table 1) for backend processing. The performance of system using IC module is further improved compared with the baseline system.

Furthermore, we can see that with the same $drop\_prob$, the performance of IC(dropblock) is generally better than IC(dropout) [14], indicating that our proposed IC(dropblock) provides more effective perturbations for convolutional layers. Specifically, among different $drop\_prob$, IC(dropblock) with $drop\_prob$=0.05 provides best F1-score of 39.30%, relative 3.13% improvement of our baseline system. To further analyze the effectiveness of TFC-A modules, we conduct several ablation experiments.

Table 3: *SED results from evaluating the IC modules.*

ConvBlock	drop_prob	Macro F1, %
Baseline convblock	-	36.17
IC(no perturbation)	0	37.74
IC(dropout) [14]	0.05	38.34
IC(dropout)	0.10	37.28
IC(dropout)	0.20	35.86
IC(dropblock)	0.05	**39.30**
IC(dropblock)	0.10	38.10
IC(dropblock)	0.20	36.49

5.2. Evaluations of different TFC-A

In the proposed attention method, different types of statistics information (*e.g.* mean, standard deviation, max) are used for TF-A and C-A. Results in Table 4, reveal quite wide differences in performance for the different types of statistics in the C-A module. Specially C-A(mean), same as Squeeze-and-Excitation [18], can improve the performance from 37.74% (baseline without C-A) to 38.68%. On the contrary, C-A(max) only achieves the performance of 25.25%, indicating that C-A(max) may loss the information of the overlapped events. In addition, C-A(mean std) incorporates the standard deviation as a second-order statistic information, and achieves F1-score of 39.00%. Finally, TF-A(mean-max) provides a slight improvement over the baseline system. This indicates that the class-wise information tends to be relatively strongly encoded in the channel dimension.

Table 4: *SED performance of different TFC-A modules.*

TFC-A module	Macro F, %
-	37.74
C-A (mean)	38.68
C-A (max)	25.25
C-A (mean std)	**39.00**
TF-A (mean max)	38.04
Parallel (TFC-A)	37.96
C-A + TF-A	37.54
TF-A + C-A	**39.50**

We also evaluate the performance of combining TF-A or C-A modules in parallel or sequential fashion, as shown in Table 4. We can see that sequential arrangement (TF-A + C-A) can achieve the F1-score of 39.50%, performing best with the same settings. Interestingly, the reverse ordered sequential arrangement (C-A + TF-A), performs worst. From the previous experiments we already know the importance of the C-A module. It seems that performing TF-A first strengthens the information on the time-frequency domain but it is necessary to apply C-A to emphasize the important channels. In the converse case, the C-A first sequential attention is overruled by the subsequent TF-A operation, thus reducing performance. The parallel TFC-A also provides slight improvement to the baseline system, but worse than the sequential attention module(TF-A + C-A).

6. Conclusion

In this paper, a novel perturbation based semi-supervised learning method, combining batch normalization with dropblock, is investigated. This not only provides perturbation to the convolutional layer but also whitens their inputs, to improve semi-supervised SED performance. Experimental results reveal the proposed dropblock based IC modules outperform conventional dropout, providing more effective perturbations to the convolutional layers. Furthermore, statistical pooling based attention module is used to explicitly model inter-dependency between time-frequency and channel domains. Among mean, standard deviation and max computed on different dimensions, we sequentially apply time-frequency attention, followed by channel attention, performing best. By combining the IC module and TFC-A module, the final F1-score of 39.8%, significantly outperforms the 32.4% achieved by the previously published winning system. In future, we hope to exploit other perturbation and attention types for semi-supervised SED.

7. Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant No. U1613211) and Key Science & Technology Project of Anhui Province(18030901016).

8. References

[1] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 405–409, 2016.

[2] A. Southern, F. Stevens, and D. Murphy, "Sounding out smart cities: Auralization and soundscape monitoring for environmental sound design," *The Journal of the Acoustical Society of America*, vol. 141, no. 5, pp. 3880–3880, 2017.

[3] S. Goetze, J. Schroder, S. Gerlach, D. Hollosi, J.-E. Appell, and F. Wallhoff, "Acoustic monitoring and localization for social care," *Journal of Computing Science and Engineering*, vol. 6, no. 1, pp. 40–50, 2012.

[4] Y. Wang, S. Rawat, and F. Metze, "Exploring audio semantic concepts for event-based video retrieval," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1360–1364.

[5] I. McLoughlin, H.-M. Zhang, Z.-P. Xie, Y. Song, and W. Xiao, "Robust sound event classification using deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 540–552, Mar. 2015.

[6] J. Yan, Y. Song, W. Guo, L.-R. Dai, I. McLoughlin, and L. Chen, "A region based attention method for weakly supervised sound event detection and classification," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 755–759.

[7] C. Rosenberg, M. Hebert, and H. Schneiderman, "Semi-supervised self-training of object detection models." *WACV/MOTION*, vol. 2, 2005.

[8] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[9] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Advances in neural information processing systems*, 2017, pp. 1195–1204.

[11] Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang, "Smooth neighbors on teacher graphs for semi-supervised learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8896–8905.

[12] J. Yan, Y. Song, L.-R. Dai, and I. McloughLin, "Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection," in *ICASSP 2020*, 2020, pp. 1195–1204.

[13] L. Lin, X. Wang, H. Liu, and Y. Qian, "What you need is a more professional teacher," *arXiv preprint arXiv:1906.02517*, 2019.

[14] G. Chen, P. Chen, Y. Shi, C.-Y. Hsieh, B. Liao, and S. Zhang, "Rethinking the usage of batch normalization and dropout in the training of deep neural networks," *arXiv preprint arXiv:1905.05928*, 2019.

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

[16] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "Dropblock: A regularization method for convolutional networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 727–10 737.

[17] H. Oja and K. Nordhausen, "Independent component analysis," *Encyclopedia of Environmetrics*, vol. 3, 2006.

[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

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

[20] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[21] 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.

[22] X. Li, S. Chen, X. Hu, and J. Yang, "Understanding the disharmony between dropout and batch normalization by variance shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2682–2690.

[23] 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.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, p. 162, 2016.

[26] L. JiaKai, "Mean teacher convolution system for DCASE 2018 task 4," *Tech. Rep., DCASE Challenge*, 2018.