

RCT: RANDOM CONSISTENCY TRAINING FOR SEMI-SUPERVISED SOUND EVENT DETECTION

Nian Shao¹, Erfan Loweimi², Xiaofei Li^{1*}

¹ Westlake University & Westlake Institute for Advanced Study, Hangzhou, China

² Centre for Speech Technology Research (CSTR), University of Edinburgh, Edinburgh, UK

shaonian@westlake.edu.cn, e.loweimi@ed.ac.uk, lixiaofei@westlake.edu.cn

Abstract

Sound event detection (SED), as a core module of acoustic environmental analysis, suffers from the problem of data deficiency. The integration of semi-supervised learning (SSL) largely mitigates such problem. This paper researches on several core modules of SSL, and introduces a random consistency training (RCT) strategy. First, a hard mixup data augmentation is proposed to account for the additive property of sounds. Second, a random augmentation scheme is applied to stochastically combine different types of data augmentation methods with high flexibility. Third, a self-consistency loss is proposed to be fused with the teacher-student model, aiming at stabilizing the training. Performance-wise, the proposed modules outperform their respective competitors, and as a whole the proposed SED strategies achieve 44.0% and 67.1% in terms of the PSDS₁ and PSDS₂ metrics proposed by the DCASE challenge, which notably outperforms other widely-used alternatives.

Index Terms: semi-supervised learning, sound event detection, data augmentation, consistency regularization

1. Introduction

Sound conveys a substantial amount of information about the environment. The skill of recognizing the surrounding environment is taken for granted by humans while it is a challenging task for machines [1]. Sound event detection (SED) aims to detect sound events within an audio stream by labeling the events as well as their corresponding occurrence timestamps. Taking advantage of deep neural networks, promising results have been obtained for SED [2]. However, the high annotation cost poses obstacles on its further development.

Two widely applied solutions for such data deficiency problem are data augmentation (DA) and semi-supervised learning (SSL). DA artificially enlarges the training dataset size in various forms including data warping, oversampling, etc. [3], while SSL leverages abundant unlabelled data to improve the model generalization capacity. For example, in computer vision (CV), different DA methods were proposed to transform the training images including rotation, cropping, etc. [3]. To combine multiple augmentation methods, *RandAugment* [4] presents a random strategy which arbitrarily selects one transformation in each training step. On the other hand, *mixup* [5, 6] conducts a linear interpolation of two classes of data points to oversample the dataset, by which the mixed samples could push the decision boundaries into low-density regions. Essentially, DA regularizes the model by constraining the predicted labels to be invariant to any noise applied on the inputs. Such idea is known as *consistency regularization* (CR) in SSL. When using CR for training, the model predictions are constrained to be invariant to

any noise not only on the inputs [7] but also on the hidden states [8, 9]. However, there exists a potential risk known as *confirmation bias* when the consistency loss is too heavily weighted in training [10]. To alleviate such risk, *MeanTeacher* [10] applies a consistency constraint in the model parameter space, which holds an exponential moving average (EMA) of the training student model to generate pseudo labels for unlabelled data. Other techniques such as *interpolation consistency training* (ICT) [7], *unsupervised data augmentation* (UDA) [11] further combine mixup or *RandAugmet* with *MeanTeacher* and CR, obtaining state-of-the-art SSL models, respectively.

The efforts on semi-supervised SED achieved promising results [12–16], thanks to the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge task 4¹, which establishes a systematically organized semi-supervised dataset [2] containing a reasonable amount of weakly-labelled (clip-level annotated) and unlabelled audio clips. The recent top-rank systems not only apply audio-specified DA methods including *SpecAugment* [17], time shift [16], pitch shift [18], etc., but also draw lessons from the SSL methods such as CR, *MeanTeacher*, etc. The top-rank system in 2018 DCASE challenge [12] introduces *MeanTeacher* into the semi-supervised SED, which becomes a role system for many variant systems [13–16]. Similar to ICT [7], shift consistency training (SCT) [16] proposes a way of unifying mixup, time shift and CR, obtaining a compatibly better semi-supervised SED model. Utilizing SCT and ICT, Zheng et al. [13] obtained the top rank in DCASE 2021 challenge task 4. Although multiple audio-specified DA methods [16–18] are leveraged in these models, the proper adaptation and effective combination of these SSL techniques are not scrutinised to be optimally applied in audio processing.

In this work, concerned with the proper usages of SSL techniques in audio processing, we propose a SSL strategy for SED, which consists of three novel modules:

i) *Hard mixup*. Vanilla mixup [5] is trivially applied in many audio processing studies [19–21], while its adaptability for audio processing is not fully investigated. In vanilla mixup, the mixed label is a soft convex combination of the labels from original samples. Instead, *Hard mixup* preserves the hard label of the original samples, as the mixup of sound events yields concurrent sound events due to the additive property of audio signal. E.g. The mixed label [0.2, 0, 0.8] in vanilla mixup would be presented as [1, 0, 1] in *hard mixup*. *Hard mixup* is the first algorithm which investigates the proper usage of the mixup method in the audio processing tasks.

ii) *RandomWarping*. Direct combination of various DA methods is not guaranteed to result in a performance gain, because of the complexity of finding an optimal set in a large hyperparameter searching space [3]. Although many efforts have

* Corresponding author.

¹<http://dcase.community/challenge2022>

been made toward unifying various data augmentation schemes in CV [4, 22], they could not be trivially adapted for audio processing, since methods such as time shift [16] or pitch shift [18] are specifically designed for audio. RandomWarping is among the first attempts toward unifying data augmentation methods for audio processing. We perturb each training sample with a randomly selected transformation, which allows taking advantage of different types of augmentation techniques in a unified way. It is a simple yet effective policy, while finding the optimal magnitude for each transformation can be challenging. We empirically find the optimal magnitude values for three data augmentation methods and achieve a consistent performance gain.

iii) *Self-consistency loss*. One of the challenges in SSL is designing the unsupervised loss for unlabelled data. In ICT [7] for mixup and SCT [16] for time shift, the MeanTeacher model and student model respectively process the original and augmented samples. In this paper, we propose an additional self-consistency loss to the MeanTeacher loss, which constrains the student model to give consistent predictions for original and augmented samples. Such self-consistency constraint always holds regardless of the correctness of the predictions, and thus would stabilize the training procedure.

The combination of these three modules is referred to as *random consistency training* (RCT), and the flowchart is shown in Fig. 1. With better adaptability for corresponding audio signals, each proposed module experimentally outperform its competing methods in the literature. As a whole, the proposed SSL strategy notably outperforms its counterparts, and achieves top performance on the DCASE 2021 challenge dataset.

2. The Proposed Method

SED is defined as a multi-class detection problem where the onset and offset timestamps of multiple sound events should be recognized from the input audio clips. We denote the time-frequency domain audio clips as $\mathbf{X}_i^{(l)} \in \mathbb{R}^{T \times K}$, where T is the number of frames (same for all clips in experiment) and K is the dimension of LogMel filterbank features. Three types of data annotations are used for training data, i.e. weakly labelled, strongly labelled and unlabelled, which are indicated by superscript $l \in \{w, s, u\}$, respectively. i denotes the index of data sample among a total of $N^{(l)}$ data points of one annotation type. Let C and $\mathbf{Y}_i^{(l)}$ be the number of sound event classes and data labels, respectively. The weakly labelled and strongly labelled data have the clip-level and frame-level labels denoted by $\mathbf{Y}_i^{(w)} \in \mathbb{R}^C$ and $\mathbf{Y}_i^{(s)} \in \mathbb{R}^{T' \times C}$, respectively. Since the required time resolution for SED is much lower than the one of sound frames, pooling layers are applied in the CNN, resulting in a coarser time resolution T' rather than T for the predictions.

The baseline model is a CRNN [12], consisting of a 7-layer CNN with Context Gating layers [23], cascaded by a 2-layer bidirectional GRU. An attention module is added at the end to produce different levels of predictions [12] and MeanTeacher [10] is employed for SSL. As shown in Fig. 1, a teacher model is obtained by an EMA of the student CRNN model to provide pseudo labels for unlabelled samples. The training loss is $\mathcal{L} = \mathcal{L}_{\text{Supervised}} + \mathcal{L}_{\text{MeanTeacher}}$, where $\mathcal{L}_{\text{Supervised}}$ is the cross-entropy loss for the labelled data, and $\mathcal{L}_{\text{MeanTeacher}}$ is the MeanTeacher mean square error (MSE) loss for the unlabelled data.

2.1. Random data augmentation for audio

RandAugment [4] is proposed as an efficient way of combining different types of image transformations. In RCT, we take such

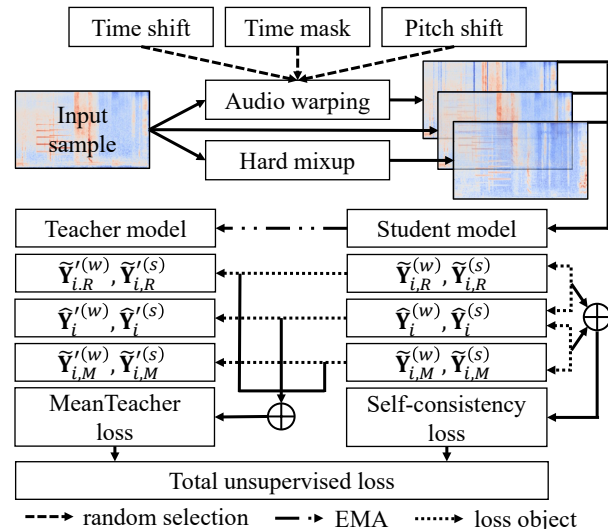


Figure 1: Flowchart of RCT: both hard mixup and audio warping are first applied for data augmentation; MeanTeacher [10] and self-consistency are used for SSL training. Subscripts R and M stand for RandomWarping and hard mixup, respectively.

idea to construct general audio warping methods for consistency regularization. Random data augmentation is accomplished by combining it with the proposed hard mixup, which means each training sample is augmented by hard mixup and random warping, respectively. As shown in Fig. 1, the batch size is thus tripled in each training step.

Hard mixup: The vanilla mixup [5] conducts an interpolation of two data points belonging to different classes, aiming to smooth the decision boundary. Such operation is feasible for images, whereas the interpolation of two audio clips produces a new clip due to the additive property of the sounds. As a result, combinations of multiple audio clips could be regarded as a realistic sample containing concurrent sound events, and should be recognized as an audio clip involving all sound events. Thus, the audio mixup is proposed to directly add multiple samples together, and the mixture is labelled with all the classes in all original samples. Since each sound clip reflect the realistic energies of all sound sources, the energy of the mixed sample is remained unchanged. Moreover, we noticed that combining more than two audio clips brings extra benefits. That is, it further condenses the distribution of sound events and can help the model toward better discriminating the sound events. Therefore, we randomly add two or three samples together in hard mixup.

RandomWarping: We use three audio warping methods in this work, and their warping magnitudes are set to $d \in \{1, 2, \dots, 9\}$. The only hyperparameter requires to be tuned is one unique d_{max} for all audio warping methods. For each mini batch, one warping method is randomly chosen with a random magnitude d uniformly distributed in $[1, d_{\text{max}}]$. This magnitude controls the audio warping intensities of the following methods:

- i) *Time shift* [16] circularly shifts each audio clip along time axis with a duration of $1 \times d$ seconds.
- ii) *Time mask* [17] randomly selects $5d$ intervals from the audio clip to be masked to 0. Since the shortest sound event lasts for 0.5 s, the length of each mask interval is set to 0.1 s.
- iii) *Pitch shift* [18] randomly raises or lowers the pitch of the audio clip by $1/2 \times d$ semitones, where both pitches and formants are stretched.

The magnitude unit of each method (1 s in time shift, 0.1 s in time mask, 1/2 semitone in pitch shift) is empirically selected.

2.2. Self-consistency training

The MeanTeacher loss [10] used in SED [12] has already shown a notable capacity in mining information from unlabelled data. To further leverage the unlabelled data, we propose to apply *self-consistency* regularization in addition to the MeanTeacher loss. Let $\hat{\mathbf{Y}}_i^{(w)} \in \mathbb{R}^C$ and $\tilde{\mathbf{Y}}_i^{(w)} \in \mathbb{R}^C$ denote the weak (clip-level) predictions of original and augmented samples of the student model, respectively. Similarly let $\hat{\mathbf{Y}}_i^{(s)} \in \mathbb{R}^{T' \times C}$ and $\tilde{\mathbf{Y}}_i^{(s)} \in \mathbb{R}^{T' \times C}$ for the strong (frame-level) predictions. Self-consistency regulates the model by an extra MSE term added to the loss function

$$\begin{aligned} \mathcal{L}_{SC} = & r(step) \frac{1}{N^{(w)}C} \sum_i^{N^{(w)}} \|\mathcal{D}_{aug}^{(w)}(\hat{\mathbf{Y}}_i^{(w)}) - \tilde{\mathbf{Y}}_i^{(w)}\|_2^2 \\ & + r(step) \frac{1}{N^{(s)}CT'} \sum_i^{N^{(s)}} \|\mathcal{D}_{aug}^{(s)}(\hat{\mathbf{Y}}_i^{(s)}) - \tilde{\mathbf{Y}}_i^{(s)}\|_2^2, \end{aligned} \quad (1)$$

where $\|\cdot\|_2$ denotes the Euclidean norm for a vector/matrix, $r(step)$ is a ramp-up function varying along the training step. $\mathcal{D}_{aug}^{(w)}$ and $\mathcal{D}_{aug}^{(s)}$ denote transformations applied on the predictions of original samples, as the labels should be correspondingly changed for augmented samples. For pitch shift and time mask, there is no need to change the labels. Time shift should accordingly shift the strong labels along time axis. As for hard mixup, the mixed audio clip includes all the sound classes presented in the original audio clips. However, the labels for the combined sound classes cannot be trivially obtained by adding the predictions of original samples, since the summation of two soft-predictions/labels is meaningless. Instead, we define the label transformation for hard mixup as

$$\mathcal{D}_{mixup}^{(l)}(\hat{\mathbf{Y}}_i^{(l)}) = \vee_{i \in \mathcal{M}} \text{harden}(\hat{\mathbf{Y}}_i^{(l)}), \quad (2)$$

where \vee denotes element-wise OR operation, \mathcal{M} is an arbitrary set consists of two or three data samples used in hard mixup, and $\text{harden}(\cdot)$ is an element-wise binary hardening function which ceils (or floors) the matrix elements to 1 (or 0) if the elements are larger than 0.95 (or smaller than 0.05). This transformation first hardens the predictions of original samples, from which the active/inactive sound classes are combined.

The prediction of original and augmented samples ($\hat{\mathbf{Y}}_i$ and $\tilde{\mathbf{Y}}_i$) are both used for gradient updating, which means they are considered as the pseudo labels for each other in training. This is also one reason of choosing symmetric MSE function as an extra component in the loss function. The total loss \mathcal{L} used for training the CRNN model will be

$$\mathcal{L} = \mathcal{L}_{Supervised} + \mathcal{L}_{MeanTeacher} + \mathcal{L}_{SC}, \quad (3)$$

where $\mathcal{L}_{MeanTeacher}$ is the average of MeanTeacher MSE losses for the original sample and two augmented samples (hard mixup and random warping), as shown in Fig. 1. The unsupervised loss is composed of the MeanTeacher loss and self-consistency regularization. Such self-consistency constraint between the original and augmented samples always holds regardless of the correctness of the predictions. This is different from ICT [7] that merges the MeanTeacher loss and the consistency loss as one single loss. In ICT, the MeanTeacher loss is set as the MSE loss between the prediction of augmented and original samples, where the predictions are obtained by the teacher model and student model, respectively. Such pseudo labels highly rely on the correctness of the MeanTeacher predictions, while incorrect pseudo labels may mislead the student model and consequently reduce the training efficiency.

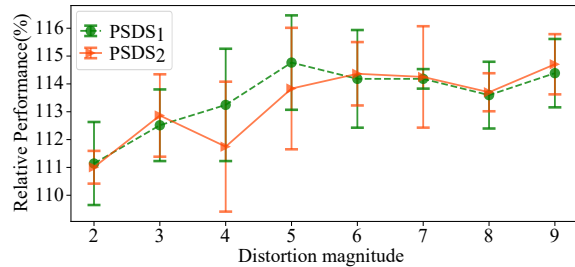


Figure 2: The relative performance gain as a function of maximum transformation magnitude (d_{max}). The transformation magnitude $d \sim U[1, d_{max}]$. Relative performance gain is computed using the baseline performance ($PSDS_1 = 34.74\%$, $PSDS_2 = 53.66\%$). The markers and vertical lines represent the mean and standard deviation computed using three trials.

Table 1: Ablation study for RCT. Different modules are added step by step and each score is obtained by averaging three trials.

Model	PSDS ₁ (%)	PSDS ₂ (%)
Baseline	34.7	53.7
+ Vanilla mixup [5]	34.9	57.9
+ Hard mixup	36.4	57.4
+ RandomWarping	38.1	58.5
+ ICT consistency [7]	38.0	59.2
+ Self-consistency	40.1	61.4

3. Experimental Results and Discussion

Our codes for this work have been released on our website². We use the baseline model [12] on DCASE 2021 Task 4 dataset³ to test the performance of the proposed method. The dataset consists of 1578 weakly-labelled, 10000 synthesized strongly-labelled and 14412 unlabelled audio clips. Each 10-second audio clip is resampled to 16 kHz and frame blocked with a length of 128 ms (2048 samples) and a hop length of 16 ms (256 samples). After 2048-point fast Fourier transform, 128-dimensional LogMel features are extracted for each frame, converting the 10-second audio clip into a 626×128 spectrogram. All samples are normalized to $[-1, 1]$ before training.

The batch size is set to 48, consisting of 12 weakly-labelled, 12 strongly-labelled and 24 unlabelled data points. The learning rate ramps up to 10^{-3} until epoch 50 and is scheduled by Adam optimizer [24] until the end of training, i.e. 200 epochs. The weight for MeanTeacher and self-consistency losses, $r(step)$, linearly ramps up from 0 to 2 at epoch 50 and is then kept unchanged. The system performance is evaluated through polyphonic sound detection scores (PSDSs) [25] according to the DCASE 2021 challenge guidelines. The metrics takes both response speed (PSDS₁) and cross-trigger performance (PSDS₂) into account; the larger the better for both metrics.

3.1. Ablation study on SED

In the random warping policy, the only hyperparameter that needs to be grid-searched is the maximum transformation magnitude d_{max} . As shown in Fig. 2, the performance improves as the maximum transformation magnitude increases until about 5

²<https://github.com/Audio-Westlake/URCT>

³<http://dcase.community/challenge2021/task-sound-event-detection-and-separation-in-domestic-environments>

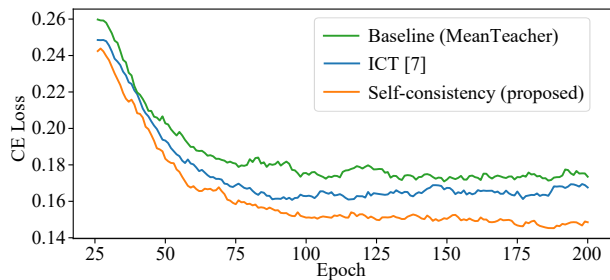


Figure 3: Cross-entropy loss of strongly-supervised validation data when training with or without self-consistency loss, comparing with the ICT scheme [7].

or 6; accordingly we set d_{\max} to 5.

Table 1 shows the result of ablation study, in which the proposed modules are individually added step by step. As seen, the proposed schemes including hard mixup, RandomWarping and self-consistency, lead to noticeable positive contributions. To further investigate the efficacy of the proposed methods, we also conduct experiments to replace hard mixup and self-consistency by vanilla mixup [5] and ICT-like consistency [7], respectively. While vanilla mixup is slightly better in cross-trigger (PSDS₂), hard mixup gives more significant gain in response time (PSDS₁). The proposed self-consistency outperforms the ICT-like consistency in both metrics.

3.2. Comparison with other semi-supervised strategies

To evaluate the proposed SSL strategy collectively, we conduct a thorough comparison with other widely used SSL strategies, including ICT [7], SCT [16] and their combination, employing the same baseline network used for the proposed strategy. Table 2 shows the comparison results, and Fig. 3 gives the validation loss curves of the models with baseline MeanTeacher, ICT and self-consistency. Compared with the baseline MeanTeacher model, ICT largely improves the performance by its proposed teacher-student consistency loss, while the always-hold self-consistency loss further improves the performance, which can be reflected by a lower validation loss curve shown in Fig. 3. SCT performance is not as good as ICT, which indicates that the time shift is not as efficient as interpolation (mixup). Combining ICT and SCT does not outperform ICT alone, which implies that the naive addition of ICT and SCT [16] is not an effective way to combine multiple different augmentations schemes. In contrast, as shown in Table 1, the proposed strategy is able to efficiently combine multiple different augmentations. Overall, the proposed method remarkably outperforms ICT and SCT, due to the strength of each proposed module and the efficient stochastic combination of them.

3.3. Comparison with DCASE2021 submissions

To further assess the efficacy of the proposed method and conduct fair comparisons with heavily processed DCASE 2021 submitted models, we also took advantage of some existing post-processing and ensembling techniques in our model. A temperature factor of 2.1 is used for inference temperature tuning as in [13]. Median filter is applied to smooth the frame-level predictions. Following [26], the length of each median filter is calculated by $\text{length}_{\text{class}} = 0.55 \times \text{avg_duration}_{\text{class}} / \text{duration}_{\text{frame}}$ and manually search for the best value with a range of ± 2 . The length of class-wise median filters are set to $\{3, 28, 7, 4, 7, 22, 48, 19, 10, 50\}$, for the event classes of $\{\text{alarm bell ringing, blender, cat, dishes, dog, elec-}$

Table 2: Comparing the proposed SSL strategy with other alternatives. Each score is obtained by averaging three trials.

Model	PSDS ₁ (%)	PSDS ₂ (%)
Baseline [12]	34.7	53.7
SCT [16]	36.0	55.6
ICT [7]	37.7	57.7
ICT+SCT [16]	37.0	58.7
RCT (proposed)	40.1	61.4

Table 3: Comparing the proposed system with DCASE2021 top-ranked submissions. All models are named in the form of network architecture plus the SSL strategy.

Model	PSDS ₁ (%)	PSDS ₂ (%)
CRNN (baseline) [12]	34.7	53.7
FBCRNN+MLFL [20]	40.1	59.7
CRNN+IPL [15]	40.7	65.3
CRNN+DA [21]	41.9	63.8
CRNN+HeavyAug. [14]	43.4	63.9
RCRNN+NS [19]	45.1	67.9
SKUnit+ICT/SCT [5]	45.4	67.1
CRNN+RCT (proposed)	44.0	67.1

tric shaver toothbrush, frying, running water, speech, vacuum cleaner}, respectively. Moreover, model ensembling was applied to fuse the predictions of multiple differently trained models. We trained eleven models with different variants of RCT: substituting time masking with frequency masking [17]; adding FilterAug [14] into audio warping choices; randomly selecting one or two methods in audio warping; and, reducing the weight of the MeanTeacher loss. We found that all different variants achieve reasonable performances. This demonstrates the flexibility of RCT and its capacity in incorporating new audio transformations into the framework with a low tuning cost overhead.

The proposed system is compared with DCASE2021 top-ranked submissions in Table 3. The scores of DCASE2021 submissions are directly quoted from the challenge results. The proposed system noticeably outperforms all other systems employing the baseline CRNN network, which verifies the superiority of the proposed RCT strategy. Furthermore, as seen the performance of the proposed system is very close to the two first-ranked submissions [13, 19]. They both use more powerful networks, i.e. SKUnit and RCRNN, which were able to largely improve the performance [13, 19]. The proposed framework is independent of the network and can be easily applied along with more advanced architectures to achieve higher performance.

4. Conclusion

In this paper, we developed a novel semi-supervised learning (SSL) strategy, named random consistency training (RCT), for sound event detection (SED) task. The proposed method improves several core modules of SSL, including unsupervised training loss and data augmentation schemes. It leads to achieving competitive performance on the DCASE 2021 challenge dataset. RCT provides a generic framework which can be effectively employed along with more advanced augmentation schemes and architectures. Besides, since RCT is not task-specific, it can potentially be applied in various audio and image processing tasks which is another broad avenue for future work.

5. References

- [1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational Analysis of Sound Scenes and Events*. Springer, 2018.
- [2] N. Turpault, R. Serizel, A. Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, 2019, p. 253.
- [3] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.
- [4] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical automated data augmentation with a reduced search space," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE Computer Society, 2020, pp. 3008–3017.
- [5] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018.
- [6] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," in *International Conference on Learning Representations (ICLR)*, 2018.
- [7] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 3635–3641.
- [8] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," *Advances in Neural Information Processing Systems (NIPS)*, vol. 27, pp. 3365–3373, 2014.
- [9] L. Samuli and A. Timo, "Temporal ensembling for semi-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2017.
- [10] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.
- [11] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," *Advances in Neural Information Processing Systems (NIPS)*, vol. 33, 2020.
- [12] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," *Detection and Classification of Acoustic Scenes and Events*, 2018.
- [13] X. Zheng, H. Chen, and Y. Song, "Zheng ustc team's submission for dcase2021 task4 – semi-supervised sound event detection," DCASE2021 Challenge, Tech. Rep., June 2021.
- [14] H. Nam, B.-Y. Ko, G.-T. Lee, S.-H. Kim, W.-H. Jung, S.-M. Choi, and Y.-H. Park, "Heavily augmented sound event detection utilizing weak predictions," DCASE2021 Challenge, Tech. Rep., June 2021.
- [15] Y. Gong, C. Li, X. Wang, L. Ma, S. Yang, and Z. W. Wu, "Improved pseudo-labeling method for semi-supervised sound event detection," DCASE2021 Challenge, Tech. Rep., June 2021.
- [16] C.-Y. Koh, Y.-S. Chen, Y.-W. Liu, and M. R. Bai, "Sound event detection by consistency training and pseudo-labeling with feature-pyramid convolutional recurrent neural networks," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 376–380.
- [17] D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.
- [18] B. McFee, E. Humphrey, and J. Bello, "A software framework for musical data augmentation," in *16th International Society for Music Information Retrieval Conference (ISMIR)*, 2015, pp. 248–254.
- [19] N. K. Kim and H. K. Kim, "Self-training with noisy student model and semi-supervised loss function for dcase 2021 challenge task 4," DCASE2021 Challenge, Tech. Rep., June 2021.
- [20] G. Tian, Y. Huang, Z. Ye, S. Ma, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, K. Ouchi, J. Ebberts, and R. Haeb-Umbach, "Sound event detection using metric learning and focal loss for dcase 2021 task 4," DCASE2021 Challenge, Tech. Rep., June 2021.
- [21] R. Lu, W. Hu, D. Zhiyao, and J. Liu, "Integrating advantages of recurrent and transformer structures for sound event detection in multiple scenarios," DCASE2021 Challenge, Tech. Rep., June 2021.
- [22] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *CVPR*, 2019, pp. 113–123.
- [23] M. Antoine, L. Ivan, and S. Josef, "Learnable pooling with context gating for video classification," *CoRR*, vol. abs/1706.06905, 2017. [Online]. Available: <http://arxiv.org/abs/1706.06905>
- [24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
- [25] A. Mesaros, T. Heittola, and T. Virtanen, "Metrics for polyphonic sound event detection," *Applied Sciences*, vol. 6, no. 6, 2016.
- [26] Y. Liu, C. Chen, J. Kuang, and P. Zhang, "Semi-supervised sound event detection based on mean teacher with power pooling and data augmentation," DCASE2020 Challenge, Tech. Rep., June 2020.