INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Resolution Consistency Training on Time-Frequency Domain for Semi-Supervised Sound Event Detection

Won-Gook Choi, and Joon-Hyuk Chang

Department of Electronic Engineering Hanyang University, Seoul, Republic of Korea
onlyworld94@hanyang.ac.kr, jchang@hanyang.ac.kr

Abstract

The fact that unlabeled data can be used for supervised learning is of considerable relevance concerning polyphonic sound event detection (PSED) because of the high costs of frame-wise labeling. While semi-supervised learning (SSL) for image tasks has been extensively developed, SSL for PSED has not been substantially explored due to data augmentation limitations. In this paper, we propose a novel SSL strategy for PSED called resolution consistency training (ResCT), combining unsupervised terms with the mean teacher using different resolutions of a spectrogram for data augmentation. The proposed method regularizes the consistency between the model predictions for different resolutions by controlling the sampling rate and window size. Experimental results show that ResCT outperforms other SSL methods on various evaluation metrics: event-f1 score, intersection-f1 score, and PSDSs. Finally, we report on some ablation studies for the weak and strong augmentation policies.

Index Terms: sound event detection, semi-supervised learning, data augmentation, multi-resolutional training

1. Introduction

Polyphonic sound event detection (PSED) detects whether predefined target sounds appear in an audio signal. PSED is divided into two types according to the detection level: weak and strong target detection (also called clip- and frame-level, respectively). Strong target detection has the advantage that the system could make the target sounds appear or not frame-by-frame. However, the frame-wise labeling of this detection type incurs a high cost. For this reason, most PSED systems are trained under semi-supervised learning (SSL) methods.

SSL is a learning strategy that effectively uses unlabeled data for training when labeled data are limited. Consistency regularization training (CRT) is an SSL approach that aims to optimize a system to predict similar results under the manifold assumption using a few perturbations on data or a model. For recent PSED, mean teacher (MT) [1] that uses *a teacher-student structure* for unlabeled data points is a widely used CRT method [2–7]. Furthermore, MT is variously expanded by combining data augmentation methods, such as the random consistency training (RCT) [2], interpolation consistency training (ICT) [6], and shift consistency training (SCT) [7].

In the vision and language processing task, other CRT methods have been developed: unsupervised data augmentation (UDA) [8] and FixMatch [9], which regularize consistency using data augmentations and artificial labels on *a single network*. Both strategies result in similar class distributions between *weakly (or non-) and strongly augmented data*. Empirical studies showed that RandAugment [10], CTAugment [11], and

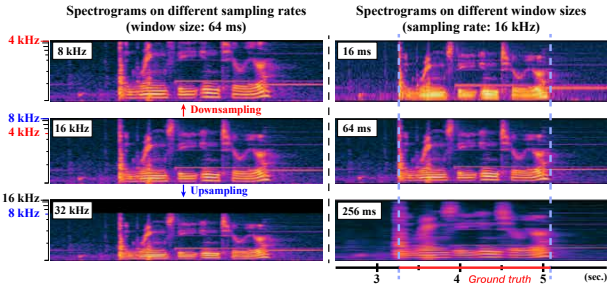

Figure 1: *Differences of log-mel spectrograms according to only controlling sampling rates (**Left**) and window sizes (**Right**).*

AutoAugment [12] perform well as strong augmentation policies in image classification, and back translation [13] has been effective for text classification. Recently, the RandAugment-like method, RCT, has been conquered in audio task, but has not been explored with the CRT between weakly and strongly augmented data. *Grollmisch et al.* [14] compared augmentation methods for FixMatch-based clip-level audio classification. However, these augmentation methods were severely distorted in time domain target (e.g., rotation, SpecAugment [15], etc.); thus, it is hard to adopt in PSED.

This study proposes a novel SSL strategy for PSED, resolution consistency training (ResCT), which smooths decision boundaries using *different time-frequency resolutions* by *combining MT and FixMatch structures*. Multi-resolution approaches have been explored in many audio tasks, such as PSED, speech enhancement, and neural vocoder for designing model fusion [16], progressive generative model [17], and joint training scheme [18], respectively, using different sampling rates, window sizes, or frame shift sizes. Different from the previous studies, this work explores multi-resolution augmentations to regularize consistencies to temporal and spectral perturbations among the different time-frequency resolutions.

A digital signal has limited frequency bands according to sampling rates, and a spectrogram has different time-frequency resolutions depending on window sizes. Based on these properties, ResCT adopts a weak augmentation as different sampling rates and a strong augmentation as different sampling rates and window sizes with stochastic values. Also, we reform the training procedure by combining MT and FixMatch methods, which consist of supervised, unsupervised, and teacher-student terms for effective ResCT using the proposed augmentation policies.

The experimental results show that the proposed method surpassed the other SSL strategies *especially on detection the onset and offset* of targets. Also, we compared the performances of the different strong augmentation strategies: the proposed method and the extant augmentations. At the end of this pa-

10.21437/Interspeech.2023-350

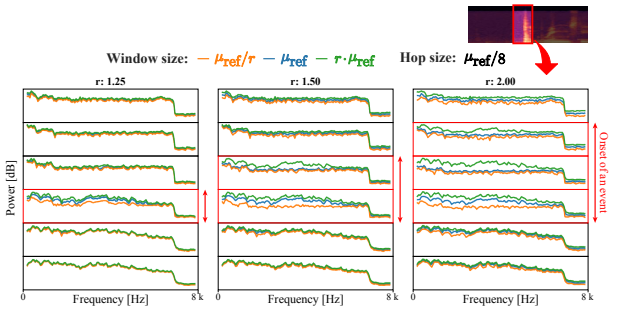

Figure 2: *Spectral analysis over six frames near the onset of an event. The spectrums over the red boxes are of event inactive frames, and below the red boxes are of event active frames. Regardless of the window size, trends of the spectrums are similar except for the onset interval. In this analysis, μ_{ref} was 128 ms.*

per, we discuss the most effective multi-resolution augmentation policy through some ablation studies.

2. Consistency regularization training

2.1. Consistency training with a single network

UDA and FixMatch are well-known CRT strategies using a single network. Both use unlabeled data to enforce consistency between raw or weakly and strongly augmented data points (unsupervised term, L_{usp}) while producing the decision boundary using labeled data (supervised term, L_{sup}), making unlabeled data useful.

$$L_{total} = L_{sup} + L_{usp}. \tag{1}$$

For example, in an image, simple cropping, flipping, or shifting augmentation is used for weak augmentation, and RandAugment, CTAugment, or AutoAugment is used for strong augmentation. Note that although the strong augmentation makes data distortion difficult, the data labels should not be distorted severely.

2.2. Consistency training with teacher-student networks

MT is a teacher-student training approach in which only the student network is trained, and the teacher network is updated with an exponential moving average (EMA) of the student's parameters at every training step (i.e., the teacher is not updated with backpropagation). The following two loss terms are used to train the network:

$$L_{total} = L_{sup} + r(t)L_{TS}, \tag{2}$$

where L_{sup}, L_{TS}, and $r(t)$ denotes the supervised loss between the outputs and targets of the student network, consistency loss between the teacher and student network outputs, and the exponential ramp-up function, respectively. The consistency loss term makes the decision boundary smoother by regularizing the perturbation of the model and data points.

By combining the data augmentation methods and MT, improved consistency learning methods have been proposed: ICT, SCT and RCT. ICT maintains the consistency between the student's output of a mixed data point and the mixed teacher's output for image classification. Similarly, SCT maintains the consistency between the student's output of a frame-shifted data point and the frame-shifted teacher's output for PSED. RCT maintains the consistencies using a hard mixup and an audio warping. The audio warping is randomly chosen among time shift, time mask, and pitch shift.

3. Resolution consistency training

In this section, we introduce ResCT, which regularizes the model to create a consistent prediction of the same audio signal, even if the resolutions of the hand-crafted features are different. The main proposals follows: first, we propose a policy of multi-resolution augmentation for CRT. For consistency training, we leverage two variables when extracting a spectrogram: the sampling rate and window size, which could make spectral and temporal perturbations. The maximum frequency of a digital signal is limited to half the sampling rate; hence, if audio is down-sampled, the high-frequency band narrows. If we consider a spectrogram as an image, this manipulation is similar to image padding, cropping, and stretching augmentation (Fig. 1. Left). When extracting the spectrogram from a waveform, the spectral resolution is higher if a window is longer, but the time resolution worsens; in other words, temporal perturbations arise (Fig. 1. Right, Fig 2). If the scenario is vice versa, spectral perturbations arise. These temporal and spectral perturbations could smoothen a decision boundary for both time (detecting either onset, ongoing, or offset) and target events.

Second, we optimize the network by adopting both MT and FixMatch structures. The proposed ResCT comprises three terms to utilize the unlabeled data: supervised, unsupervised, and teacher-student terms (Fig. 3). The following subsections describe the details of the policy of leveraging a sampling rate and window size and the three terms mentioned above.

3.1. Multi-resolution policy for consistency training

Sampling rates and window sizes for the augmentation are sampled from a log-Gaussian distribution in which the mean and standard deviation are logarithmic values of the reference values (i.e., non-augmented data) and distortion ratios on a Gaussian distribution, respectively.

$$\ln x \sim \mathcal{N}(\ln \mu_{\text{ref}}, (\ln r)^2), \tag{3}$$

where x, μ_{ref}, and r denote the random variable, reference value, and ratio, respectively. For example, if μ_{ref} and r of the window size are 128 ms and 2, the probability that the random variable x is sampled from 64 ms to 256 ms is approximately 68.27%.

The spectral analysis depicted in Fig. 2 shows how much the temporal perturbations arise near the onset of an event according to the ratio. If the window size is eight times the hop size, only a frame had perturbations when the ratio was 1.25. Therefore, in this study, the reference value of window size and ratio were set at 128 ms and 1.25, respectively. In addition, the reference value of the sampling rate and ratio were empirically set at 16 kHz and 1.25, and we clipped all sampled values on half and twice ratios to prevent severe distortions.

3.2. Resolution consistency training

On the supervised term of ResCT, the network is trained with different-resolution spectrograms having the same label. The unsupervised term regularizes the consistency between weakly and strongly augmented data, which have different multi-resolution policies. The teacher-student term regularizes the detection probabilities between the teacher and student networks on the same strongly augmented data. Let $(l, y) \in \mathcal{L}$ and $u \in \mathcal{U}$

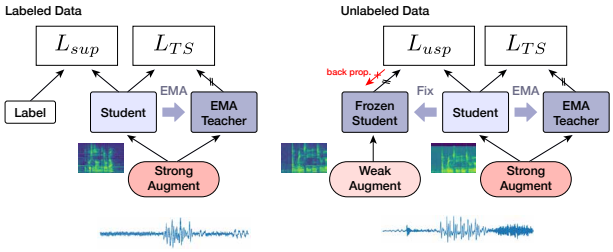

Figure 3: *Overall structure of resolution consistency training.*

be the labeled dataset and unlabeled dataset respectively, where l, y, and u denote a labeled sample, label, and unlabeled sample, respectively. The loss function of ResCT comprises three loss terms: a supervised loss L_{sup}, unsupervised loss L_{usp}, and teacher-student loss L_{TS}. First, L_{sup} is just the binary cross-entropy (BCE) loss on the strong-augmented labeled data:

$$L_{sup} = \mathcal{H}(f_w(l_{\mathcal{A}}), y_w) + \mathcal{H}(f_s(l_{\mathcal{A}}), y_s), \tag{4}$$

where $f(\cdot)$, $\mathcal{H}(\cdot)$, and the subscripts $\mathcal{A}$, w and s represent the student network, BCE, and the strongly augmented data, clip- and frame-level, respectively. In contrast, L_{usp} is the BCE loss on strong-augmented unlabeled data:

$$L_{usp} = \mathcal{H}(f_w(u_{\mathcal{A}}), \bar{f}_w(u_{\boldsymbol{\alpha}})) + \mathcal{H}(f_s(u_{\mathcal{A}}), \bar{f}_s(u_{\boldsymbol{\alpha}})), \tag{5}$$

where $\bar{f}(\cdot)$ and the subscript $\boldsymbol{\alpha}$ denote the frozen student network and weakly augmented data, respectively. Herein, the prediction of $u_{\boldsymbol{\alpha}}$ plays a role as an artificial label for the unlabeled sample. Lastly, L_{TS} is the mean squared error (MSE) loss between the predictions of teacher and student networks:

$$\begin{aligned} L_{TS} =& \mathbb{E}||f_w(l_{\mathcal{A}}) - \bar{g}_w(l_{\mathcal{A}})||^2 + \mathbb{E}||f_s(l_{\mathcal{A}}) - \bar{g}_s(l_{\mathcal{A}})||^2 + \\ & \mathbb{E}||f_w(u_{\mathcal{A}}) - \bar{g}_w(u_{\mathcal{A}})||^2 + \mathbb{E}||f_s(u_{\mathcal{A}}) - \bar{g}_s(u_{\mathcal{A}})||^2, \end{aligned} \tag{6}$$

where $\bar{g}(\cdot)$ is the MT network for the student. The total loss is obtained by summing the three loss terms as follows:

$$L_{Total} = L_{sup} + L_{usp} + r(t)L_{TS}. \tag{7}$$

4. Experiments and Results

4.1. Dataset and experimental setup

The DESED database[1] was used for evaluating our proposed method. It consists of real recordings from Audioset [19] and synthetic audios, and each audio is labeled on the frame-level, clip-level, or unlabeled. The details of the size of the dataset are shown in Table 1. Each audio clip was 10 s long with either 44.1 kHz or 16 kHz and had ten predefined target sounds.

For the experiments, each audio was down-mixed to 16 kHz. Subsequently, we extracted the reference log-mel spectrograms and the corresponding strong labels with the window size, shift size, and mel-filter banks with 128 ms, 16 ms, and 128, respectively. We optimized the networks with an AdamW [20] optimizer, and the weight decay was 0.001. The learning rate was exponentially warmed up to 0.001 for the first 50 epochs and annealed to 0 for the other 50 epochs. Regardless of the weak and strong augmentation, the mixup was applied. Convolutional recurrent neural network (CRNN) was used for the network. We substituted the CNN part of the DCASE challenge baseline [21] with EfficientNetv2 [22] (EffiCRNN). The detailed structure of CNN is described in Table 2. The total number of parameters was 2.15 M.

[1] https://dcase.community/challenge2022/task-sound-event-detection-in-domestic-environments

Table 1: *Details of the number of audio clips of the DESED. (·) denotes the number of batch size. (R: real, S: synthesized audio)*

Split	Frame-level	Clip-level	Unlabeled
Train	R: 3,470 (8), S: 10,000 (8)	R: 1,578 (4)	R: 14,412 (40)
Eval.	R: 1,168	-	-

Table 2: *Specification of the CNN part of EffiCRNN. Except for the first layer of each stage, the filter sizes and strides were (3x1) and (1, 1).*

Stage	Operator	#Channels	Stride	#Layers
0	Conv 5x5	16	(1, 1)	1
1	Fused-MBConv1 3x3	24	(2, 2)	2
2	Fused-MBConv4 3x3	32	(2, 2)	4
3	Fused-MBConv4 3x3	48	(2, 1)	4
4	MBConv4 3x3	64	(2, 1)	4
5	MBConv4 3x3	128	(2, 1)	4
6	MBConv4 3x3	128	(2, 1)	4
7	Avg. pool & Conv 1x1	128	(1, 1)	1

4.2. Evaluation metrics

For the validation of PSED systems, we used four metrics: event-f1 score [23], intersection-f1 score, and the polyphonic sound detection score (PSDS) [24] with two settings. Event-f1 score is a collar-based evaluation metric to validate the onset and offset of detected sounds. PSDS is AUC of polyphonic sound detection ROC curve [24], which validates how much the detected sounds overlapped with the ground truth. Also, the intersection-f1 score is PSDS metric-based f1 score used in DCASE challenges[1]. The intersection tolerance of the PSDS1 was higher than that of PSDS2; therefore, the PSDS1's criteria were more severe so it could evaluate how correctly the system detects timestamps [25, 26]. All settings for metrics were the same as the recent DCASE challenge task 4[1].

4.3. Comparison to other methods

To validate the superiority of the proposed method, we compared it with other SSL methods (Table 3). ResCT outperformed the others on all scores except the PSDS2. Importantly, the *event-f1 score* and *PSDS1* were notably enhanced. It demonstrates that a few temporal perturbations increase robustness of how the system captures the onset and offset of detected sounds during temporal ResCT [25,26]. In addition, as shown in Fig. 4, ResCT was effective in alleviating overfitting. In other words, the decision boundaries become smoother than that of other SSL strategies.

The proposed multi-resolutional augmentations were compared to other augmentations that did not significantly distort the temporal labels (Table 4). Among the single augmentations, the proposed method generally outperformed other augmentations. When the proposed method was combined with frequency masking, the performance improved for event-f1 score

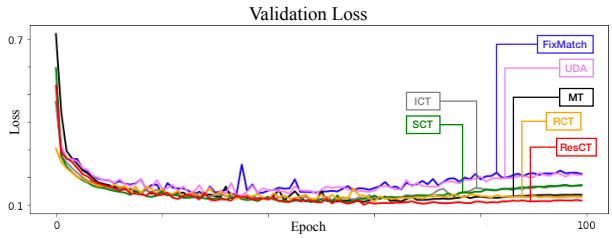

Figure 4: *Validation loss of systems as a monitor while training.*

Table 3: *Detecting performances for the different SSL strategies. FixMatch and UDA used the same proposed augmentation method of the ResCT. The thresholds for the confidence based masking of FixMatch and UDA were set to 0.9 for activated frames, and 0.1 for inactivated frames. Also, the temperature of sigmoid for sharpening was set to 0.4 for UDA. The configurations of RCT were the same as in [2]. (Inter.-f1: intersection-f1)*

Methods	Event-f1	Inter.-f1	PSDS1	PSDS2
Supervised	32.50	54.74	15.32	25.15
MT	47.42	72.97	35.1	61.59
UDA	46.39	70.65	32.15	54.83
FixMatch	47.55	69.96	32.43	54.93
ICT	46.45	68.98	34.03	59.09
SCT	49.10	72.70	36.49	59.91
RCT	48.13	73.27	34.62	**65.32**
ResCT	**52.28**	**74.88**	**37.69**	61.26

and PSDS1, whereas the Gaussian noise was less effective.

4.4. Ablation studies

We investigated which combination and sampling distribution would be better for the ResCT. The performances according to the combinations of resolution policies are shown in Table 5. Typically, controlling the sampling rate to the weak augmentation had better performance than controlling the window size. In other words, the weak augmented data roles an artificial label for the unlabeled data; thus, controlling the window size could produce inconsistent temporal labels.

Further, how much the spectrogram should be distorted for the ResCT is shown in Table 6. We experimented with three probability distributions; log-Gaussian, Beta (0.5, 0.5) and uniform distributions (the difference in probability density functions of the three distributions is shown in Fig. 5). The Gaussian distribution had the best performance, but that of the Beta distribution was the worst (*even worse than the other methods shown in Table 3*); these performances demonstrate that fewer

Table 4: *Detecting performances for the different strong augmentation policies. **Res.**, **FM** and **GN** denotes the proposed policy, frequency masking and adding Gaussian noise. The size of masking was under 20 bands, and the Gaussian noise was randomly added from 6 to 30 SNRs.*

Augmentation	Event-f1	Inter.-f1	PSDS1	PSDS2
Res.	**52.28**	**74.88**	**37.69**	61.26
FM	50.46	73.20	37.52	**61.45**
GN	48.49	71.15	35.26	58.34
Res. + FM	**53.03**	**74.75**	37.79	59.76
Res. + GN	51.65	74.42	36.88	60.46
Res. + FM + GN	51.8	74.02	**38.31**	**61.71**

Table 5: *Ablation study on different resolution policies for weak- and strong-augmentation. We controlled the combinations so that strong augmentations became more distorted than weak ones. (**Win**: Window sizes, **Sr**: Sampling rates)*

Weak	**Strong**	Event-f1	Inter.-f1	PSDS1	PSDS2
-	**-**	49.61	71.21	33.54	52.89
-	**Sr**	49.82	71.83	34.55	56.22
-	**Win**	49.38	71.58	33.09	53.17
-	**Sr+Win**	49.37	72.03	35.18	56.15
Win	**Sr**	50.39	70.09	36.00	57.14
Win	**Sr+Win**	49.24	71.56	35.41	55.47
Sr	**Win**	51.78	74.50	37.48	60.03
Sr	**Sr+Win**	**52.28**	**74.88**	**37.69**	**61.26**

Table 6: *Ablation study on sampling policies from different probability distributions for weak- and strong-augmentation. Compared to the uniform distribution, the density of the Gaussian is high at the center, but that of Beta is high at both ends.*

Distribution	Event-f1	Inter.-f1	PSDS1	PSDS2
Log-Gaussian	**52.28**	**74.88**	**37.69**	**61.26**
Log-Beta (0.5, 0.5)	41.99	66.15	31.59	48.06
Log-uniform	46.05	67.41	33.25	53.28

perturbations made decision boundaries smoother, but too many perturbations created too many distortions to regularize consistency and mismatch with the ground truth, degrading performance.

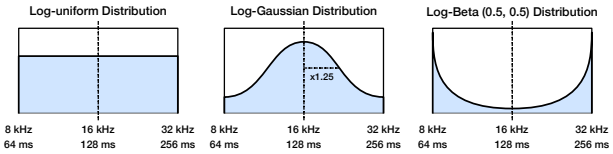

Figure 5: *Probability density functions (PDF) for log-uniform, Gaussian, and Beta (0.5, 0.5) distribution. For the log-Gaussian PDF, most of the observations cluster around the mean; however, for the log-Beta (0.5, 0.5) PDF, the probabilities for values are higher further from the mean.*

5. Conclusion

In this study, we introduced ResCT, which makes model predictions consistent across multi-resolutions with an unsupervised term combined MT algorithm. The time and frequency resolutions of a spectrogram were changed by controlling sampling rates and window sizes. We demonstrated that a little temporal and spectral noise made the PSED system more robust, especially in detecting the onset and offset of sound events. In addition, we expect that the ResCT's structure and proposed augmentation methods could be further developed by combining with other SSL methods.

6. Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation)

7. References

[1] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Advances in neural information processing systems*, vol. 30, 2017.

[2] N. Shao, E. Loweimi, and X. Li, "RCT: Random consistency training for semi-supervised sound event detection," in *Proc. Interspeech*, 2022, pp. 1541–1545.

[3] Y. Liang, Y. Long, Y. Li, and J. Liang, "Selective Pseudo-labeling and Class-wise Discriminative Fusion for Sound Event Detection," in *Proc. Interspeech*, 2022, pp. 1496–1500.

[4] S. Park, S. R. Kothinti, and M. Elhilali, "Temporal coding with magnitude-phase regularization for sound event detection," in *Proc. Interspeech*, 2022, pp. 1536–1540.

[5] X. Zheng, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "An effective mutual mean teaching based domain adaptation method for sound event detection," in *Proc. Interspeech*, 2021, pp. 556–560.

[6] V. Verma, A. Lamb, J. Kannala, Y. Bengio, and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *Proc. the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 3635–3641.

[7] 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 *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2021, pp. 376–380.

[8] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Proc. Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6256–6268.

[9] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Proc. Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 596–608.

[10] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, "Randaugment: Practical data augmentation with no separate search," *arXiv preprint arXiv:1909.13719*, vol. 2, no. 4, p. 7, 2019.

[11] D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *Proc. International Conference on Learning Representations* (*ICLR*), 2019.

[12] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[13] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *arXiv preprint arXiv:1511.06709*, 2015.

[14] S. Grollmisch and E. Cano, "Improving semi-supervised learning for audio classification with fixmatch," *Electronics*, vol. 10, no. 15, p. 1807, 2021.

[15] 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," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[16] D. de Benito-Gorron, S. Segovia, D. Ramos, and D. T. Toledano, "Multiple feature resolutions for different polyphonic sound detection score scenarios in dcase 2021 task 4." in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), 2021, pp. 65–69.

[17] H. Y. Kim, J. W. Yoon, S. J. Cheon, W. H. Kang, and N. S. Kim, "A multi-resolution approach to gan-based speech enhancement," *Applied Sciences*, vol. 11, no. 2, p. 721, 2021.

[18] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 6199–6203.

[19] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2017, pp. 776–780.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. International Conference on Learning Representations* (*ICLR*), 2018.

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

[22] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *Proc. International Conference on Machine Learning*, 2021, pp. 10 096–10 106.

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

[24] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta, and S. Krstulović, "A framework for the robust evaluation of sound event detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2020, pp. 61–65.

[25] G. Ferroni, N. Turpault, J. Azcarreta, F. Tuveri, R. Serizel, Ç. Bilen, and S. Krstulović, "Improving sound event detection metrics: insights from dcase 2020," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), 2021, pp. 631–635.

[26] W.-G. Choi and J.-H. Chang, "Confidence regularized entropy for polyphonic sound event detection," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events* (*DCASE*), Nancy, France, Nov. 2022.