DCASE 2024 TASK 4: SOUND EVENT DETECTION WITH HETEROGENEOUS DATA AND MISSING LABELS

Samuele Cornell[1,*], *Janek Ebbers*[2,*], *Constance Douwes*[3],
Irene Martín-Moratő[4], *Manu Harju*[4], *Annamaria Mesaros*[4], *Romain Serizel*[3]

[1]Carnegie Mellon University, USA [2]Mitsubishi Electric Research Laboratories, USA
[3] Universite de Lorraine, CNRS, Inria, Loria, Nancy, France [4] Tampere University, Finland

ABSTRACT

The Detection and Classification of Acoustic Scenes and Events Challenge Task 4 aims to advance sound event detection (SED) systems in domestic environments by leveraging training data with different supervision uncertainty. Participants are challenged in exploring how to best use training data from different domains and with varying annotation granularity (strong/weak temporal resolution, soft/hard labels), to obtain a robust SED system that can generalize across different scenarios. Crucially, annotation across available training datasets can be inconsistent and hence sound labels of one dataset may be present but not annotated in the other one and vice-versa. As such, systems will have to cope with potentially missing target labels during training. Moreover, as an additional novelty, systems will also be evaluated on labels with different granularity in order to assess their robustness for different applications. To lower the entry barrier for participants, we developed an updated baseline system with several caveats to address these aforementioned problems. Results with our baseline system indicate that this research direction is promising and is possible to obtain a stronger SED system by using diverse domain training data with missing labels compared to training a SED system for each domain separately.

Index Terms— Sound event detection, missing labels, efficiency, weak supervision, heterogeneous data

1. INTRODUCTION

It can be argued that, with current deep learning based techniques, the ability to leverage as much training data as possible is as important as the pursue of novel (in the methodological sense) techniques. For example, the effectiveness of modern large-language models (LLMs) relies mostly on the scale of the training data rather than on their deep neural network (DNN) architecture. The same is true for automatic speech recognition (ASR) models, with recent works [1–3] demonstrating that a great deal of robustness, as well as zero-shot and emerging capabilities [1], come both from the scale of the model and, crucially, the size of the training set.

However, leveraging data at scale has its own set of challenges. This is particularly true for SED where readily available data and metadata is not readily obtainable from web sources. While self-supervised learning (SSL) techniques [4–7] can help to circumvent this issue, still, supervised data is necessary for fine-tuning. For this latter, the only viable option right now is hand annotation, which is very expensive and difficult to scale as SED requires temporal endpoints together with the class label. Hence, to lower the annotation burden, temporally *weak* annotations (i.e. presence or not of a sound event inside a particular audio clip of several seconds without precise endpoints) are often used in conjunction with a smaller portion of temporally precise (i.e. *strong*) annotated recordings [8, 9]. These latter are particular important, as it has been demonstrated [10, 11] that increasing the amount of strong-labeled examples brings considerably benefits in terms of performance, despite the obvious drawbacks of increasing the annotations costs. As such, in the recently proposed MAESTRO [12] dataset, a sliding window approach to the annotation procedure was developed. This approach, together with crowdsourcing, allows for better scaling in the annotation stage. In fact, in MAESTRO temporally strong labels are obtained by overlap-add of several temporally weak annotations.

This discrepancy in the annotation temporal granularity, since 2018, has been explored extensively in the past DCASE Task 4 challenges [9, 13–17], with DESED [18, 19] being the main dataset used through these all past editions.

However, another crucial issue is that, between different datasets, not only the temporal granularity (temporally strong vs. weak labels) can vary but also the consistency in the annotation procedure. I.e. which classes are considered as events of interest and which are instead disregarded, or again, if annotation confidence (i.e. the use of *soft* labels) is available or not. This direction has been largely underexplored in previous DCASE Task 4 challenges but is essential towards the goal of leveraging as much as training data as possible and is the main novelty introduced this year.

2. MOTIVATION

This year the DCASE Challenge Task 4 aims at addressing 2 different aspects that are related to the aforementioned problem of leveraging diverse training data with missing and (temporally and/or posterior-wise) weak annotation. Each of these aspects answer fundamental research questions which are formulated in the following.

2.1. Can we combine datasets from diverse domains with different annotations to improve performance ?

As said in Section 1, one of the challenges of combining different datasets for SED is the fact that the two datasets may not have consistent annotation with one another. In extreme cases, the datasets might not even share any common sound event classes. Instead of training a SED model on each dataset separately an intriguing approach is to just train one model on all available datasets. Intuitively, if the two datasets have sound classes that overlap or, at least, some classes that could be mapped from one another (e.g. when one event is a sub-class of another event [20–22]), then we expect that using all datasets should afford better performance compared to training a model for each separately. However, since annotation can be inconsistent and some events that are annotated in one dataset may be present but not annotated in the other, the training procedure and

[*] These authors contributed equally to this work

possibly even the SED model must be modified to account for this issue. In Section 6 we describe how we addressed this when developing this year baseline system and in Section 7.2 we present also some results which indicate that this research direction is promising and indeed leads to large performance gains. Our hope is that participants will devise novel and more effective ways to address this important problem.

2.2. What is the best way to exploit soft labels ? Are they useful to improve performance ?

Some datasets, such as MAESTRO, due to their data annotation protocol, have soft labels expressing the annotators overall confidence of the presence or not of a particular sound event. In [12] it was shown that it is possible to train an effective SED system using such soft labeled annotation and two possible loss functions: binary cross entropy (BCE) and mean square error (MSE) were explored as well as different post-processing techniques. In particular, the choice of the loss function was found to affect the model performance on more rare occurring sound event classes. Several research questions however arise when soft labels are combined with strong labels from other datasets and with soft labels from pseudo labels obtained from the model (e.g. via mean-teacher [23]). I.e. it would be interesting to assess if annotation confidence metadata is useful for training a robust SED system when training data is scaled and if also other approaches e.g. filtering may be helpful or not.

3. CHALLENGE DATASETS

This year challenge keeps using the DESED dataset, in order to be comparable with previous editions, but also adds MAESTRO as another dataset participants can use and on which performance will be evaluated. Both are described in detail in the following.

DESED consists of 10 seconds length audio clips either recorded in a domestic environment or synthesized to reproduce such an environment. It features annotated sound events from 10 different classes: alarm_bell_ringing blender, cat, dishes, dog, electric_shaver_toothbrush, frying, running_water, speech, vacuum_cleaner. The synthetic part of the dataset is generated with Scaper [24] with foreground events obtained from the Freesound datasets [25] while backgrounds are extracted from YouTube videos under Creative Commons license and from the Freesound subset of the MUSAN dataset [26]. Such synthetic set is divided into an evaluation and training part. More information is available in [15]. The real-world recording part is instead derived from AudioSet [27] and it comprises of a temporally-weakly annotated set (1578 clips), a totally unlabeled set (14412 clips) and also a strongly annotated portion from [10] (3470 clips).

MAESTRO Real, which has been proposed in [12] and used in the past DCASE 2023 Task 4 (track B) challenge, consists of a development (6426 clips) and an evaluation part of long-form real-world recordings. This dataset contains multiple temporally-strong annotated events with soft labels from 17 classes. However, in this challenge, out of these, only 11 are considered in evaluation as the other 6 do not occur with confidence over 0.5. These classes are: birds_singing, car, people_talking, footsteps, children_voices, wind_blowing, brakes_squeaking, large_vehicle, cutlery_and_dishes, metro_approaching, metro_leaving. As said, this data was annotated using crowdsourcing and the procedure introduced in [28], where temporally-weak labeling is used in conjunction to a sliding window approach to derive events temporal localization. Multiple annotators outputs are aggregated via MACE [29]. The recordings are derived from TUT Acoustic Scenes 2016 [30] dataset and are between 3 to 5 minutes long.

4. RULES

Rules are largely similar to previous year edition. However this year we allow participants to always use external data and pretrained models [1] without reporting also results for approaches that are fully trained only with DESED and MAESTRO. Another important difference is that, this year, since we have two scenarios, we prohibit domain identification. In fact we want participants to focus on approaches that can generalize across various scenarios without a-priori knowledge of which subset of sound classes can be present.

5. EVALUATION

SED evaluation assesses a system's capability of recognizing and temporally localizing sound events. Currently three different event-matching approaches exist namely collar- [31], intersection- [32,33] and segment-based [31], which differ in the way they compare predicted and ground truth temporal locations of sound events. In recent years, intersection-based evaluation has gained popularity as an event-based metric favoring detection of reasonably connected events, while being less sensitive to annotation ambiguities compared to collar-based evaluations. Further, there is a high variation in SED application requirements, with some applications requiring a high recall, others a high precision, and yet others may even let the user control sensitivity. Hence, an SED evaluation metric ideally aggregates performance over various operating modes.

Therefore, the polyphonic sound detection score (PSDS) [32, 34] has been used as primary metric in this task since 2021. It evaluates the normalized partial area under the PSD-ROC curve, where the PSD-ROC is the average of class-wise intersection-based ROC curves plus a penalty on inter-class standard deviation. PSDS parameters are the detection tolerance criterion ρ_{DTC} (the required intersection of a detected event with ground truth events to not be counted false positive (FP)), the ground truth intersection criterion ρ_{GTC} (the required intersection of a ground truth event with non-FP detected event), the penalty weight α_{ST} on inter-class standard deviation, and the maximum FP-rate e_{max} up to which the area under curve is computed [2]. In previous editions PSDS1 and PSDS2 have been evaluated, which differ in their parameters. This year we are considering only PSDS1 for evaluation with $\rho_{\text{DTC}} = \rho_{\text{GTC}} = 0.7$, $\alpha_{\text{ST}} = 1.$, $e_{\text{max}} = 100\,\text{FPs/hour}$, as PSDS2 is tuned more as an audio tagging than an SED metric. Event's onset and offset times required for PSDS computation, however, are only available for DESED data and classes which is why PSDS1 is only evaluated on this fraction of the evaluation set.

For MAESTRO, segment-based labels (segment length of one second) are provided, and we use the segment-based mean (macro-averaged) partial area under ROC curve (segMPAUC) as the primary metric instead with a maximum FP-rate of 0.1. segMPAUC is computed w.r.t. hard labels (using a binarization threshold of 0.5) for the 11 classes listed in Sec. 3.

To have a common processing of DESED and MAESTRO data during inference, we split MAESTRO recordings, which comprise several minutes, into clips of 10 seconds with a clip overlap of

[1] Allowed data and model resources are listed in the challenge website

[2] Cross-trigger parameters are not mentioned as not considered this year.

Data \ Classes	DESED	MAESTRO
DESED	PSDS1	not evaluated
MAESTRO	not evaluated	segMPAUC

Table 1: Primary metrics for different data and class fractions.

50%. DESED and MAESTRO clips are anonymized and shuffled in the evaluation set to prevent manual domain identification (cf. task rules in Sec. 4). At evaluation time, we reconstruct recording-level predictions from the MAESTRO clips by computing, for each class, a scalar posterior score in each segment. To do so, submitted (short-time) class posterior scores are averaged over the duration of a segment and further averaging segment-level scores of the same segment from overlapping clips.

Table 1 summarizes which primary metric is used on which data and class fractions. Additionally, we report segMPAUC on DESED, collar-based F_1-score on DESED, and macro-averaged segment-based F_1-scores and error rates (ERs) for a detection threshold of 0.5 and with optimal detection thresholds. All metrics are evaluated using sed_scores_eval[3]. As in previous editions, we use both the predictions from three independent training runs and bootstrapped evaluation [35] to compute a mean and standard deviation for each of the metrics. Here, 20 different bootstrap samples (whereby we ensure that each clip is overall sampled equally often) are evaluated for each of the three runs yielding 60 results to compute statistics from. As ranking metric the sum of the primary metrics' means $\overline{\text{PSDS1}}+\overline{\text{segMPAUC}}$ is used. Note, that both metrics are taken from the same system, as, in contrast to previous editions, both metrics focus on SED here.

Energy efficiency is another important factor in SED systems. Hence, as in the previous two editions, we ask participants to report the energy consumption of their system during both training and testing stages using the CodeCarbon package [36]. We also ask participants to report the energy consumption for training the baseline model on 10 epochs as well as for inference with the baseline model on the development set. This procedure has to be performed on the same hardware as used for their system training/inference such that energy consumption can be normalized among different hardware and provide fairer comparisons [17]. In addition, this year we ask not only CodeCarbon's total energy consumption, which is calculated as the sum of the three components (GPU, CPU, RAM), but also the energy from the GPU component alone. In fact, we found that CPU and RAM consumption were both included by CodeCarbon in previous DCASE Task 4 challenges, and are now also interested an accurate picture of the GPU energy. Having a more precise energy consumption estimation could allow to better assess the relationship between the number of multiply-accumulate (MAC) operations, the number of parameters and energy consumption from the GPU. In Section 7, Table 3 energy consumption figures are for the baseline are reported.

6. DCASE 2024 CHALLENGE TASK 4 BASELINE SYSTEM

The baseline system is directly inherited from the previous DCASE Task 4 challenges [17, 37] and consists of a convolutional recurrent neural network (CRNN) network which also employs self-supervisedly learned features from BEATs pre-trained model [6]. The CRNN model has a convolutional neural network (CNN) encoder of 7 convolutional layers with batch normalization, gated linear unit and dropout, followed by a bi-directional gated recurrent unit (biGRU) layer. Before this latter, BEATs features are concatenated with the CNN extracted ones. Average pooling is applied to BEATs features prior concatenation to make the sequence length the same as the one from the CNN encoder. Clip-wise and frame-wise posteriors are then derived using an attention pooling [38]. The CNN encoder is fed log-Mel filterbank energies extracted with a 128 ms window and 16 ms stride from 16 kHz audio. During training the BEATs model is kept frozen, Mixup [39] regularization strategy is employed and the mean-teacher framework [23] is used in order to leverage unlabeled and weakly-labeled data. Baseline code and pre-trained checkpoints are available online[4].

For this year challenge we introduced two incremental improvements and, to deal with the aforementioned missing labels problem also some ad-hoc modifications to the training procedure. Regarding the minor improvements for this year baseline we use SpecAugment-style [40] time-wise masking on the pre-trained model extracted features and, independently, on the features extracted from the CNN encoder. We denote this strategy as *dropstep* in Section 7.1. Another difference is that for post-processing we employ a multi-class median filter where each class has a different median filter length.

6.1. Dealing with partially annotated data

The training procedure had to be modified in several places in order to deal with the missing labels problem.

6.1.1. Cross mapping sound event classes

First, as a pre-processing step, we map some DESED events to similar classes in MAESTRO. More in detail, we have in DESED "speech" which is a super-class for "people_talking, children_voices, announcement" in MAESTRO, "dishes" which corresponds to "cutlery_and_dishes" and also "dog" which is a super-class for "dog_bark". Note that these mapping are from MAESTRO to DESED but not vice-versa as DESED ones are mostly super-classes of MAESTRO ones. Intuitively, with this strategy, when computing the loss on MAESTRO e.g. for a clip with the event "people_talking" having confidence 0.5, we also drive the network output posterior corresponding to "speech" class to 0.5.

6.1.2. Loss computation

The model is trained using BCE loss function on DESED real-world strongly, synthetic and weakly labeled examples as well as on MAESTRO soft labeled examples. MSE is instead used for the mean-teacher pseudo-labeling loss component which is applied on both weak and unlabeled data from DESED. When computing the loss for both components on a particular clip we avoid computing the loss for the network outputs corresponding to the classes that do not correspond to the clip original dataset. For example, for MAESTRO, we do not compute the loss for DESED output logits except for class that have been cross-mapped as explained before.

6.1.3. Attention-pooling masking

The attention pooling mechanism [38] employed in the final layer of the baseline model applies the softmax function over classes. Before taking the softmax, the values corresponding to unlabeled classes (not belonging to the current clip dataset) are masked to minus infinite in order to prevent to attend to them.

[3] https://github.com/fgnt/sed_scores_eval

[4] github.com/DCASE-REPO/DESED_task/recipes/dcase2024_task4_baseline

6.1.4. *Mixup*

Mixup [39] regularization strategy is applied for MAESTRO and DESED independently as labels are missing and the two cannot be mixed together in a reliable manner.

6.2. Hyperparameters tuning

We adopt a dual-phase approach to hyperparameters tuning in order to ease the computational burden of the overall tuning procedure. In the first step, we tune the network and training parameters [5]. This requires training the model from scratch for each set of selected hyperparameters. In detail we tune the number of biGRU layers and its hidden state size, learning rate, dropout and dropstep parameters, warmup epochs and gradient clipping value. In a second step, the network is kept frozen and we use the best model as found in the first step and tune only the multi-class median filter. This second step requires only to perform inference on the dev-test portions of the data[6]. Such dual-phase approach allows to dramatically reduce the required number of training runs compared to tuning everything together from scratch, since a slight change in the median filter length for a particular class has a significant effect on the performance of the overall system, leading to a very noisy hyperparameter tuning procedure. The hyperparameter tuning procedure was performed using the Optuna toolkit [41] using multi-objective tree-structured Parzen estimator [42] with dev-test PSDS1 + segMPAUC as the objective function.

7. EXPERIMENTAL RESULTS

7.1. Baseline improvements

In Table 2 top-panel, we report an ablation study to motivate the baseline system changes described in Section 6. We can observe that all the proposed changes bring substantial improvement. In particular, the dual-phase Optuna-based hyperparameter tuning (- HypTune ablations) appears to be quite effective. Adding a median filter (- HypTune 2 ablation, unprocessed scores) seems crucial, while having a multi-class median filter (- MC-Median ablation), improves performance only marginally. Compared to this latter, the dropstep regularization strategy has a more significant effect (- dropstep ablation).

Model	PSDS1 ↑	segMPAUC ↑
	Dev-test (DESED)	Dev-test (MAESTRO)
Random Init	0.0	0.02
Baseline	0.491	0.731
- dropstep	0.479	0.706
- HypTune1	0.458	0.669
- HypTune2	0.391	0.702
- MC-Median	0.485	0.714
- DESED	0.0	0.642
- MAESTRO	0.483	0.115
- CrossMap	0.469	0.722

Table 2: Baseline improvements ablation study on dev-test and effect of training the system only on DESED or MAESTRO data. For MAESTRO, we used 90% overlap when reconstructing the long-form audio.

Total kWh		GPU kWh	
Train	Dev-test	Train	Dev-test
1.180	0.119	0.113	0.013

Table 3: Baseline energy consumption in kWh for training and inferring on the development set on one A100 (40GB)

7.2. Leveraging heterogeneous datasets with missing labels

In Table 2 bottom-panel we report an ablation study to assess how removing one of the two datasets (MAESTRO or DESED) affects the overall performance of the SED system. We can see that, in both instances where the other dataset is removed, whether it is DESED (- MAESTRO ablation) or MAESTRO (- DESED ablation), the performance on the remaining dataset also drops. However, the performance drop is small if MAESTRO is removed. This is likely due to the fact that DESED is much larger and thus the effect of removing/adding MAESTRO is modest. The strategy described in Section 6 of mapping some MAESTRO classes to some DESED classes is considerably effective (- CrossMap ablation) in particular for DESED as one would expect (some MAESTRO classes are mapped to corresponding DESED super-classes). What is rather surprising is, instead, the fact that if DESED is removed (- DESED ablation), the performance on MAESTRO drops quite dramatically. In fact, as described in Section 6, during training, when both datasets are used, the loss on the classes that do not belong to the dataset from which the input audio is taken are masked, thus e.g. MAESTRO outputs are completely ignored when the input audio comes from DESED (we do not map any class from DESED to MAESTRO). We hypothesize that the addition of DESED data boosts significantly the performance on MAESTRO because it may help the model to learn how to extract a more meaningful and generalizable representation especially in the earlier layers of the network, acting as a regularization strategy (especially important as MAESTRO is small compared to DESED). This hypothesis may also explain why if we remove the class mapping (- CrossMap ablation) the performance on MAESTRO is still superior to using MAESTRO alone.

8. CONCLUSIONS

In this paper we presented the DCASE 2024 Task 4 challenge which addresses the important problem of leveraging multiple data sources for training SED systems. Datasets can differ by the temporal resolution of the labels e.g. temporally *strong* or *weak* labels or by the fact that annotator confidence may be present (e.g. *soft* labels) or not, or again, by which sound classes are actually considered during the annotation process. To spur research towards addressing these issues, this year task involves two datasets DESED and MAESTRO on which participants systems are benchmarked, while external data and pre-trained models can also be leveraged. Due to the aforementioned annotation inconsistencies participants will need to devise novel and effective ways to cope with the fact that sound events that are considered in DESED may be present in MAESTRO but are not annotated and vice versa. To ease the challenge participation entry barrier, an updated baseline system was developed. Results from such baseline suggest that leveraging more data, if the aforementioned problems are addressed in a reasonable way, is always beneficial. In fact, we show that it is possible to obtain a system trained on multiple datasets which is stronger than single systems that are trained on each dataset/scenario independently.

[5] Script available at: dcase2024_task4_baseline/optuna_pretrained.py

[6] The optimized class-wise median filters lengths are in dcase2024_task4_baseline/confs/default.yaml

9. REFERENCES

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[2] Y. Peng, J. Tian, B. Yan, D. Berrebbi, X. Chang, X. Li, J. Shi, S. Arora, W. Chen, R. Sharma, *et al.*, "Reproducing whisper-style training using an open-source toolkit and publicly available data," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[3] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[4] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.

[5] A. Baade, P. Peng, and D. Harwath, "Mae-ast: Masked autoencoding audio spectrogram transformer," 2022.

[6] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in *International Conference on Machine Learning*. PMLR, 2023, pp. 5178–5193.

[7] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 708–28 720, 2022.

[8] 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.

[9] R. Serizel, N. Turpault, H. Eghbal-Zadeh, and A. Parag Shah, "Large-Scale Weakly Labeled Semi-Supervised Sound Event Detection in Domestic Environments," in *DCASE Workshop*, 2018.

[10] S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[11] F. Ronchini, R. Serizel, N. Turpault, and S. Cornell, "The impact of non-target events in synthetic soundscapes for sound event detection," *arXiv preprint arXiv:2109.14061*, 2021.

[12] I. Martín-Morató, M. Harju, P. Ahokas, and A. Mesaros, "Training sound event detection with soft labels from crowdsourced annotations," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[13] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," in *DCASE Workshop*, 2020.

[14] N. Turpault, R. Serizel, A. Parag Shah, and J. Salamon, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *DCASE Workshop*, 2019.

[15] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *Proc. of ICASSP*, 2020.

[16] F. Ronchini, S. Cornell, and N. e. a. Turpault, "DCASE 2021 Task 4 Challenge," https://dcase.community/challenge2021, 2021.

[17] F. Ronchini, S. Cornell, R. Serizel, N. Turpault, E. Fonseca, and D. P. Ellis, "Description and analysis of novelties introduced in dcase task 4 2022 on the baseline system," *DCASE Workshop*, 2022.

[18] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *Proc. of ICASSP*, 2020.

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

[20] H. Shrivastava, Y. Yin, R. R. Shah, and R. Zimmermann, "Mt-gcn for multi-label audio-tagging with noisy labels," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 136–140.

[21] G. Wichern, B. Mechtley, A. Fink, H. Thornburg, and A. Spanias, "An ontological framework for retrieving environmental sounds using semantics and acoustic content," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2010, pp. 1–11, 2010.

[22] A. Shah, L. Tang, P. H. Chou, Y. Y. Zheng, Z. Ge, and B. Raj, "An approach to ontological learning from weak labels," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[23] 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.

[24] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *Proc. of WASPAA*, 2017.

[25] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

[26] D. Snyder, G. Chen, and D. Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[27] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[28] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 902–914, 2023.

[29] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with mace," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 1120–1130.

[30] A. Mesaros, T. Heittola, and T. Virtanen, "Tut database for acoustic scene classification and sound event detection," in *EUSIPCO*, 2016.

[31] ——, "Metrics for polyphonic sound event detection," *Applied Sciences*, 2016.

[32] Ç. Bilen, G. Ferroni, and F. e. a. Tuveri, "A framework for the robust evaluation of sound event detection," in *Proc. of ICASSP*, 2020.

[33] 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*. IEEE, 2021, pp. 631–635.

[34] J. Ebbers, R. Haeb-Umbach, and R. Serizel, "Threshold independent evaluation of sound event detection scores," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 1021–1025.

[35] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. Chapman and Hall/CRC, 1994.

[36] V. Schmidt, K. Goyal, and A. J. et al., "CodeCarbon: Estimate and Track Carbon Emissions from Machine Learning Computing," 2021.

[37] F. Ronchini, J. Ebbers, F. Angulo, D. Perera, S. Essid, and R. Serizel, "DCASE 2023 Task 4a Challenge," https://dcase.community/challenge2023, 2023.

[38] L. JiaKai, "Mean teacher convolution system for dcase 2018 task 4," DCASE2018 Challenge, Tech. Rep., 2018.

[39] H. Zhang, M. Cisse, and Y. N. e. a. Dauphin, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[40] 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.

[41] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[42] Y. Ozaki, Y. Tanigaki, S. Watanabe, M. Nomura, and M. Onishi, "Multiobjective tree-structured parzen estimator," *Journal of Artificial Intelligence Research*, vol. 73, pp. 1209–1250, 2022.