A HYBRID SYSTEM OF SOUND EVENT DETECTION TRANSFORMER AND FRAME-WISE MODEL FOR DCASE 2022 TASK 4

Technical Report

Yiming Li[1,2], *Zhifang Guo*[1,2], *Zhirong Ye*[1,2], *Xiangdong Wang*[1,†], *Hong Liu*[1], *Yueliang Qian*[1], *Rui Tao*[3], *Long Yan*[3], *Kazushige Ouchi*[3]

[1] Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, eamon.y.li@gmail.com, {guozhifang21s, yezhirong19s, xdwang, hliu, ylqian}@ict.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Toshiba China R&D Center, Beijing, China, {taorui, yanlong}@toshiba.com.cn, kazushige.ouchi@toshiba.co.jp

ABSTRACT

In this technical report, we describe in detail our system for DCASE 2022 Task4. The system combines two considerably different models: an end-to-end Sound Event Detection Transformer (SEDT) and a frame-wise model (MLFL-CNN). The former is an event-wise model which learns event-level representations and predicts sound event categories and boundaries directly, while the latter is based on the widely-adopted frame-classification scheme, under which each frame is classified into event categories and event boundaries are obtained by post-processing such as thresholding and smoothing. For SEDT, self-supervised pre-training using unlabeled data is applied, and semi-supervised learning is adopted by using an online teacher, which is updated from the student model using the EMA strategy and generates reliable pseudo labels for weakly-labeled and unlabeled data. For the frame-wise model, the ICT-TOSHIBA system of DCASE 2021 Task 4 is used, which incorporates techniques such as focal loss and metric learning into a CNN model to form the MLFL-CNN model, adopts mean-teacher for semi-supervised learning, and uses a tag-condition CNN model to predict final results using the output of MLFL-CNN. Experimental results show that the hybrid system considerably outperforms either individual model, and achieves psds1 of 0.420 and psds2 of 0.783 on the validation set without external data. The code is available at *https://github.com/965694547/Hybrid-system-of-frame-wise-model-and-SEDT*.

Index Terms— Sound Event Detection Transformer, Online Pseudo-labelling, Hybrid System

1. INTRODUCTION

Sound Event Detection (SED) aims at identifying the category of foreground sound events as well as their corresponding onset and offset timestamps. Task4 of the DCASE challenge has been focusing on weakly supervised SED for several years. The DCASE 2022 Task4 [1] is a follow up of DCASE 2021 Task4 [2], while having some novel characteristics. In addition to exploring a heterogeneous development dataset containing unlabeled data, synthetic data and weakly labeled data, participants are encouraged to incorporate external dataset or pre-trained embeddings. As last year, the SED system will be evaluated by Polyphonic Sound Detection Score (PSDS) [3] under two different real-life settings.

For weakly supervised SED, most existing works follow the Multiple Instance Learning (MIL) framework, and formulate SED as a seq2seq classification task. They usually design Convolutional Neural Networks (CNNs) or Convolutional Recurrent Neural Networks (CRNNs) to obtain frame-level classification probability and then apply pooling mechanism to aggregate frame-level predictions to event-level results. However, such methods do not take sound events as a whole, which may impose limitations on their detection performance. Recently, an event-wise model, namely SEDT, is proposed to handle such problems [4]. It models SED as a set prediction problem, which directly maps audio spectrogram to a set of candidate events, thus freeing SED models from trivial post-processing, namely frame-level thresholding or median filtering. Empirical study has shown that SEDT can achieve competitive performance compared with its frame-wise counterparts [4]. Moreover, we find that the two models can supplement each other, as they solve SED task in different ways. Therefore, combining them together may be an intuitive approach to reach promising sound event detection performance.

In this report, we describe our system participating in DCASE 2022 Task 4. It is a combination of SEDT and frame-wise CNN model. For SEDT, special-designed training formulas, including supervised learning, self-supervised learning and semi-supervised learning, are studied to help it learn from the heterogeneous development dataset. For frame-wise CNN model, metric learning is applied to narrow the domain gap between real and synthetic data, mean-teacher framework is conducted to provide supervision for unlabeled data and a tag-conditioned CNN model is used to generate final predictions based on audio tags. After obtaining each well-trained model, we explore the fusion strategy and post-processing methods of the ensemble model. By using the methods above, the hybrid system achieves competitive results on validation dataset.

2. SEMI-SUPERVISED SEDT

2.1. Sound Event Detection Transformer

An overview of SEDT is shown in Fig. 1. It represents each sound event as a vector $y_i = (c_i, b_i)$, where c_i is the event category and

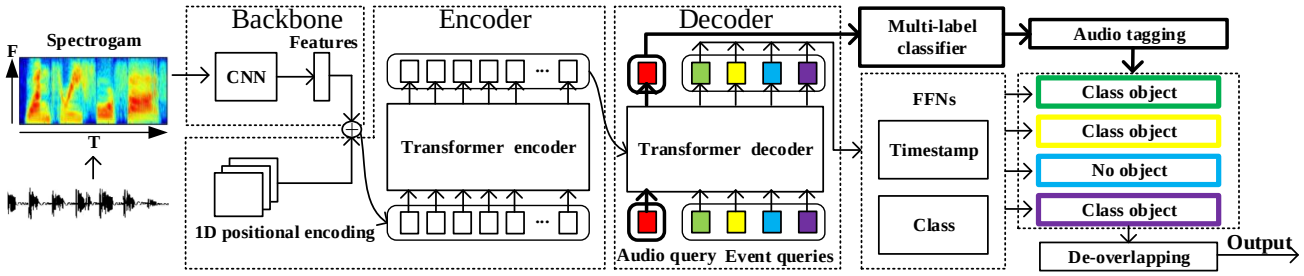

Figure 1: Overview of Sound Event Detection Transformer

b_i denotes the temporal boundary, and directly seeks a mapping between input features and ground-truth events. Given the input spectrogram, the backbone CNN is adopted to extract its feature map, which is then added with one-dimensional positional encoding and fed into transformer encoder for further feature process. The transformer decoder takes $N+1$ learned embeddings (N event queries and 1 audio query) as input, where each event query gathers information of a potential event from the encoder output feature via encoder-decoder cross-attention mechanism to generate event-level representations, and audio query gathers the whole audio information to generate clip-level representations. Finally, FFNs are utilized to transform the event-level representations and clip-level representations from the decoder into event detection and audio tagging results, which are then fused together to get the candidate detection results. De-overlapping is implemented on overlapped candidate events of the same category. Specifically, it only reserves the events with the highest class probability. More details can be found in [4].

2.2. Supervised learning for SEDT

SEDT incorporates event-level loss and clip-level loss to optimize its event detection and audio tagging performance. For strongly-labeled data, both loss terms will be involved during the SEDT model training, while for weakly-labeled data, the event-level loss will be excluded since the strong annotations are not available.

Event-level loss. SEDT introduces a novel label assignment scheme before computing event-level loss: it tries to find a matching $\hat{\sigma}_i$ between each event prediction $\hat{y}_i$ and its corresponding ground-truth annotation y_i through Hungarian algorithm. To equip SEDT with sound event classification and localization ability, the loss for SEDT supervised training is formulated as the weighted linear combination of localization loss $\mathcal{L}_{\text{loc}}$ and classification loss $\mathcal{L}_{\text{cls}}$. For each event prediction, the two loss functions are calculated as:

$$\mathcal{L}_{\text{loc}} = \sum_{i}^{N} \left(\lambda_{\text{IOU}} \mathcal{L}_{\text{IOU}} \left(b_i, \hat{b}_{\hat{\sigma}(i)} \right) + \lambda_{\text{L1}} \left\| b_i - \hat{b}_{\hat{\sigma}(i)} \right\|_1 \right) \quad (1)$$

$$\mathcal{L}_{\text{cls}} = \frac{1}{N} \sum_{i=1}^{N} -\log \hat{p}_{\hat{\sigma}(i)} \left(c_i \right) \quad (2)$$

where λ_{IOU} and λ_{L1} are corresponding weights for IOU Loss and L1 Loss.

Clip-level loss. The audio tagging loss is defined as the binary cross-entropy between the clip-level class label l_{tag} and predicted audio tagging y_{tag}:

$$\mathcal{L}_{\text{at}} = \text{BCE} \left(l_{\text{tag}}, y_{\text{tag}} \right) \quad (3)$$

2.3. Self-supervised learning for SEDT

To better use the unlabeled dataset and external datasets, such as AudioSet and SINS, we adopt a self-supervised learning method to pre-train SEDT on unlabeled data, which is named as Self-supervised Pre-training SEDT (SP-SEDT). Specifically, we randomly crop spectrogram along the time axis to obtain several patches, and then pre-train the model to predict corresponding temporal boundaries of the patches. To preserve the category information in SP-SEDT, classification loss and patch feature reconstruction loss are also adopted as sub-objective terms. By means of such pretext task, we hope that SEDT can localize sound event and maintain most category-related features at the same time. More details of SP-SEDT can be found in [5].

2.4. Semi-supervised learning for SEDT

Pseudo-labelling [6] is one of the mainstream approaches of semi-supervised learning. It requires a well-trained model to generate pseudo labels on unlabeled data, so that in the next stage, the converged model can be re-optimized on both labelled data and unlabeled data jointly. Based on that, we propose an improved pseudo-labelling method for the Semi-Supervised learning of SEDT(SS-SEDT). SS-SEDT splits the training process into two stages: the burn-in stage and the teacher-guided stage. In the burn-in stage, SEDT is simply trained on the labeled dataset to initialize the model. At the beginning of the teacher-guided stage, the initialized model is copied into two models (a student model and a teacher model), and then the teacher model generates pseudo labels on unlabeled data so that the student model can gain knowledge from both labeled data and unlabeled data. To guarantee the quality of the pseudo labels, we revisit the following off-the-shelf techniques, and apply them in the teacher-guided process.

- **EMA**: Unlike previous methods, we resort to a progressing teacher model to generate pseudo labels. The teacher model is updated from the student model through Exponential Moving Average (EMA). Thus, it can be viewed as implicit ensemble models and provide more accurate pseudo labels.
- **Strong and weak augmentation**: Strong and weak augmentation has been widely applied in the field of semi-supervised image classification [7]. Inspired by that, we adopt similar idea in the teacher-guided stage, during which weakly-augmented (frequency mask and frequency shift) spectrograms are fed into the teacher model to obtain pseudo labels and the student model make predictions on the strong augmented (frequency mask, frequency shift, time mask and gaussian noise) version of the same data batch.

- **Mixup** [8]: We mix labeled data with ground-truth annotations and unlabeled data with pseudo annotations together, which is supposed to improve the model robustness to pseudo annotation noise and alleviate the overfitting problem in model training.
- **Focal loss** [9]: Focal loss is adopted to handle the unbalanced event categories in SED, without which the model may be overwhelmed by easily classified samples and produce biased outputs. It should be noted that focal loss is merely used in the teacher-guided stage, we believe such pattern may help our model learn from easy to difficult.

3. FRAME-WISE CNN MODEL

The pipeline of the frame-wise CNN model is illustrated in Fig. 2. At first, MLFL-CNN is preliminarily trained with weakly labeled data and strongly labeled synthetic data to acquire basic event detection and audio tagging ability. Then, it attaches pseudo strong labels to the weakly-labeled and unlabeled data, and the model is jointly trained with all these data in a self-training manner. Finally, the trained MLFL-CNN provides audio tags and strong pseudo labels for the weakly-labeled data and unlabeled data to train the tag-conditioned CNN [10], which gives the final SED results.

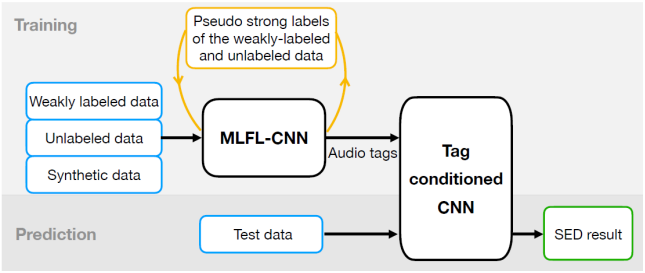

Figure 2: Overview of the frame-wise model

The MLFL-CNN model contains three branches. The first branch is the embedding-level attention pooling branch based on the MIL framework, which is the same with [11]. The second branch is the sound event detection branch which is introduced to exploit the strong labels of synthetic data and uses the focal loss as its loss function. The third branch is the domain adaptation branch which uses metric learning by inter-frame distance contrastive loss, more details of which can be found in [12]. During training process, the MLFL-CNN adopts the mean-teacher architecture and pseudo-labelling framework simultaneously [13]. It combines clip-level loss (for weakly-labeled data), frame-level loss (for data with strong labels and pseudo strong labels), inter-frame distance contrastive loss (for real data and synthetic data), and consistency loss together. And the tag-conditioned CNN takes spectrograms and audio tags predicted by the MLFL-CNN as inputs, and uses the strong labels of synthetic data and pseudo strong labels of real data as ground-truth to train.

4. FUSION OF THE TWO MODELS

4.1. Preliminary: Class-specific PSDS

The essence of PSDS is to obtain a function $r(e)$ of effective TP rate (eTPR) changing with effective FP rate (eFPR), and calculate the integral of this function over $(0, e_{\max})$, where $e_{\max}$ represents the maximum value of eFPR value [3]:

$$\mu_{\mathrm{TP}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}} r_{\mathrm{TP},c} \quad \sigma_{\mathrm{TP}} = \sqrt{\frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\left(r_{\mathrm{TP},c}-\mu_{\mathrm{TP}}\right)^2} \tag{4}$$

$$\text{eTPR:} \quad r(e) \triangleq \mu_{\mathrm{TP}}(e) - \alpha_{ST} * \sigma_{\mathrm{TP}}(e) \tag{5}$$

$$\mathrm{PSDS} \triangleq \frac{1}{e_{\max}}\int_0^{e_{\max}} r(e)de \tag{6}$$

The meanings of above notations are consistent with [3]. According to the above definition, we can decouple the eTPR according to the sound event category and redefine the PSDS value of given category:

$$\mu_{\mathrm{TP,c}} = r_{\mathrm{TP,c}} \quad \sigma_{\mathrm{TP,c}} = r_{\mathrm{TP,c}} - \mu_{\mathrm{TP,c}} \tag{7}$$

$$\text{eTPR}_c: \quad r_c(e) \triangleq \mu_{\mathrm{TP,c}}(e) - \alpha_{ST} * \sigma_{\mathrm{TP,c}}(e) \tag{8}$$

$$\mathrm{PSDS}_c \triangleq \frac{1}{e_{\max}}\int_0^{e_{\max}} r_c(e)de \tag{9}$$

where PSDS_c, eTPR_c, $\mu_{\mathrm{TP,c}}$ and $\sigma_{\mathrm{TP,c}}$ are corresponding class-wise indicators for specific event class c.

4.2. Model fusion method

The core of model fusion is to calculate the class-wise fusion coefficients of each model's prediction during the evaluation stage. Assume that there are N models $m_i (i = 1, 2, \ldots N)$, for each sound event class c, the PSDS of model m_i on c is denoted as $\mathrm{PSDS}_{i,c}$. Then the fusion coefficient of model i on category c is defined as:

$$w_{i,c} = \frac{\mathrm{PSDS}_{i,c}}{\sum_{i=1}^{N}\mathrm{PSDS}_{i,c}} \tag{10}$$

Therefore, for specific event category c, the final fusion probability $\hat{p_c}$ is formulated as the weighted linear combination of each model's predicted probability $p_{i,c}$:

$$\hat{p_c} = \sum_{i=1}^{N} w_{i,c} * p_{i,c} \tag{11}$$

It is noteworthy that the above PSDS in Eq.(10) can be interpreted as PSDS1 or PSDS2 for this year's DCASE task4, so two different sets of parameters $w_{i,c}$ can be obtained on the development set and utilized to improve PSDS1 and PSDS2 respectively. Besides, in our submitted hybrid system, event-level predictions of SEDT are firstly converted into frame-level probability, before being fused with frame-wise model to obtain the final results.

5. POST-PROCESSING

In order to reduce the noise in frame-level probability and make sound events continuous, it is necessary to perform a smoothing operation on the frame-level probability. Smoothing operation usually adopts mean filter or median filter. Currently, median filtering with a fixed window length or with the average length of each event calculated on the development set is generally utilized [14]. In this paper, we perform median filtering and mean filtering (with larger window size) on frame-level probabilities in sequence, and propose a method to search for optimal class-wise window lengths on the development set.

Table 1: The PSDS on validation set

System	Extra data	PSDS1	PSDS2
Baseline 1		0.336	0.536
Baseline 2	✓	0.351	0.552
System 1	✓	0.449	0.645
System 2	✓	0.115	0.816
System 3		0.420	0.618
System 4		0.099	0.783

Table 2: Ablation study on techniques in SS-SEDT

MU	FL	SWA	EMA	PSDS1	PSDS2
	✓	✓	✓	0.372	0.570
✓		✓	✓	0.349	0.540
✓	✓		✓	0.369	0.566
✓	✓	✓		0.357	0.538
✓	✓	✓	✓	0.388	0.573

Specifically, for a given event class c, we enumerate window length wl_c from the shortest frame length 1 to the maximum length 500, and find the optimal window length ${wl_c}^*$:

$$ {wl_c}^* = \arg\max_{wl_c} \frac{\mathrm{PSDS}_c}{\mathrm{PSDS}} \tag{12} $$

Similar to Section 4, two different sets of optimal window length ${wl_c}^*$ can be obtained on the development set and applied to optimize PSDS1 and PSDS2 respectively.

6. EXPERIMENT

6.1. Experiment Setup

For SEDT not using external data, we firstly pre-train it on unlabeled real subset (14412 clips), then simply train it on the weakly labeled training set (1578 clips) and synthetic 2019 subset (2045 clips) during burn-in stage, and finally use weakly labeled training set, synthetic 2019 subset, synthetic 2021 subset (10000 clips), and unlabeled real subset to conduct teacher-guided learning. For SEDT using external data, the two main differences compared to the above lie in: 1.models are pre-trained on both unlabeled real subset and SINS subset (72894 clips); 2. an additional strongly labeled set (3470 clips) is further included in the teacher-guided stage. The detailed settings of training hyper-parameters and configurations can be found in [15].

For frame-wise model not using external data, the training set contains the weakly labeled training set, the unlabeled training set, and synthetic 2021 subset. While for systems using external data, we add the same strongly labeled set taken from AudioSet to the original strong labeled set. The detailed settings of training hyper-parameters and configurations can be found in [16].

6.2. Results of Submitted Systems

Table. 1 shows the performance of our submitted systems, all of which are fused models of ensemble frame-wise CNN models and ensemble SEDT. Among them, system 1 and 2 incorporate external data to train the models, while system 3 and 4 do not. Besides, model fusion and window tuning methods proposed in Section 4 and Section 5 are utilized in system 1, 3 to improve their PSDS1 and in system 2, 4 to improve their PSDS2. As shown in Table. 1, our hybrid systems outperform the baseline considerably whatever the usage of external data.

Table 3: Ablation study on window tuning and model fusion

Id	Model	MF	WT	PSDS1	PSDS2
1	Single SEDT			0.415	0.582
2	Ensemble SEDT			0.431	0.607
3	Single frame			0.349	0.668
4	Ensemble frame			0.392	0.673
5	Hybrid system	✓		0.437	0.740
6	Hybrid system	✓	✓	0.449	0.816

6.3. Ablation Study

6.3.1. Techniques in SS-SEDT

To verify the effectiveness of techniques in SS-SEDT, we conduct ablation study using SS-SEDT without external data. Table. 2 shows the results of models trained without specific technique, where MU, FL, SWA denotes Mixup, Focal Loss, Strong and Weak Augmentation mentioned in Section 2.4, and the model without SWA means that the input spectrograms of teacher model and student model are both weakly augmented. It can be seen that all techniques can improve the performance of SS-SEDT and SS-SEDT can finally reach a PSDS1 of 0.388 and a PSDS2 of 0.573 while incorporating all techniques.

6.3.2. Window tuning and model fusion

To investigate the effects of window tuning and model fusion strategy, we conduct ablation study using SEDT and frame-wise model trained with external data. Table. 3 compares the performance between models under different settings. In the above table, MF and WT denotes Model Fusion and Window Tuning methods proposed in Section 4 and Section 5 respectively, and frame-wise model is abbreviated to "frame". Among all these models, model 2 and 4 are ensemble models of top 1-5 single models, while hybrid system represents the fused model of ensemble SEDT and ensemble frame-wise model. By comparing model 1, 2 with model 3, 4, it is obvious that SEDT can achieve higher PSDS1 while frame-wise model is better at PSDS2. Moreover, by comparing model 5 with model 2, 4, we can see that while SEDT and frame-wise model have their own edges, they can complement each other, since the hybrid system achieve further improvements of 0.006 and 0.067 compared to single ensemble models. By comparing model 6 with model 5, the effectiveness of window tuning can be validated, since model 6 provides the best PSDS1 (0.449) and PSDS2 (0.816).

7. REFERENCES

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

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

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

[4] Z. Ye, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, and K. Ouchi, "Sound event detection transformer: An event-based end-to-end model for sound event detection," *arXiv preprint arXiv:2110.02011*, 2021.

[5] Z. Ye, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, and K. Ouchi, "Sp-sedt: Self-supervised pre-training for sound event detection transformer," *arXiv preprint arXiv:2111.15222*, 2021.

[6] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[7] 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," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[8] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.

[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[10] J. Ebbers and R. Haeb-Umbach, "Forward-backward convolutional recurrent neural networks and tag-conditioned convolutional neural networks for weakly labeled semi-supervised sound event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop*, 2020, pp. 41–45.

[11] L. Lin, X. Wang, H. Liu, and Y. Qian, "Specialized decision surface and disentangled feature for weakly-supervised polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1466–1478, 2020.

[12] Y. Huang, L. Lin, X. Wang, H. Liu, Y. Qian, M. Liu, and K. Ouchi, "Learning generic feature representation with synthetic data for weakly-supervised sound event detection by inter-frame distance loss," *arXiv preprint arXiv:2011.00695*, 2020.

[13] R. Tao, L. Yan, K. Ouchi, and X. Wang, "Couple learning: Mean teacher method with pseudo-labels improves semi-supervised deep learning results," *arXiv preprint arXiv:2110.05809*, 2021.

[14] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning convolution system for dcase 2019 task 4," in *Workshop on Detection and Classification of Acoustic Scenes and Events 2019*, 2019, p. 134.

[15] https://github.com/Anaesthesiaye/sound_event_detection_transformer.

[16] G. Tian, Y. Huang, Z. Ye, S. Ma, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, K. Ouchi, and R. Ebbers, Janek Haeb-Umbach, "Sound event detection using metric learning and focal loss for dcase 2021 task 4," DCASE2021 Challenge, Tech. Rep., June 2021.