Improving Sound Event Detection with Auxiliary Foreground-Background Classification and Domain Adaptation

Michel Olvera, Emmanuel Vincent, Gilles Gasso

IMPROVING SOUND EVENT DETECTION WITH AUXILIARY FOREGROUND-BACKGROUND CLASSIFICATION AND DOMAIN ADAPTATION

Michel Olvera[1]*, Emmanuel Vincent*[1]*, Gilles Gasso*[2]

[1] Université de Lorraine, CNRS, Inria, Loria, F-54000 Nancy, France
[2] LITIS EA 4108, Université & INSA Rouen Normandie, 76800 Saint-Étienne du Rouvray, France
michel.olvera@inria.fr

ABSTRACT

In this paper we provide two methods that improve the detection of sound events in domestic environments. First, motivated by the broad categorization of domestic sounds as foreground or background events according to their spectro-temporal structure, we propose to learn a foreground-background classifier jointly with the sound event classifier in a multi-task fashion to improve the generalization of the latter. Second, while the semi-supervised learning capability adopted for training sound event detection systems with synthetic labeled data and unlabeled or partially labeled real data aims to learn invariant representations for both domains, there is still a gap in performance when testing such systems on real environments. To further reduce this data mismatch, we propose a domain adaptation strategy that aligns the empirical distributions of the feature representations of active and inactive frames of synthetic and real recordings via optimal transport. We show that these two approaches lead to enhanced detection performance in terms of the event-based macro F1-score on the DESED dataset.

Index Terms— Sound event detection, foreground-background classification, semi-supervised learning, domain adaptation

1. INTRODUCTION

Over the past five years, the interest in environmental acoustic scenes has increased considerably among researchers in the field of audio signal processing, largely driven by the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge and Workshop series [1–3]. As a result, many tasks involving the automatic analysis of ambient sounds have progressed substantially. This includes sound event detection (SED), a core task for home surveillance or assisted living [4–6].

To this end, DCASE Challenge Task 4 encourages the development of methods that contribute to the advancement of SED methods that are trained in a semi-supervised way with a heterogeneous dataset [7] including a set of synthetic soundscapes with annotations indicating the sound events class labels and timestamps (strong labels), as well as a set of real recordings, mostly unlabeled and where only a small subset contains at most information about the active sound classes in the recordings (weak labels). The objective of the task is to find for a given soundscape the class of the active sounds as well as their onset and offset time. Many improvements to the Task 4 baseline system have been proposed: augmentation schemes to improve generalization [8, 9]; changes in the acoustic front-end with alternative time-frequency representations to log-Mel spectrograms [10] or time-frequency resolutions for each sound event class [11]; modifications to the backbone architecture with appended multi-task branches and post-processing techniques to refine the outputs [12–14].

In this work we propose two methods to improve SED in domestic environments. First, we propose the classification of sounds by their spectro-temporal content as foreground or background events as an auxiliary task for SED. This broad categorization of sound events was shown to be useful for deep neural network-based source separation systems to differentiate rapidly varying spectro-temporal features of short duration sound events from slowly varying features of long duration sounds [15, 16]. The proposed foreground-background classifier is jointly trained with the SED branch in a multi-task fashion and the combination of both branches is also investigated. The second improvement is a domain adaptation strategy to reduce the mismatch between synthetic and real recordings. Our proposed strategy aligns the empirical distributions of the feature representations of active and inactive frames of synthetic and real data via optimal transport [17, 18]. Altogether the proposed methods lead to enhanced performance in terms of the event-based macro F1-score on the Domestic Environment Sound Event Detection Dataset (DESED) validation and public evaluation sets [19, 20].

The remainder of this article is organized as follows. In Section 2 we present the proposed improvements to the SED task. Experimental evaluation is discussed in Section 3 and lastly, we conclude the paper in Section 4.

2. PROPOSED METHODS

2.1. Foreground-background classification

Let $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ be the input, output and latent space. For the SED task we denote the soundscape time-frequency representation by $x \in \mathcal{X}$ with corresponding annotations $y \in \mathcal{Y}$. We have access to a synthetic dataset with strong labels $\mathcal{D}^{\mathcal{S}} = \{(x_i^{\mathrm{s}}, y_i^{\mathrm{s}})\}_{i=1}^{n_{\mathrm{s}}}$ and two datasets of real recordings: a weakly labeled dataset $\mathcal{D}^{\mathcal{W}} = \{(x_i^{\mathrm{w}}, y_i^{\mathrm{w}})\}_{i=1}^{n_{\mathrm{w}}}$ and an unlabeled dataset $\mathcal{D}^{\mathcal{U}} = \{x_i^{\mathrm{u}}\}_{i=1}^{n_{\mathrm{u}}}$.

The SED model is a Mean Teacher model [21] in which both the student and the teacher models have the same convolutional-recurrent neural network (CRNN) architecture. We use the CRNN from the student model as a representation mapping $g : \mathcal{X} \to \mathcal{Z}$, where the log-Mel spectrograms are mapped to the latent space. The

This work was made with the support of the French National Research Agency, in the framework of the project LEAUDS "Learning to understand audio scenes" (ANR-18-CE23-0020). Experiments presented in this paper were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (https://www.grid5000.fr).

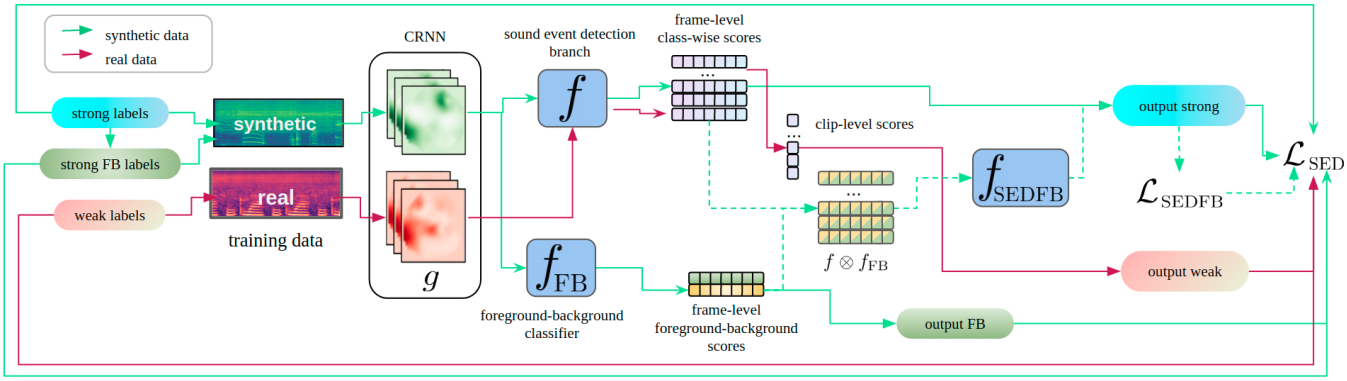

Figure 1: Proposed model with colored data flow for synthetic and real data. For simplicity, the diagram depicts only the student model and the associated classification costs. The dash-lined path represents the training scheme with the fusion of the sound event detection and foreground-background classification branches.

SED model is represented by the function $f : \mathcal{Z} \rightarrow \mathcal{Y}$ that maps the latent representations to the output space.

2.1.1. FB branch

Motivated by the broad categorization of sound events into foreground and background according to their spectro-temporal structure, we propose a foreground-background (FB) auxiliary classifier $f_{\text{FB}} : \mathcal{Z} \rightarrow \mathcal{Y}^{\text{FB}}$ that maps the latent space to foreground-background labels. We learn this classifier jointly with the SED model in a multi-task fashion, hypothesizing that these different yet related classification tasks will help improve the network's generalization capability. Analogously, for the teacher model we denote by g', f' and f'_{FB}, the CRNN embedding function, SED and FB branches, respectively. Figure 1 shows the proposed system depicting the student model.

To train the FB classifier in the multi-task paradigm, we derived foreground-background ground-truth annotations y_i^{fb} from the strong labels y_i^{s} of the synthetic data by combining the sound event labels in two categories: foreground: (*alarm - bell ringing, speech, cat, dog*, *dishes*) and background (*blender, vacuum cleaner, frying, electric shaver - toothbrush*, *running water*). The SED model is optimized by minimizing

$$\begin{aligned}\mathcal{L}_{\text{SED}} = L(y_i^{\text{s}}, f(g(x_i^{\text{s}}))) + \lambda L_{\text{strong}}(f(g(x_i)), f'(g'(x_i)))+ \\ L(y_i^{\text{w}}, f(g(x_i^{\text{w}}))) + \lambda L_{\text{weak}}(f(g(x_i)), f'(g'(x_i)))+ \\ L(y_i^{\text{fb}}, f_{\text{FB}}(g(x_i^{\text{s}}))) + \lambda L_{\text{strong}}(f_{\text{FB}}(g(x_i)), f'_{\text{FB}}(g'(x_i)))\end{aligned} \quad (1)$$

where $L(\cdot,\cdot)$ is a binary cross-entropy classification loss, and $L_{\text{strong}}(\cdot,\cdot)$ and $L_{\text{weak}}(\cdot,\cdot)$ are mean-square error consistency costs which are differentiable on their second parameter over strong (frame-level) and weak (clip-level) scores, respectively. The consistency weight λ is tied to all consistency costs.

2.1.2. SEDFB branch

Going beyond the proposed FB classification branch, we explored its fusion with the SED branch into a detection branch (SEDFB) to refine outputs. This branch is represented by a function $f_{\text{SEDFB}} : \mathcal{Y} \times \mathcal{Y}^{\text{FB}} \rightarrow \mathcal{Y}$ (f'_{SEDFB} for the teacher model). The input for the SEDFB branch is the outer product of the outputs from the SED and FB branches $w_i = f(g(x_i)) \otimes f_{\text{FB}}(g(x_i))$, as this fusion creates a representation containing information from the joint interaction of the SED and FB classifiers. The following classification-consistency cost pair is added to the training objective in (1):

$$\mathcal{L}_{\text{SEDFB}} = L(y_i^{\text{s}}, f_{\text{SEDFB}}(w_i^{\text{s}})) + \lambda L_{\text{strong}}(f_{\text{SEDFB}}(w_i), f'_{\text{SEDFB}}(w_i)). \quad (2)$$

The overall cost involving the SEDFB branch is given by

$$\mathcal{L} = \mathcal{L}_{\text{SED}} + \mathcal{L}_{\text{SEDFB}}. \quad (3)$$

2.1.3. Output smoothing

We used two methods to post-process the SED frame-level scores. The first method corresponds to smoothing the binary multi-label frame-level scores with a median filter of 0.45 s. The second approach consists of Hidden Markov Model (HMM) decoding. Following the same procedure as in [14], we determined the optimal transition probabilities for each sound event class using the validation set. We contrast the contribution of both post-processing schemes in Section 3.3.

2.2. Domain adaptation for sound event detection

From the unsupervised domain adaptation perspective, we regard the synthetic dataset with strong labels as the source domain $\mathcal{S} = \mathcal{D}^{\mathcal{S}}$, and the combination of real recordings from the weakly and unlabeled dataset as the target domain $\mathcal{T} = \mathcal{D}^{\mathcal{W}} \cup \mathcal{D}^{\mathcal{U}}$. We denote as x^{s} and x^{t} the soundscapes from $\mathcal{S}$ and $\mathcal{T}$, respectively.

In contrast to adversarial adaptation approaches that introduce a domain discriminator to reduce the distribution discrepancy between domains [14, 22, 23], our proposed strategy relies on optimal transport for its ability to find correspondences between samples by exploiting the geometry of the underlying space. We adopt the DeepJDOT framework [18], to correct the mismatch between the distributions of learned feature representations in the two domains.

2.2.1. Joint distribution optimal transport

Let $\mu_{\text{s}} = \sum_{i=1}^{n_{\text{s}}} a_i \delta_{g(x_i^{\text{s}}), y_i^{\text{s}}}$ and $\mu_{\text{t}} = \sum_{i=1}^{n_{\text{t}}} b_i \delta_{g(x_i^{\text{t}}), y_i^{\text{t}}}$ be two empirical distributions on the product space $\mathcal{Z} \times \mathcal{Y}$, where $\delta_{g(x_i), y_i}$

is the Dirac function at position $(g(x_i), y_i) \in \mathcal{Z} \times \mathcal{Y}$, and a_i and b_i are uniform probability weights, i.e. $\sum_{i=1}^{n_s} a_i = \sum_{i=1}^{n_t} b_i = 1$. The associated cost for the i-th source and j-th target element can be expressed as a weighted combination of costs in the latent and label spaces

$$d(g(x_i^s), y_i^s; g(x_j^t), y_j^t) = \alpha c(g(x_i^s), g(x_j^t)) + \beta \mathcal{L}(y_i^s, y_j^t) \quad (4)$$

where $c(\cdot,\cdot)$ is the squared ℓ_2 distance, $\mathcal{L}(\cdot,\cdot)$ is a cross-entropy loss that enforces regularity between the source and target domain labels, and α and β are two scalar values. Since no labels y_j^t are available in the target domain they are replaced with pseudo-labels $f(g(x_j^t))$ obtained from the classifier $f : \mathcal{Z} \to \mathcal{Y}$. We seek for a transportation coupling $\gamma \in \mathbb{R}^{n_s \times n_t}$ in the space $\Gamma(\mu_s, \mu_t)$ of joint probability distributions with marginals $\gamma \mathbf{1}_{n_t} = \mu_s$ and $\gamma^T \mathbf{1}_{n_s} = \mu_t$, where $\mathbf{1}_d$ is a d-dimensional vector of ones, and a pair of mapping functions g and f that minimize

$$\min_{\gamma \in \Gamma(\mu_s, \mu_t), g, f} \sum_{i,j} \gamma_{i,j} d(g(x_i^s), y_i^s; g(x_j^t), f(g(x_j^t))). \quad (5)$$

We follow a two-step procedure to solve this optimization problem. In the first step, we compute the optimal coupling matrix γ with fixed model parameters f and g,

$$\min_{\gamma \in \Gamma(\mu_s, \mu_t)} \sum_{i,j} \gamma_{ij} (\alpha ||g(x_i^s) - g(x_j^t)||^2 + \beta \mathcal{L}(y_i^s, f(g(x_j^t)))). \quad (6)$$

In the second step, with fixed γ, we update the models g and f as

$$\min_{g,f} \mathcal{L}_s + \sum_{i,j} \gamma_{ij} (\alpha ||g(x_i^s) - g(x_j^t)||^2 + \beta \mathcal{L}(y_i^s, f(g(x_j^t)))) \quad (7)$$

where $\mathcal{L}_s$ correspond to the classification cost on the source domain to avoid losing performance on synthetic data.

2.2.2. *Sampling strategy*

For each data batch we sample all active and inactive frames from the source and target domains as indicated by the strong labels and the pseudo-labels. For both domains we only keep active frames where no sound event overlap occurs, so that the optimal transport takes place between the empirical distributions of the sound classes of both domains. The number of sampled active frames per class can vary considerably from batch to batch for synthetic and real data, which can lead to the absence of certain classes in one type of data for some batches. To account for this imbalance problem we only keep active frames from common classes to both domains, and then balance all classes by resampling them randomly, taking as many elements as there are in the class with fewer elements. Similarly, the number of inactive frames in the source and target domains varies from batch to batch, so we sample randomly each set by taking as many inactive frames as there are in the set with fewer elements.

2.2.3. *Pseudo-label refinement*

To improve the reliability of the pseudo-labels assigned to real data, we leverage the provided annotations of the weakly labeled set to refine pseudo-labels on this subset. The refinement process consists of fusing the frame-level outputs of the SED branch f on soundscapes from $\mathcal{D}^{\mathcal{W}}$ with their clip-level annotations by an element-wise multiplication. The target domain pseudo-labels are thus updated for

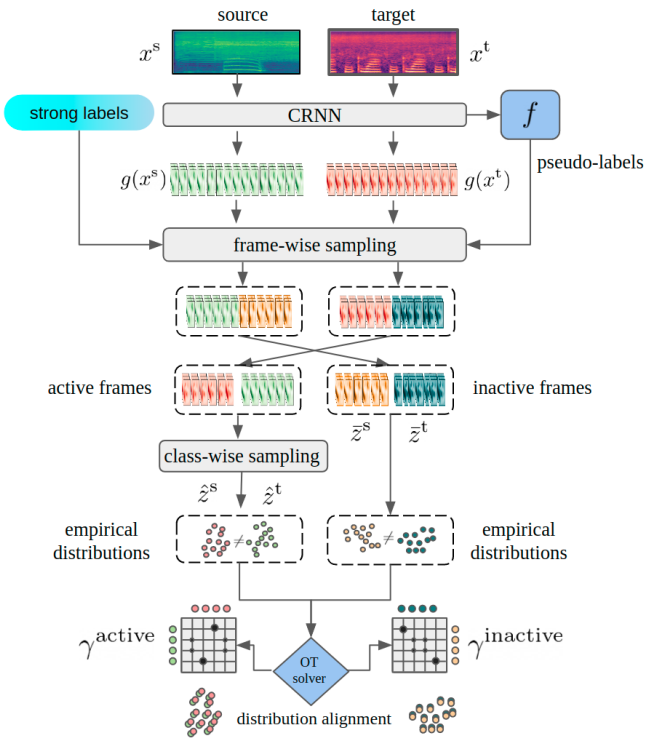

Figure 2: Proposed adaptation method to correct domain mismatch.

weakly labeled data as $\hat{y}_j^t = f(g(x_j^w)) \odot y_j^w$, $j = 1, \ldots, n_w$. This operation constrains the estimated labels to contain at most the same classes present in the weakly labeled soundscapes. Filtering out all extra classes helps reduce false positives and allows more reliable pseudo-labels to be obtained for the proposed sampling strategy and domain adaptation process.

2.2.4. *Training objectives*

We denote as $\hat{z}^s$ and $\hat{z}^t$ the sampled active frames and as $\bar{z}^s$ and $\bar{z}^t$ the sampled inactive frames from the source and target domain latent representations z^s and z^t, respectively. After an initial pre-training stage using (3), we construct the following objective function to account for the mismatch between the empirical distributions of active and inactive learned feature representations

$$\mathcal{L}_s + \mathcal{L}_{\text{active}} + \mathcal{L}_{\text{inactive}} \quad (8)$$

where

$$\mathcal{L}_s = \frac{1}{n_s} \sum_{i=1}^{n_s} L(y_i^s, f(g(x_i^s))) + \frac{1}{n_s} \sum_{i=1}^{n_s} L(y_i^{fb}, f_{FB}(g(x_i^s))) \quad (9)$$

corresponds to the first and third classification cost terms of the training classification cost in (1). As only the student model undergoes adaptation, no consistency losses are included in the above objective to train the source domain classifier. The cost function $\mathcal{L}_a$ corresponds to the distribution alignment loss of active frames

$$\mathcal{L}_{\text{active}} = \frac{1}{|C_{\text{active}}|} \sum_{i,j}^{N_{\text{active}}} \gamma_{ij}^{\text{active}} (\alpha ||\hat{z}_i^s - \hat{z}_i^t||^2 + \beta \mathcal{L}(y_i^s, \hat{y}_j^t)) \quad (10)$$

where $|C_{\text{active}}|$ is the cardinality of the subset of labels $C_{\text{active}} \in C$ representing the total number of active classes in the batch. The second term in (10) enforces consistency between the target domain pseudo-labels and source domain labels. The cost function $\mathcal{L}_{\text{inactive}}$ accounts for the alignment of the marginal distributions of the learned representations of inactive frames in both domains:

$$\mathcal{L}_{\text{inactive}} = \sum_{i=1}^{N_{\text{inactive}}} \gamma_{ij}^{\text{inactive}} (\alpha ||\vec{z}_i^{\text{s}} - \vec{z}_i^{\text{t}}||^2). \quad (11)$$

Figure 2 depicts the proposed frame-level domain adaptation strategy based on optimal transport for the SED task.

3. EXPERIMENTS AND RESULTS

3.1. Model

The selected model architecture is the same as the baseline system of DCASE 2020. The CNN part is composed of 7 layers with 16, 32, 64, 128, 28, 128, 128 filters, respectively. A kernel of size 3x3 was used with max-pooling [2, 2], [2, 2], [1, 2], [1, 2], [1, 2], [1, 2] and [1, 2], respectively. A gated linear unit activation is applied to the convolution operations. The RNN part is composed of 2 layers of 128 bidirectional gated recurrent units. The output of the CRNN is followed by a dense layer with sigmoid activation to produce frame-level (strong) class-wise posteriors. Clip-level (weak) scores are obtained by multiplying the aforementioned linear layer with a dense layer with softmax activation followed by mean temporal aggregation. The FB branch consists of a dense layer with sigmoid activation, which acts upon the outputs of the RNN block. The SEDFB branch is composed of a bidirectional RNN with 128 gated recurrent units and a dense layer with sigmoid activation.

3.2. Dataset and training procedure

We conducted experiments on the Domestic Environment Sound Event Detection Dataset (DESED) dataset [19, 20], composed of a training set of 2,584 synthetic audio clips generated by Scaper [24], 1,578 real soundscapes with clip-level annotations and 14,412 unlabeled real recordings. In the training stage, the model was trained for 200 epochs with the Adam optimizer, a dropout value of 0.5, and a gradually increasing learning rate with a max value of 10^{-3} [25]. The consistency weight λ was set to 1. In the adaptation stage, the student model was adapted for 300 epochs. We used cost weights $\alpha = 0.2$, $\beta = 5.0$, and the contribution of the source domain classifier cost $\mathcal{L}_s$ to the total adaptation cost was multiplied by 100. The learning rate was fixed to 10^{-4}. Experiments with optimal transport were performed using the Python Optimal Transport package [26].

3.3. Results

In Table 1 we compare results obtained by the proposed methods on the validation and public evaluation sets of the DESED dataset in terms of the event-based macro F1 score. The model labeled as Baseline correspond to the baseline system of DCASE 2020 Challenge Task 4. Although the baseline system of the 2021 edition comprises the same architecture as Baseline, it cannot be compared with our methods as it was trained on a different synthetic dataset with data augmentation. For each evaluation set we show performance with median filtering (+MF) or HMM smoothing (+HMM) post-processing.

Adding the FB branch is beneficial to the SED task as it improved results on the validation set compared to Baseline by 8.3%.

Table 1: Performance on the validation and public evaluation sets.

Method	F1 score		F1 score	
	val +MF	val +HMMs	eval +MF	eval +HMMs
Baseline	34.8		38.1	
+ DA	42.41	43.89	44.8	47.12
+ FB	43.12	45.42	46.06	49.38
+ FB + DA	45.68	47.77	50.79	53.10
+ SEDFB	46.15	46.20	48.40	49.79
+ SEDFB + DA	**47.61**	47.75	**52.12**	53.30
DCASE 1	45.13	**48.07**	50.58	**53.35**
DCASE 2	45.15	47.08	50.28	52.23

Further enhancement was achieved by refining outputs with HMM smoothing, as performance increased by 10.6%. Moreover, its fusion with the SED branch into a combined SEDFB branch brought an additional gain of 3% with median filtering. A similar trend is observed for the public evaluation set.

Model adaptation with the proposed strategy increased performance over Baseline by 7.6% as a standalone method, and by 10.8% and 12.8% when combined with the FB and SEDFB branches. These results prove the effectiveness of the system in reducing mismatch between synthetic and real data. Compared to the validation set, performance was about 2% larger on the public set, which might be due to the fact that the empirical distribution of the active and inactive frames of this set resembles more that of the provided real training data on which adaptation was carried out.

HMM smoothing as a post-processing method yielded greater improvement to the scores over median filtering for all proposed models except for the SEDFB-based methods, in which +MF and +HMM provided similar scores, implying that the SEDFB branch plays a similar role as HMM decoding in the modeling of time-varying spectra of sound events.

DCASE 1 and DCASE 2 are model ensembles comprising three and two + FB + DA systems from different training runs, respectively. Ensembling is achieved by simply averaging the model outputs. These models correspond to the submissions made to the DCASE 2021 Challenge Task 4 and are labeled as *Olvera_INRIA_task4_SED_1* and *Olvera_INRIA_task4_SED_2*, respectively. Both systems showed competitive performance in terms of the event-based macro F1-score on the evaluation and public evaluation sets among 65 systems.

4. CONCLUSION

In this paper we proposed two methods that enhance the detection of domestic sound events. Motivated by the categorization of the spectro-temporal characteristics of domestic sounds as foreground or background, we proposed the use of an auxiliary foreground-background classifier that is jointly trained with the sound event classifier to improve generalization. Furthermore, we proposed to incorporate an adaptation stage based on the joint distribution optimal transport of feature embeddings and labels to account for the acoustic mismatch between the available synthetic and real training data. We showed that the multi-task training scheme together with the adaptation stage substantially improved the performance of the baseline system.

5. REFERENCES

[1] M. D. Plumbley, C. Kroos, J. P. Bello, G. Richard, D. P. Ellis, and A. Mesaros, Eds., *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. Tampere University of Technology. Laboratory of Signal Processing, 2018.

[2] M. Mandel, J. Salamon, and D. P. W. Ellis, Eds., *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*. New York University, 2019.

[3] N. Ono, N. Harada, Y. Kawaguchi, A. Mesaros, K. Imoto, Y. Koizumi, and T. Komatsu, Eds., *Proceedings of the Fifth Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2020)*, 2020.

[4] P. van Hengel and J. Anemüller, "Audio event detection for in-home care," in *Int. Conf. on Acoustics (NAG/DAGA)*, 2009.

[5] R. M. Alsina-Pagès, J. Navarro, F. Alías, and M. Hervás, "homesound: Real-time audio event detection based on high performance computing for behaviour and surveillance remote monitoring," *Sensors*, vol. 17, no. 4, p. 854, 2017.

[6] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[7] N. Turpault and R. Serizel, "Training sound event detection on a heterogeneous dataset," in *Proc. DCASE*, 2020, pp. 200–204.

[8] X. Li, "Semi-supervised sound event detection using random augmentation and consistency regularization," *arXiv preprint arXiv:2102.00154*, 2021.

[9] K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, "Conformer-based sound event detection with semi-supervised learning and data augmentation," in *Proc. DCASE*, 2020, pp. 100–104.

[10] A. Copiaco, C. Ritz, N. Abdulaziz, and S. Fasciani, "A study of features and deep neural network architectures and hyperparameters for domestic audio classification," *Applied Sciences*, vol. 11, no. 11, p. 4880, 2021.

[11] D. de Benito-Gorron, D. Ramos, and D. T. Toledano, "A multi-resolution approach to sound event detection in dcase 2020 task4," in *Proc. DCASE*, 2020, pp. 36–40.

[12] Y. Huang, L. Lin, S. Ma, X. Wang, H. Liu, Y. Qian, M. Liu, and K. Ouchi, "Guided multi-branch learning systems for sound event detection with sound separation," in *Proc. DCASE*, 2020, pp. 61–65.

[13] L. Cances, T. Pellegrini, and P. Guyot, "Multi task learning and post processing optimization for sound event detection," IRIT, Université de Toulouse, CNRS, Toulouse, France, Tech. Rep., 2019.

[14] S. Cornell, M. Olvera, M. Pariente, G. Pepe, E. Principi, L. Gabrielli, and S. Squartini, "Domain-adversarial training and trainable parallel front-end for the dcase 2020 task 4 sound event detection challenge," in *Proc. DCASE*, 2020, pp. 26–30.

[15] M. Olvera, E. Vincent, R. Serizel, and G. Gasso, "Foreground-background ambient sound scene separation," in *Proc. EUSIPCO*, 2020, pp. 281–285.

[16] D. D. Varma, R. Padmanabhan, and A. Dileep, "Learning to separate: Soundscape classification using foreground and background," in *Proc. EUSIPCO*, 2020, pp. 21–25.

[17] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy, "Joint distribution optimal transportation for domain adaptation," in *Proc. NIPS*, 2017, pp. 3733–3742.

[18] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty, "Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation," in *Proc. ECCV*, vol. 11208, 2018, pp. 467–483.

[19] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. DCASE*, 2019, pp. 253–257.

[20] R. Serizel, N. Turpault, A. Shah, and J. Salamon, "Sound event detection in synthetic domestic environments," in *Proc. ICASSP*, 2020, pp. 86–90.

[21] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017.

[22] L. Yang, J. Hao, Z. Hou, and W. Peng, "Two-stage domain adaptation for sound event detection," in *Proc. DCASE*, 2020, pp. 230–234.

[23] H. Park, S. Yun, J. Eum, J. Cho, and K. Hwang, "Weakly labeled sound event detection using tri-training and adversarial learning," in *Proc. DCASE*, 2019, pp. 184–188.

[24] J. Salamon, D. MacConnell, M. Cartwright, P. Li, and J. P. Bello, "Scaper: A library for soundscape synthesis and augmentation," in *Proc. WASPAA*, 2017, pp. 344–348.

[25] L. Delphin-Poulat and C. Plapous, "Mean teacher with data augmentation for dcase 2019 task 4," Orange Labs Lannion, France, Tech. Rep., 2019.

[26] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, "Pot: Python optimal transport," *Journal of Machine Learning Research*, vol. 22, no. 78, pp. 1–8, 2021.