Cosmopolite Sound Monitoring (CoSMo): A Study of Urban Sound Event Detection Systems Generalizing to Multiple Cities

Florian Angulo, Slim Essid, Geoffroy Peeters, Christophe Mietlicki

▶ **To cite this version:**

Florian Angulo, Slim Essid, Geoffroy Peeters, Christophe Mietlicki. Cosmopolite Sound Monitoring (CoSMo): A Study of Urban Sound Event Detection Systems Generalizing to Multiple Cities. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun 2023, Rhodes Island, Greece. pp.1-5, 10.1109/ICASSP49357.2023.10095833 . hal-04093374

HAL Id: hal-04093374
https://hal.science/hal-04093374

Submitted on 10 May 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

COSMOPOLITE SOUND MONITORING (COSMO) : A STUDY OF URBAN SOUND EVENT DETECTION SYSTEMS GENERALIZING TO MULTIPLE CITIES

Florian Angulo [⋆†] *Slim Essid* [⋆] *Geoffroy Peeters* [⋆] *Christophe Mietlicki* [†]

[⋆] LTCI - Télécom Paris, Institut Polytechnique de Paris
[†] Bruitparif

ABSTRACT

Measuring noise in cities and automatically identifying the corresponding sound sources are a crucial challenge for policymakers. Indeed, such information helps addressing noise pollution and improving the well-being of urban dwellers. In recent years, researchers have provided annotated datasets recorded in two major cities to foster the development of urban sound event detection (SED) systems. This paper presents an in-depth study of the behaviour of state-of-the-art SED systems well suited to our problem, combining three far-field real recordings datasets which can be used jointly during training. In our evaluation, we highlight the performance gaps existing between simple and hard recording examples based on the salience of sound events and the polyphony of the recordings. We provide new proximity annotations for this analysis. We evaluate the ability of urban SED systems to generalize across cities with varying degrees of training supervision. We show that such generalization is hindered mostly by the difficulties current urban SED systems have to detect sound events with low salience along with sound events in highly polyphonic soundscapes.

***Index Terms*—** Sound Event Detection (SED), Far-field urban audio recordings, urban sound monitoring,

1. INTRODUCTION

Noise pollution in big cities is one of the most challenging issues to address by urban policymakers to improve the quality of life of urban citizens. Data-driven approaches have recently been investigated [1, 2, 3, 4] to provide automated reports and refine simulation-based noise maps. The task of estimating the presence of sound classes of interest in an audio recording is called sound event detection (SED). The Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge Task 4 on synthetic and real domestic sound recordings [5] has provided in the recent years a framework to develop systems that perform well both on clip-level SED and frame-level SED. In the context of urban SED, The first studies focused on synthetic urban soundscapes or monophonic recordings [6]. However, there is no guarantee that systems trained on such data can generalize well if deployed in a real world scenario. Therefore, great efforts have been made to provide annotated real polyphonic recordings of urban sounds [1, 2], which is essential to train deep neural networks, the current state-of-the-art approach [7].

The ultimate goal of urban SED is to be universal and reliably applicable to any city. Due to the financial cost and human labor required, curating annotated real recordings of urban soundscapes for every city in the world is impractical. Nonetheless, state-of-the-art urban sound event detectors need annotated data to be trained. Therefore, it is crucial to investigate how systems trained on recordings from one or a few cities can generalize to other cities. In the context of audio-visual understanding of urban traffic scenes, the Urbansas dataset [8] was recently proposed to benchmark audio-visual vehicle tracking systems across different cities.

In the context of general-purpose urban sound monitoring, even though an ontology has been proposed with the Audioset initiative [9], which is exploited in many derived audio classification tasks, cross-evaluation between different urban sound monitoring datasets remains hard. Indeed, it is necessary that a common set of classes be large enough and shared across datasets to perform a fair evaluation, which is often not the case. Sounds of New York City (SONYC) [1] is the first dataset to propose a hierarchical taxonomy targeting those sound classes that are the main sources of urban noise pollution. More recently, SINGA:PURA [2], a dataset of real urban audio recordings from Singapore with strongly-annotated (frame-level) data, has been proposed. The authors chose to follow the same taxonomy as that of the SONYC dataset, with a proposed extension incorporating more urban sound classes, which creates an opportunity to test urban SED systems' generalization across different cities.

Though the DCASE challenge 2020 Task 5[1] has favoured the development of advanced urban SED systems, it only considered the SONYC data. The challenge ranked submitted systems on global and per-class metrics obtained on the evaluation set of SONYC and showed that further improvements could still be reached, especially on class-wise performances. This has further motivated us to seek a better explanation of when urban SED systems under-perform the most. We thus believe that it is still particularly relevant to study in this context the impact of polyphony and proximity of sound events occurring in the urban soundscapes, as well as systems' generalization to cities unseen during training.

Contributions We offer an in-depth study of the behaviour of urban SED systems and test their generalization abilities across cities, combining data from SONYC and SINGA:PURA. Following the DCASE challenge task 4 SED system's design, we experiment with systems performing SED on far-field and non-synthetic urban sound recordings from two different cities. To the best of our knowledge, this framework is, in particular, the first to exploit proximity annotations to evaluate urban sound classifiers. We also investigate how our systems behave under various levels of polyphony. We distribute our code as well as the proximity annotations of the SONYC evaluation set which we created for this study.[2]

This work was funded by Bruitparif and granted access to the HPC resources of IDRIS under the allocation 2022-AD011013394 made by GENCI.

[1] https://dcase.community/challenge2020/task-urban-sound-tagging-with-spatiotemporal-context

[2] https://github.com/florian-angulo/CoSMo/

2. THE COSMO FRAMEWORK

While previous works on sound event detection in real urban soundscapes only focused on a single dataset, we propose CoSMo: a cross-dataset experimental framework. It is currently composed of three existing datasets:

1) SONYC-UST-v2 [1]: 51.4 hours of weakly-annotated audio data recorded in New York with 56 single-microphones.

2) SINGA:PURA labelled set [2]: 18.2 hours of strongly-annotated audio data recorded in Singapore with 10 single-microphones and 4 seven-microphone arrays.

3) SINGA:PURA unlabelled set [2]: 201 hours of unlabelled audio data recorded in the same conditions as SINGA:PURA labelled.

The label taxonomy used is a parameter of our framework. One can choose to use the original SONYC-UST taxonomy (8 coarse classes, 23 fine classes) or one can use the extension proposed by SINGA:PURA (14 coarse classes, 40 fine classes).[3]

Dataset Splits. For SONYC-UST, we use the split into training, validation and evaluation sets as proposed by the authors [1]. For SINGA:PURA [2], the authors do not propose such a split. We therefore replicate the strategy used for splitting SONYC-UST to SINGA:PURA: a) sensors used in the validation set must not be used in the training set b) recordings of the evaluation set must not occur on the same day as any recordings from the training set or validation set. c) The class distributions of the three subsets must be similar. The best split we found checking all these requirements is the following: the validation set is comprised of events recorded after August 21th with sensors [b827eb0ebf2f, b827eb7680c5 and b827eb3e52b8], the evaluation set is comprised of events recorded before and including August 21th and the training set is comprised of events recorded after August 21th without sensors [b827eb0ebf2f, b827eb7680c5 and b827eb3e52b8]. With this dataset split, we replicate the usual train, validation and evaluation ratio of 70%, %10, 20%, respectively.

"Proximity" annotations. For both SONYC and SINGA:PURA, annotated sound event labels come with additional information on the proximity of the events (near or far). For SINGA-PURA, these "proximity" annotations are complete and verified by a scientific team [2]. For SONYC, these annotations are provided by volunteers and the major part of the event labels are not associated with a "proximity" label. Also for the provided ones, a significant disagreement exists between annotators. Thus, we leave the exploitation of these noisy proximity annotations during training for future work. Still, since for our analysis, we need consistent proximity annotations, we re-annotated the "proximity" of labeled sound events of the evaluation set of SONYC . We used the following annotation rule: if the sound event is salient or easily distinguishable from the ambient noise and simultaneous active sound sources, we label it as "near". Otherwise, we label it as "far". Clearly, this had better be referred to as "salience" annotation, but we choose for simplicity not to change the way it is referred to, especially as, in practice, on this dataset "salience" and "proximity" are well correlated. We discard from our annotations any ambiguous scenarios where the same sound class occurs "near" and "far" at different times in the clip. As a consequence, the union of the "near" and "far" subsets is not equal to the whole set.

[3] https://zenodo.org/record/5645825

3. CONSIDERED SYSTEMS

For our experiments we consider three classification systems, based on our assessment of the current state of the art in the DCASE field [6, 10]. While we could have focused only on systems used in the DCASE Challenge on the SONYC dataset, our choices are also motivated by the existence of strongly annotated data and large amounts of unlabeled data from SINGA:PURA, which calls for systems making both frame-level and clip-level predictions and also calls for semi-supervised learning paradigms to exploit unlabelled data. The first system corresponds to the baseline of the DCASE Challenge Task 4 [5], a convolutional recurrent neural network (CRNN) using the Mean-Teacher semi-supervised learning paradigm [11]. We consider the convolutional part as the *feature extraction* pipeline and the recurrent part (a two-layer bidirectional Gated Recurrent Unit followed by a Linear layer with attention pooling to aggregate frame-level predictions into a clip-level prediction) as the *classification head*. While more advanced systems were proposed on sound event detection challenges such as transformer-based or conformer-based architectures [12], we prefer to use this well-established and well-studied architecture [6].

The second and third systems use pretrained embeddings for the feature extraction stage but the same classification head. These embeddings are computed with the best performing systems submitted in the HEAR challenge [10] on environmental sound event detection downstream tasks (ESC-50, DCASE 2016 Task 2, FSD50K) [10]. The second system therefore uses PaSST (Patchout faSt Spectrogram Transformer) [13], a vision transformer trained on Audioset in a supervised way, as its feature extractor. The third system uses openL3 [14], a CNN trained on Audioset in an unsupervised way with an audio-visual correspondence task.

For the rest of the study we name the first system "CRNN" , the second system "embedding (PaSST) classifier" and the third system"embedding (openL3) classifier".

4. EXPERIMENTS

We design our experiments to answer the following questions: how well can the chosen systems generalize across cities? What is the impact of proximity on performances? What is the impact of polyphony on performances? Does a state-of-the-art pretrained feature extractor perform as well as one trained-from-scratch?

4.1. Training of the CRNN

We take inspiration from the framework of the baseline system used in DCASE task 4 2021 [5]. We compute the Mel-spectrogram from the 32 kHz audio with a Hamming window of 60 ms, a hop size of 16 ms and 128 Mel filters.

Frontends. We test three different processings of the input Mel-spectrogram. The first one is a simple pointwise log-transformation. The second one is PCEN [15] with the hyperparameters advised by an in-depth study of this function [16]: $\alpha = 0.8$, $r = 0.25$, $\epsilon = 10^{-6}$, $\delta = 10$ and $T = 800$ ms. The authors propose a window size of $T = 60$ ms for bird sound event detection. We select empirically a longer window because chirp rates of urban sound events are generally slower. Even though, there does not exist an ideal window size which can enhance at best every sound classes of interest, as shown in [17], we found empirically $T = 800$ ms to be a good trade-off. To prevent losing information caused by PCEN de-emphasizing background noise and stationary sounds, we propose a third processing which combines the log-transform and PCEN into a

2-channel input. We then apply the same normalization scheme as in [18]: we standardize each Mel band of each channel independently using their mean and standard deviation computed without outliers.

Batch content. In a training batch, we can include weakly-labelled examples from SONYC, strongly-labelled ones from SINGA:PURA, weakly-labelled (derived from the strong-labels : "a class is active at the clip-level if it is active during at least one frame of the clip) ones from SINGA:PURA and unlabelled examples of SINGA:PURA. In our experiments, unlabelled examples only contribute to the regularization provided by the Mean-Teacher [11] paradigm. As we can combine datasets of different sizes in the training batch, we ensure that no matter the training batch configuration, the number of annotated examples used per epoch is the same. We set the batch size of annotated examples to 32. If we use unlabelled examples from SINGA:PURA in the training batch, we take 16 examples per batch.

Training loss components. If examples from both annotated datasets are present in a training batch, we give them equal weight in the supervised loss. If we supervise frame-level predictions on SINGA:PURA, we give respectively a weight of 60% and 40% to the losses computed respectively on frame-level predictions loss and on clip-level predictions. If we use the Mean-Teacher [11] regularization, the self-supervised loss computed between the student predictions and the teacher predictions is added to the total loss with a scaling factor warming up exponentially from 0 to 2 in 30 epochs.

Training details. For all our experiments, we supervise the systems with the fine-grained taxonomy of SONYC [1]. We use the Adam optimizer [19] with a learning rate of 5×10^{-4}. To mitigate the issue of class imbalance, our loss is computed with the focal loss [20]. Using this loss instead of the binary cross-entropy loss helps improving class-averaged performances because it reduces the influence of overrepresented classes (e.g. "engine") in the loss. We apply data augmentations to the final spectrograms with Mixup [21] with a probability of 50% and frequency masking [22] of up to 48 bins. Training is done in a maximum of 100 epochs and is interrupted if no improvement is observed for 20 epochs on the objective metric (macro-averaged AUC-PR) computed on the validation set. The validation set is comprised of the labelled datasets used during training.

Evaluation sets. Irrespective of the training batch configuration, we evaluate the systems on the official evaluation set of SONYC and on the previously defined evaluation set of SINGA:PURA. We evaluate them on the coarse-level hierarchy by converting fine-grained class predictions into coarse-grained ones. Prior to evaluation, we adjust the decision threshold of each class on the validation set, using the GHOST [23] algorithm.

4.2. Training of the embedding classifier

We take inspiration from the evaluation of downstream tasks with precomputed embeddings in the HEAR challenge [10]. With its official implementation, we compute PaSST [13] timestamp embeddings with its default parameters: a hop size of 50 ms and an embedding size of 1295 (768 from the projection layer + 527 from the Audioset classification head). Likewise, we use the official implementation of openL3 [14] and its default parameters: a hop size of 100 ms and and embedding size of 512).

For a fair comparison, we use the same training protocol described in the previous subsection (without using unlabelled data, the Mean Teacher paradigm or any data augmentation) and we use the same classification head as the system trained from scratch. With this classification head, we reached superior performances compared to the Multi-Layer Perceptron used for the evaluation on downstream tasks of the HEAR challenge.

4.3. Metrics and Parameters

Performance metrics. In this study, we focus our systems evaluation on clip-level SED. We measure the performances of the chosen systems using the micro-averaged and macro-averaged scores of AUC-PR (Area Under the Curve - Precision Recall), and F1-measure. AUC-PR scores are computed globally and are used to compare the systems to the state-of-the-art. F1 scores are computed by mini-batches of 10 samples, presented in boxplots and analyzed in our discussion.

Proximity-related evaluations. Using the "proximity" annotations, we create subsets containing only "near" or "far" ground-truths. If events in a clip share the same proximity, evaluation is straight-forward. However, when a clip contains both "near" and "far" events, we need to post-process predictions and ground-truths to disambiguate the evaluation. Because mixtures are not synthetic, we cannot remove sound events when they do not correspond to the proximity of the subset. Therefore, we need to adapt *a posteriori* the predictions and the ground-truths. If the system correctly predicts a sound event whose proximity value is not under study, we do not want to penalize it so we zero out the predicted value for this class consistently with the fact that we had removed it as well from the ground-truth. This processing can artificially increase precision scores. Score metrics obtained on those subsets are still useful in our analysis despite reflecting performances only partly.

Polyphony-related evaluations. We create subsets based on a value of "event polyphony". For each clip, we define a "weak polyphony" value equal to the number of different sound sources present in the groundtruth. We use three subsets named "monophony" ($p = 1$), "low polyphony" ($p = 2$) and "high polyphony" ($p \geq 3$). Polyphony is derived from the annotations as the number of different classes labelled in the clip.

5. DISCUSSION

Performances w.r.t state-of-the-art solutions. We ensure that the systems that we exploit perform well enough on the SONYC test set with a matching train/test paradigm so that we can conduct a reliable analysis of cross-city generalization (in an unmatched train/test paradigm) solely based on our systems performances. We refer to the last edition of the SONYC-UST challenge [1] where the submitted systems were trained with the same amount of training data and evaluated on the same set as the CRNN and the embedding classifiers. The only difference is that we did not exploit the spatio-temporal context, which was successfully used in the challenge to boost performances. The winning solution achieves a micro-averaged and macro-averaged AUC-PR of respectively 83.5% and 64.5% while the baseline system provided by the SONYC-UST team achieves scores of 74.9% and 51.0 %. Our best run comes from using CRNN with the PCEN frontend with scores of 80.8% and 60.1%. Our best run using the embedding classifiers gives scores of 78.0% and 57.3% with PaSST [13] as a feature extractor. Hence our systems, though perfectible, are competitive with the state of the art.

Influence of front-end. We indicate in Figure 1, the F1-measure obtained on the SONYC evaluation set with different frontends (log, pcen, log+pcen). The improvements brought using PCEN are not as significant as the ones observed in other audio classification tasks [15, 24]. Using both transforms as 2-channel input did not help improving performances either. We reached the same conclusions in the unmatched train/test paradigm.

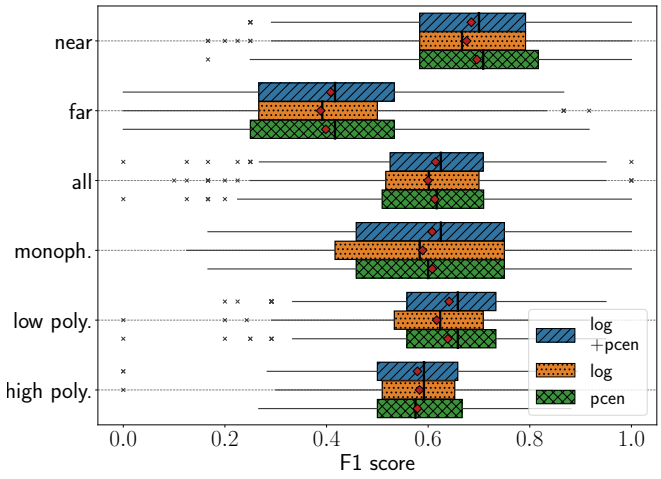

Fig. 1: CRNN performances when using different front-ends in a matched train/test paradigm. The systems are trained and evaluated on SONYC.

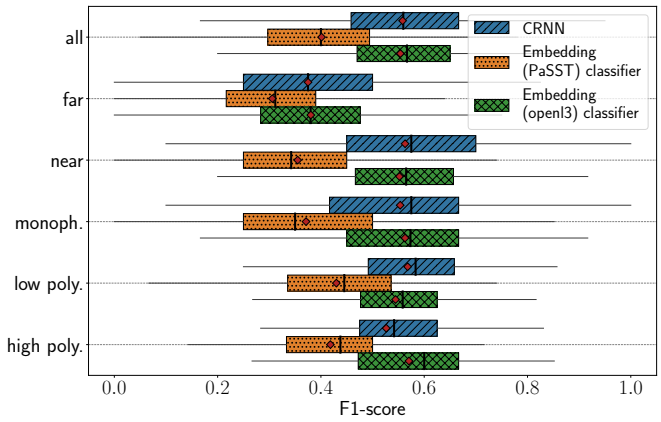

Fig. 2: CRNN and embedding classifiers performances in an unmatched train/test paradigm. The systems are trained on SONYC and evaluated on SINGA:PURA.

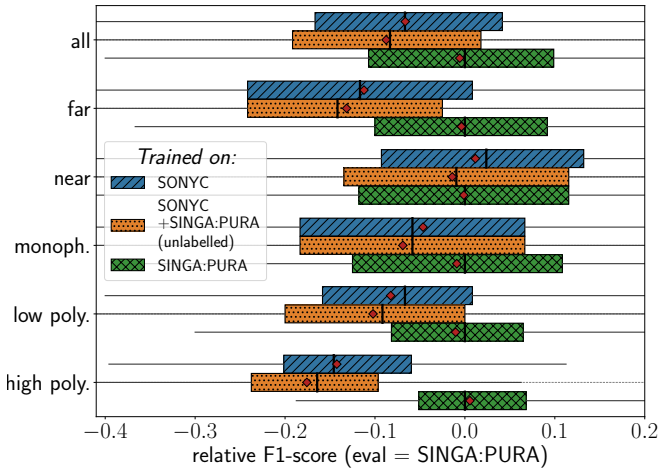

Fig. 3: CRNN performances on SINGA:PURA evaluation set under various train/test paradigms. The legend indicates which datasets were used to train the system. F1-scores are shown relative to the matched train/test paradigm.

Generalization across cities. It is important to acknowledge that cross-city generalization depends on more factors than the cities themselves (e.g. sensors, annotator processes and populations). Referring to Figure 2, we show our systems performances, trained on SONYC only, in an unmatched train/test settings. While performances are similar in a matched settings, the embedding (PaSST) classifier, underperforms compared to the other systems when confronted with a domain shift whereas the embedding (openL3) classifier is more robust and on-par with the CRNN. The embedding classifier therefore generalized better to unseen conditions with a feature extractor trained in an unsupervised way. In addition, we indicate in Figure 3, the F1 scores obtained on the SINGA:PURA evaluation set with the CRNN, when varying the training batch content. As expected, the lowest performances on the SINGA:PURA evaluation set are obtained when the CRNN is trained on SONYC only. No improvements were obtained in our experiments when using unlabelled data from SINGA:PURA with the Mean-Teacher paradigm [11]. The biggest F1 score differences between the unmatched train/test paradigm and the matched one are observed on the "far" and "polyphonic" subsets. This suggests the need for alternative domain adaptation methods to bridge the gap between both type of proximity and various polyphony levels.

Influence of proximity. No matter which system is used, we see a statistically significant degradation of performances on the SONYC "far" subset compared to the SONYC "near" subset. However, Referring to Figure 1 and Figure 2, the performance gap between the SINGA:PURA "far" subset and "near" subset is smaller. The proximity annotation strategy based on salience we used on the evaluation set of SONYC was probably a more reliable approach for our study than the one used by the SINGA:PURA annotators.

Influence of polyphony. Referring to Figure 1 and Figure 2, we observe that the CRNN system's predictions are impacted negatively as the polyphony value increases in the matched train/test paradigm. Interestingly, in the unmatched train/test paradigm, the performance gap between the CRNN trained from scratch and the embedding classifier is explained by the polyphony subsets. The embedding-based classifier fares better on polyphonic clips. This could be explained by the pretraining of openL3 on Audioset [9], which contains large amounts of polyphonic clips.

6. CONCLUSION AND FUTURE WORK

We have introduced CoSMo, a corpus allowing for an in-depth study of urban sound event detection combining three existing datasets of real far-field urban audio recordings. By evaluating baseline systems trained from scratch in a semi-supervised way or fine-tuned using pretrained general-purpose audio embeddings, we highlight performance gaps between subsets of audio recordings which are harder because of sound event proximity or high levels of polyphony. We also show that generalization across cities is hindered mostly by those hard examples. For future work, normalization and domain adaptation strategies should be explored [25]. To improve performances on "far" sound events or on highly polyphonic clips, using unsupervised sound source separation before the multilabel classification could help as it was successfully applied for bird sound classification [18]. More recent semi-supervised methods could also be used to better exploit unlabelled data [26]. We hope our insights can foster new research on urban sound event detection that will help bridging the performance gaps identified in this study.

7. REFERENCES

[1] M. Cartwright, J. Cramer, A. E. M. Méndez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon, O. Nov, and J. P. Bello, "Sonyc-ust-v2: An urban sound tagging dataset with spatiotemporal context," in *DCASE*, 2020.

[2] K. Ooi, K. N. Watcharasupat, S. Peksi, F. A. Karnapi, Z.-T. Ong, D. Chua, H.-W. Leow, L.-L. Kwok, X.-L. Ng, Z.-A. Loh, and W.-S. Gan, "A strongly-labelled polyphonic dataset of urban sounds with spatiotemporal context," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 982–988.

[3] E. Vidaña-Vila, J. Navarro, D. Stowell, and R. M. Alsina-Pagès, "Multilabel acoustic event classification using real-world urban data and physical redundancy of sensors," *Sensors (Basel, Switzerland)*, vol. 21, no. 22, pp. 7470, November 2021.

[4] F. Gontier, V. Lostanlen, N. Fortin, M. Lagrange, C. Lavandier, and J.-F. Petiot, "Polyphonic training set synthesis improves self-supervised urban sound classification," *Journal of the Acoustical Society of America*, June 2021.

[5] N. Turpault and R. Serizel, "Training Sound Event Detection On A Heterogeneous Dataset," in *DCASE Workshop*, Tokyo, Japan, Nov. 2020.

[6] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 67–83, 2021.

[7] F. Ronchini and R. Serizel, "A benchmark of state-of-the-art sound event detection systems evaluated on synthetic soundscapes," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1031–1035.

[8] M. Fuentes, B. Steers, P. Zinemanas, M. Rocamora, L. Bondi, J. Wilkins, Q. Shi, Y. Hou, S. Das, X. Serra, and J. P. Bello, "Urban sound & sight: Dataset and benchmark for audio-visual urban scene understanding," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 141–145.

[9] J. F. Gemmeke, D. P. W. 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)*, 2017, pp. 776–780.

[10] J. Turian, J. Shier, H. R. Khan, B. Raj, B. W. Schuller, C. J. Steinmetz, C. Malloy, G. Tzanetakis, G. Velarde, K. McNally, M. Henry, N. Pinto, C. Noufi, C. Clough, D. Herremans, E. Fonseca, J. Engel, J. Salamon, P. Esling, P. Manocha, S. Watanabe, Z. Jin, and Y. Bisk, "Hear: Holistic evaluation of audio representations," in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, D. Kiela, M. Ciccone, and B. Caputo, Eds. 06–14 Dec 2022, vol. 176 of *Proceedings of Machine Learning Research*, pp. 125–145, PMLR.

[11] 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, vol. 30, Curran Associates, Inc.

[12] F. Ronchini and R. Serizel, "A benchmark of state-of-the-art sound event detection systems evaluated on synthetic soundscapes," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1031–1035.

[13] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *INTERSPEECH*, 2022.

[14] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3852–3856.

[15] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5670–5674.

[16] V. Lostanlen, J. Salamon, M. Cartwright, B. McFee, A. Farnsworth, S. Kelling, and J. P. Bello, "Per-channel energy normalization: Why and how," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 39–43, 2019.

[17] C. Ick and B. McFee, "Sound event detection in urban audio with single and multi-rate pcen," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 880–884.

[18] T. Denton, S. Wisdom, and J. R. Hershey, "Improving bird classification with unsupervised sound separation," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 636–640.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007.

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

[22] 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*, 2019, pp. 2613–2617.

[23] C. Esposito, G. A. Landrum, N. Schneider, N. Stiefl, and S. Riniker, "Ghost: Adjusting the decision threshold to handle imbalanced data in machine learning," *Journal of Chemical Information and Modeling*, vol. 61, no. 6, pp. 2623–2640, 2021.

[24] V. Lostanlen, J. Salamon, A. Farnsworth, S. Kelling, and J. P. Bello, "Robust sound event detection in bioacoustic sensor networks," *PLoS ONE*, vol. 14, 2019.

[25] M. Olvera, E. Vincent, and G. Gasso, "On the impact of normalization strategies in unsupervised adversarial domain adaptation for acoustic scene classification," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 631–635.

[26] L. Cances, E. Labbé, and T. Pellegrini, "Comparison of semi-supervised deep learning algorithms for audio classification," *EURASIP J. Audio Speech Music Process.*, vol. 2022, no. 1, sep 2022.