ARCA23K: AN AUDIO DATASET FOR INVESTIGATING OPEN-SET LABEL NOISE

Turab Iqbal, Yin Cao, Andrew Bailey, Mark D. Plumbley, Wenwu Wang

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, UK
{t.iqbal, yin.cao, andrew.bailey, m.plumbley, w.wang}@surrey.ac.uk

ABSTRACT

The availability of audio data on sound sharing platforms such as Freesound gives users access to large amounts of annotated audio. Utilising such data for training is becoming increasingly popular, but the problem of label noise that is often prevalent in such datasets requires further investigation. This paper introduces ARCA23K, an Automatically Retrieved and Curated Audio dataset comprised of over 23 000 labelled Freesound clips. Unlike past datasets such as FSDKaggle2018 and FSDnoisy18K, ARCA23K facilitates the study of label noise in a more controlled manner. We describe the entire process of creating the dataset such that it is fully reproducible, meaning researchers can extend our work with little effort. We show that the majority of labelling errors in ARCA23K are due to out-of-vocabulary audio clips, and we refer to this type of label noise as open-set label noise. Experiments are carried out in which we study the impact of label noise in terms of classification performance and representation learning.

Index Terms— Audio dataset, audio classification, label noise, machine learning, deep learning, neural networks

1. INTRODUCTION

Labelled audio data is a relatively scarce resource, yet it is vital for training audio classifiers in a supervised fashion. With the emergence of online sharing platforms such as Freesound [1] and YouTube [2], users now have access to massive amounts of annotated audio, and it is becoming increasingly popular to utilise this data for training. For classifying general sound events, early examples of web-sourced datasets include ESC-50 [3] and UrbanSound8K [4]. However, these datasets are relatively small, which is largely due to the high cost of manually verifying the data to ensure the sounds are relevant and the labels are correct. At the time of writing, the largest sound event dataset with thoroughly verified labels is FSD50K [5], which contains approximately 50 000 sounds and is the result of several years of crowdsourced labelling [1].

Given the cost of label verification, there has been interest in reducing or eliminating this aspect of dataset curation. AudioSet [2], for example, is a large-scale audio dataset comprised of over two million sounds across hundreds of classes. AudioSet classes belong to an ontology in which the classes share parent-child relationships. Although AudioSet clips have been manually verified by listeners, the process was not thorough, and many labelling errors remain [6]. The labels of other datasets, such as VGGSound [7], have not been verified at all. In the case of FSDKaggle2018 [8], FSDKaggle2019 [9], and FSDnoisy18k [10], only a small subset of the dataset has been manually verified. Nevertheless, these datasets are attractive because they are relatively large. The challenge is that the presence of labelling errors, or *label noise*, can significantly impact learning [10]. Hence, studying the effects of label noise is important.

Due to label noise, rather than the training examples being drawn from the true distribution, P, examples are drawn from a corrupted distribution, Q. In the literature, the noise process responsible for this corruption is typically assumed to be reversible, such that any incorrectly-labelled instance can be relabelled [11, 12]. This is *not* a realistic assumption when retrieving and labelling web data, as the sounds that are retrieved can be *out-of-vocabulary (OOV)* [10]. OOV sounds are sounds that do not belong to any of the classes of interest. We refer to this type of label noise as *open-set label noise*. There is currently a disconnect where much of the analysis and tools are for closed-set label noise, while open-set label noise has received little attention in this respect. While there are works that address datasets with open-set label noise [13, 14, 6, 9, 10], the analysis is limited by the lack of empirical insight.

In this paper, we introduce *ARCA23K*[1] (Automatically Retrieved and Curated Audio 23K), which is a dataset containing more than 51 hours of audio data across 23 727 Freesound clips and 70 classes taken from the AudioSet ontology. The clips comprising the training set have been retrieved and curated using an entirely automated process, while the validation set and test set are subsets of FSD50K. Given the absence of human verification, labelling errors are to be expected in the training set. In particular, many of the audio clips are out-of-vocabulary.

Our aim in creating ARCA23K is to facilitate the study of real-world, open-set label noise, including its effects on learning and how these effects can be mitigated. Unlike datasets such as FSDnoisy18k, ARCA23K allows studying label noise in a more controlled manner. Instead of manually verifying a subset of the dataset, we introduce another dataset called *ARCA23K-FSD*, which is a subset of FSD50K. ARCA23K-FSD is essentially a ‘clean’ counterpart of ARCA23K. Under certain assumptions, this setup allows controlling the amount of label noise by substituting clips from one dataset with clips from the other. A similar idea was proposed in the image domain [15].

The contributions of this paper are four-fold. First, we provide a detailed description of how the ARCA23K datasets were created and release the associated source code[2]. Our intention is to provide a method of dataset creation that is realistic while also being easily reproducible[3], such that anyone can adopt or improve our method for their own needs. Our second contribution is the release of ARCA23K itself (along with ARCA23K-FSD). As all the clips are available under a Creative Commons license, we are able to distribute the clips freely. Third, we characterise the label noise present in ARCA23K by running listening tests. Finally, we conduct experiments to examine the impact of open-set label noise on training audio classifiers, which includes comparisons to synthetic label noise and an evaluation of the representations that are learned.

[1] `https://zenodo.org/record/5117901`

[2] `https://github.com/tqbl/arca23k-dataset`

[3] Some clips on Freesound may be deleted, which we cannot control.

2. ARCA23K-FSD

In order to investigate open-set label noise, we propose two datasets: a clean dataset with training examples drawn from P and a noisy dataset with training examples from Q. The number of examples per class is set to be equal across the two datasets. By satisfying this requirement, we are able to emulate the noise process that corrupts P to give Q. More specifically, a training example drawn from the clean dataset is corrupted by substituting it with a training example of the same class drawn from the noisy dataset. The amount of label noise can then be controlled by substituting a proportionate number of training examples.

The clean dataset, ARCA23K-FSD, is a subset of FSD50K [5], which is currently the largest clean dataset of sound events available. FSD50K is comprised of more than 40 k training examples and 200 classes taken from the AudioSet ontology. In general, multiple labels are associated with each audio clip.

For simplicity, we reduced FSD50K to a single-label dataset. First, clips containing more than one type of sound were discarded. Next, to prevent class overlap, classes that were ancestors of other classes (according to the AudioSet ontology) were dropped, e.g. clips labelled as Guitar would be dropped because Acoustic guitar and Electric guitar are child classes. Finally, any sound class with an insufficient number of audio clips was removed from the dataset. The thresholds are 50 instances in the training set, 10 instances in the validation set, and 20 instances in the test set. A total of 77 classes were retained after this pruning process. Let $\mathcal{L}$ denote the set of AudioSet labels that remained.

3. ARCA23K

In this section, we describe how the clips in the ARCA23K dataset were retrieved and curated. This dataset only includes a training set, since the validation set and test set of the ARCA23K-FSD dataset are used for validation and testing, respectively. We use a keyword-based algorithm to retrieve relevant clips and label them accordingly. We will assume that we have access to the metadata of every clip in the Freesound database and that we can download the clips. The metadata includes a description of the clip and a set of *tags* that are intended to be search terms. As with FSD50K, we limit our search to clips that are between 0.3 and 30 seconds [5]. After curation, all clips are converted to 16-bit mono WAV files sampled at 44.1 kHz.

3.1. Retrieval Algorithm

The general framework for the retrieval algorithm is as follows. For every candidate Freesound clip, the tags and description are tokenised and preprocessed to give two word sequences, d_{tags} and d_{desc}, which we refer to as *documents*. For each label, $l \in \mathcal{L}$, a *query*, q_l is constructed, which also involves tokenisation and preprocessing. Given q_l, d_{tags}, and d_{desc}, a relevance score, $r(q_l, d_{\text{tags}}, d_{\text{desc}}) \in [0, 1]$, is computed, such that a higher score indicates a better match between the query and the two documents. By computing scores for each label, the most relevant label can be assigned to the clip:

$$l^* := \arg\max_l r(q_l, d_{\text{tags}}, d_{\text{desc}}). \tag{1}$$

If $r(q_{l^*}, d_{\text{tags}}, d_{\text{desc}})$ is a low score, it indicates that none of the labels are a good match according to the algorithm. For this reason, clips for which $r(q_{l^*}, d_{\text{tags}}, d_{\text{desc}}) < \tau$ are discarded, where τ is a predefined threshold.

3.1.1. Tokenisation and Preprocessing

Tags, descriptions, and labels are tokenised and preprocessed using standard practices in information retrieval [16]. Tokenisation refers to converting a sequence of characters into a sequence of words. During this process, non-words such as punctuation and numbers are discarded. Tags are already assumed to be a sequence of words, hence tokenisation is not necessary for tags.

Preprocessing is carried out by first converting the words to lower-case so that retrieval can be case-insensitive. Following this, lemmatisation is applied to canonicalise words to their lemma form, e.g. 'guitars' would be reduced to 'guitar'. In cases where the lemma depends on which word class the word belongs to (e.g. verb, noun), the shortest lemma is chosen. For instance, 'clapping' would be reduced to 'clap' because, even though 'clapping' is the lemma for the noun, 'clap' is the lemma for the verb and is the shortest. After lemmatisation, stop words such as conjunctions and prepositions are removed. Finally, any duplicate words are also removed.

3.1.2. Query Construction

Given a label l, a query, q_l, is constructed as follows. The label is first tokenised and preprocessed as per Section 3.1.1. We will refer to the resulting output as a *root query* and denote it as $\bar{q}_l$. Next, a root query is constructed for every descendant label of l. For example, the label Bowed string instrument has several descendants, such as Cello and Double bass. After constructing the root queries, the final query q_l is constructed by concatenating all of them.

3.1.3. Computing Relevance Scores

In this section, we describe how relevance scores are computed. After creating a query for each label $l \in \mathcal{L}$, the vocabulary, V, can be defined as the concatenation of all the queries (after removing any duplicates). After constructing V, one can map any sequence of words, w, into a vector, $v(w) \in \{0, 1\}^{|V|}$, such that

$$v(w)_i := \begin{cases} 1 & w_i = V_i \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The relevance score is then defined as

$$r(q, d_{\text{tags}}, d_{\text{desc}}) := \bar{v}(q) \cdot [\bar{v}(d_{\text{tags}}) + \bar{v}(d_{\text{desc}})], \tag{3}$$

where $\bar{v}(w) := v(w)/\|v(w)\|$. In other words, $r(q, d_{\text{tags}}, d_{\text{desc}})$ is the cosine similarity between $v(q)$ and $[v(d_{\text{tags}}) + v(d_{\text{desc}})]$.

3.1.4. Evaluation

The Freesound clips that are labelled by our retrieval algorithm include all Freesound clips that are between 0.3 and 30 seconds in duration. This means the clips that constitute the ARCA23K-FSD dataset are labelled by our algorithm too. It is therefore possible to compare the labels assigned by our algorithm to the ground truth labels of ARCA23K-FSD.

We used a threshold of $\tau = 0.5$, as it resulted in a reasonable compromise between precision and recall. Our algorithm retrieved 84.1 % of the ARCA23K-FSD clips and achieved an accuracy of 90.3 %. The accuracy was greater than 90 % for 51 out of 77 classes, and the average accuracy for these classes was found to be 96.4 %. For the other 26 classes, the average accuracy was found to be 67 %.

It should be noted that the ARCA23K-FSD clips are not an unbiased sample of the retrieved clips. The aim of the evaluation is

Table 1: Estimates of the proportion of ARCA23K clips that are PP/PNP/NP. The percentage of clips marked 'Unsure' is $1\,\%$.

	PP	PNP	NP
IV	$(52.7{\pm}5.8)\%$	$(2.3{\pm}1.3)\%$	$(8.7{\pm}3.5)\%$
OOV	N/A	$(1.3{\pm}0.7)\%$	$(33.3{\pm}5.6)\%$

not to determine label accuracy in general but to demonstrate that it is comparable to approaches used for existing datasets. In Section 3.3, we evaluate the accuracy of the labels by manually verifying a subset of the ARCA23K dataset.

3.2. Curation

After labelling the candidate Freesound clips using the retrieval algorithm, we used a threshold of $\tau = 0.5$ to discard clips with a low relevance score. All clips belonging to the FSD50K dataset were also discarded to prevent any overlap. The number of retrieved clips at this point totalled almost $170\,\text{k}$. Next, the number of clips per class was reduced to match ARCA23K-FSD, since our aim is to create a dataset that mirrors ARCA23K-FSD. This was done by selecting a random sample of the correct size from each class. For seven of the classes, there was an insufficient number of clips to match the ARCA23K-FSD dataset, so the clips belonging to these classes were dropped altogether. The same was done for ARCA23K-FSD, resulting in 70 classes in total for both datasets.

3.3. Noise Rate Estimation

In this section, we describe how noise rates were estimated for the ARCA23K dataset and present the results. The noise rate is defined as the percentage of incorrectly labelled audio clips in the dataset. Similar to Fonseca et al. [5], we categorise clips as either 'Present and predominant' (PP), 'Present but not predominant' (PNP), 'Not present' (NP), and 'Unsure' (U). The reader is referred to the original work for detailed definitions [5]. PNP and NP are further split based on whether the other sounds are in-vocabulary (IV) or out-of-vocabulary (OOV). For example, NP/OOV means that at least one OOV sound can be identified in the clip.

The noise rate of the dataset was estimated by selecting a random subset of the dataset and performing listening tests. We selected 100 clips for the sample and repeated the experiment three times with replacement. Each sample was processed by a different listener, i.e. three listeners participated. The first three authors of this paper carried out the tests. They were trained by familiarising themselves with the classes, which involved reading the class descriptions and listening to example clips. They were also able to listen to example clips during the test and confer with each other[4].

The results are presented in Table 1. The noise rate can be calculated by excluding the sounds categorised as U. When PNP sounds are considered as incorrect, the noise rate was found to be $(46.4{\pm}4.8)\%$ (95% confidence interval). When PNP sounds are considered as correct, the noise rate was found to be $(42.4{\pm}4.1)\%$. Based on the results in Table 1, $75.9\,\%$ of incorrectly labelled clips are OOV. For many of the NP clips, we were able to identify them as NP from the tags and description alone[5], meaning that the labelling errors were the fault of the retrieval algorithm; some were simple mistakes, while others required understanding the context, which a keyword-based retrieval algorithm cannot infer. In other cases, the uploaders' annotations were misleading or incorrect. This was more prevalent with classes such as Whoosh, swoosh, swish, which are more open to interpretation without an agreed-upon definition. Finally, we observed that many of the OOV sounds were quite similar in sound to the IV classes. For example, 462351.wav, labelled as Acoustic guitar, contains sounds of a guitar string being strummed, but it is too distorted to belong to any of the guitar classes.

4. EXPERIMENTS

In this section, we describe the experiments that were carried out and present the results. Systems are evaluated using the accuracy and the mean average precision (mAP). The mAP is approximately equal to the area under the precision-recall curve; a higher value indicates better performance. We ran each experiment five times and provide 95% confidence intervals for the scores.

4.1. System

The machine learning model used in our experiments is an 11-layer convolutional neural network based on the VGG13 architecture [17]. Our model differs from VGG13 in that it uses batch normalisation [18] and only one fully-connected layer instead of three, as we found multiple fully-connected layers to be unhelpful.

The model was trained with mel-spectrogram inputs. Prior to computing the mel-spectrograms, the audio was downsampled from $44.1\,\text{kHz}$ to $32\,\text{kHz}$, which reduced the audio's data rate without significantly affecting the results. The mel-spectrograms were then computed using a $32\,\text{ms}$ frame length, a $16\,\text{ms}$ hop length, and 64 mel bins per frame. Finally, the amplitudes of the mel-spectrograms were scaled logarithmically.

Since the audio clips in both datasets vary in duration, we padded the clips with silence. Instead of padding to a single fixed length, we used three different lengths: 5 seconds, 15 seconds, and 30 seconds. The least amount of padding was applied to each clip, e.g. a clip less than 5 seconds would be padded to 5 seconds. When selecting clips for a mini-batch, only clips of the same length were allowed. Without this multi-length approach, each clip would have to be padded to the maximum length, which would greatly increase training times.

The model was trained for 50 epochs using the cross-entropy loss function and the AdamW optimiser [19] with a learning rate of 0.0005, which was decayed by $10\,\%$ every two epochs. We used a batch size of 64, 32, and 16 for 5-, 15-, and 30-second clips, respectively. By using different batch sizes, and given the memory constraints, we were able to significantly improve training times and even the classification accuracy. During inference, we averaged the predictions of the top three epochs in order to reduce volatility.

4.2. Adding Noise

In addition to training with the ARCA23K datasets, we also added synthetic label noise to the ARCA23K-FSD dataset, which allows us to compare synthetic label noise to the real-world label noise present in ARCA23K. The synthetic label noise is closed-set rather than open-set. Let $k \in \{1, \ldots, K\}$ represent the class associated with an instance, where $K = 70$ is the number of classes. To add synthetic noise, we selected a proportion, ρ, of training examples and changed the class k of each selected example to

$$(k + i) \mod K, \tag{4}$$

[4] In practice, listeners only conferred when they were unsure.

[5] All clips were listened to in their entirety nonetheless.

Table 2: Model performance when using different training sets.

Dataset	Accuracy	mAP
ARCA23K	$(50.08{\pm}0.78)\%$	$(52.32{\pm}0.77)\%$
ARCA23K-FSD	$(61.16{\pm}0.41)\%$	$(66.28{\pm}0.59)\%$
Uniform Noise	$(38.12{\pm}1.83)\%$	$(35.76{\pm}2.10)\%$
Conditional Noise	$(38.35{\pm}0.56)\%$	$(36.82{\pm}0.62)\%$

where i is a random integer drawn from a suitable distribution. Two types of label noise were considered: uniform and class-conditional. In the case of uniform noise, i followed the uniform distribution, $U(1, K-1)$. In the case of class-conditional noise, the geometric distribution was used with $p = 0.5$. This distribution is concentrated over a small number of outcomes, which is realistic because only a small set of classes tend to be incorrectly attributed to a sound.

Finally, we ran experiments in which we replaced a proportion, ρ, of the ARCA23K-FSD training examples with ARCA23K training examples. Recall from Section 2 that this is equivalent to controlling the noise rate of ARCA23K. For each example that was replaced, the replacement example was restricted to be identically labelled. The noise rate of the resulting mixed dataset is a fraction of the noise rate estimated in Section 3.3. For example, $\rho = 0$ corresponds to a noise rate of 0, $\rho = 1$ corresponds to a noise rate of $46.4\,\%$, and $\rho = 0.5$ corresponds to a noise rate of $23.2\,\%$.

4.3. Representation Learning

As a final set of experiments, we examined how label noise affects the representations that are learned. To do this, we trained a linear classifier on embeddings derived from the output of the VGG model's penultimate layer and evaluated its performance. The VGG model was first trained as normal using a noisy dataset (either ARCA23K or ARCA23K-FSD with synthetic label noise). Next, using the output of the penultimate layer as input data, a linear classifier was trained on the (clean) ARCA23K-FSD dataset. In addition to training with the whole of ARCA23K-FSD, we trained the linear classifier with $10\,\%$, $20\,\%$, and $50\,\%$ of the dataset.

4.4. Results

The first group of results are presented in Table 2. In this table, we compare the performance of the system when trained using: (1) ARCA23K, (2) ARCA23K-FSD, (3) ARCA23K-FSD but with uniform label noise, and (4) ARCA23K-FSD but with conditional label noise. We set $\rho = 0.45$ for both (3) and (4). The results show that training with ARCA23K-FSD gives an mAP score of $66.28\,\%$, which is 14% higher than when training with ARCA23K, which suggests that the presence of open-set label noise has a considerable effect on training. However, it can be seen that the effect of synthetic label noise is much more severe, as the mAP drops below 40%.

We hypothesise that there are at least two reasons why real-world, open-set label noise has a milder effect on training. First, we believe that OOV clips are inherently less likely to harm performance compared to mislabelled IV clips, especially if the OOV clips sound very different to the IV clips. Recall that most of the incorrectly labelled clips in ARCA23K are OOV. Second, as observed in Section 3.3, a considerable number of OOV clips were found to be similar in sound to the IV clips. If these clips are labelled accordingly (e.g. 462351.wav labelled as Acoustic guitar), they can be considered as surrogates of the IV clips. Rather than being detrimental to learning, these surrogates are likely to be beneficial. Both of these hypotheses were verified to some extent in previous work [13].

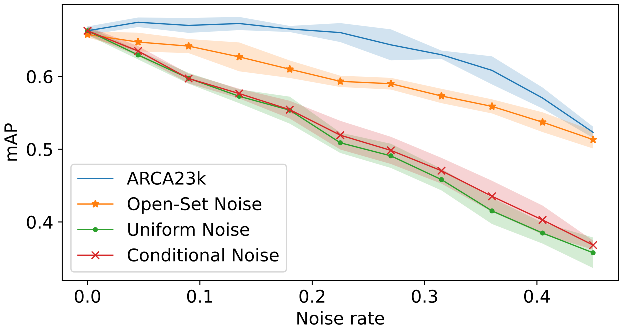

Figure 1: The mAP scores as ρ is varied from 0 to 0.45.

Table 3: mAP scores for the linear classifier. Columns indicate the percentage of ARCA23K-FSD clips used for training.

Dataset	$10\,\%$	$20\,\%$	$50\,\%$	$100\,\%$
ARCA23K	$55.27\,\%$	$58.94\,\%$	$58.91\,\%$	$59.82\,\%$
Uniform Noise	$30.86\,\%$	$37.11\,\%$	$42.29\,\%$	$45.52\,\%$
Conditional Noise	$41.15\,\%$	$46.06\,\%$	$48.11\,\%$	$50.09\,\%$
Random Weights	$7.93\,\%$	$10.11\,\%$	$13.12\,\%$	$16.23\,\%$

In Figure 1, we present the mAP scores as ρ (refer to Section 4.2) is varied from 0 to 0.45 at increments of 0.045. For all three types of noise, the performance generally decreases as ρ increases. However, while the plots appear linear for synthetic label noise, the plot for real-world, open-set label noise is non-linear. The performance decreases exponentially as the noise rate increases, albeit it is roughly the same until the noise rate exceeds $20\,\%$.

The results for the experiments described in Section 4.3 are presented in Table 3. We have also reported the performance when using a randomly-initialised VGG model to compute the embeddings. Similar to the results in Table 2, the performance is considerably worse when using synthetic label noise. When using ARCA23K to learn the representation, the performance of the linear classifier is relatively high even when training with $10\,\%$ of ARCA23K-FSD. On the other hand, the scores are still significantly lower than the score of $66.28\,\%$ in Table 2. These results show that label noise has a substantial effect on the quality of the learned representations.

5. CONCLUSION

In this paper, we introduced the ARCA23K dataset along with the companion ARCA23K-FSD dataset, which were created with the intention of studying open-set label noise. ARCA23K was created with minimal human labour by retrieving and curating clips from the Freesound database using an automated process, while ARCA23K-FSD was derived from FSD50K. We described the dataset creation process in detail and characterised the type of label noise present in ARCA23K via listening tests. Using these datasets, we were able to study the effect of label on learning in a controlled manner. We found that, while open-set label noise negatively affected performance, the impact was considerably milder than that of synthetic label noise. Furthermore, our experiments showed the extent to which label noise affects the learned representations of a model.

6. ACKNOWLEDGEMENT

This work was funded by the Engineering and Physical Sciences Research Council (EPSRC) Doctoral Training Partnership grants EP/N509772/1 and EP/R513350/1. It was also supported in part by EPSRC project EP/T019751/1 and by a Newton Institutional Links Award from the British Council with grant number 623805725.

7. REFERENCES

[1] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: A platform for the creation of open audio datasets," in *Proc. 18th Int. Society Music Information Retrieval Conf. (ISMIR)*, Suzhou, China, 2017, pp. 486–493.

[2] 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)*, New Orleans, LA, USA, 2017, pp. 776–780.

[3] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, New York, NY, USA, 2015, pp. 1015–1018.

[4] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM International Conference on Multimedia*, Orlando, FL, USA, 2014, pp. 1041–1044.

[5] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An open dataset of human-labeled sound events," *arXiv preprint arXiv:2010.00475*, Oct. 2020.

[6] B. Zhu, K. Xu, Q. Kong, H. Wang, and Y. Peng, "Audio tagging by cross filtering noisy labels," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2073–2083, Jul. 2020.

[7] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A large-scale audio-visual dataset," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.

[8] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of Freesound audio with AudioSet labels: Task description, dataset, and baseline," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, Woking, UK, 2018, pp. 69–73.

[9] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York, NY, USA, 2019, pp. 69–73.

[10] E. Fonseca, M. Plakal, D. P. W. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019, pp. 21–25.

[11] G. Patrini, A. Rozza, A. Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2233–2241.

[12] S. Sukhbaatar, J. Bruna, M. Paluri, L. Bourdev, and R. Fergus, "Training convolutional networks with noisy labels," in *International Conference on Learning Representations*, San Diego, CA, USA, 2015.

[13] T. Iqbal, Y. Cao, Q. Kong, M. D. Plumbley, and W. Wang, "Learning with out-of-distribution data for audio classification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 636–640.

[14] A. Kumar, A. Shah, A. Hauptmann, and B. Raj, "Learning sound events from webly labeled data," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, Macao, China, 2019, pp. 2772–2778.

[15] L. Jiang, D. Huang, M. Liu, and W. Yang, "Beyond synthetic noise: Deep learning on controlled noisy labels," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, vol. 119, 2020, pp. 4804–4815.

[16] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*. Cambridge University Press, 2008.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, vol. 37, Lille, France, 2015, pp. 448–456.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019.