Few-Shot Audio Classification Model for Detecting Classroom Interactions Using LaSO Features in Prototypical Networks

Rashed Iqbal, Christian Ritz, Jack Yang and Sarah Howard
University of Wollongong, NSW, Australia
E-mail: mri510@uowmail.edu.au, critz@uow.edu.au, jiey@uow.edu.au, sahoward@uow.edu.au

Abstract— **This research introduces an innovative approach to few-shot sound classification applied to classroom sound recordings that integrates Label Set Operation (LaSO) features with Prototypical Networks. Traditional audio classification methods often require extensive labeled datasets, which can be impractical in real-world scenarios where obtaining large amounts of labeled audio data is challenging. This is particularly the case for the target application of automatically annotating long recordings of classroom audio to understand student learning in classrooms. This paper proposes an enhanced few-shot learning approach based on Prototypical Networks by incorporating LaSO features, to augment the feature space for the Prototypical Network. This methodology focuses on detecting and classifying teacher and student voices for future understanding and analysis of classroom interactions. Experimental results indicate the proposed approach incorporating LaSO features significantly improves classification accuracy of a prototypical network used for few-shot learning. This work paves the way for more advanced and automated solutions in educational environments, facilitating better monitoring and understanding of classroom dynamics.**

I. Introduction

A classroom environment is filled with various sounds that indicate the range of learning activities taking place. These sounds include teacher lectures, student questions, group discussions, and background noises. For education researchers, detecting patterns in these sounds can provide deeper insights into student learning dynamics. However, manually analyzing large volumes of recordings over extended periods can be impractical. Consequently, this study investigates an automated approach for classifying sounds in classroom audio recordings.

Previous techniques for classifying classroom audio involve training neural networks using features extracted from labeled classroom sound recordings, as seen in [1, 2]. The DART method uses straightforward features based on sound power levels, similar to the neural network approaches in [1]. In contrast, time-frequency features such as the mel-spectrogram [2] are frequently used. The neural networks applied for classroom sound classification in [1, 2] include Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and RNN variants like Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks. This classification falls within the broader field of environmental sound or scene classification[3, 4].

In the recent past , the author explored utilizing sound power level features derived from the Decibel Analysis for Research in Teaching (DART) algorithm [5] to classify classroom audio captured by their developed system [6]. Other scholars have examined different techniques for analyzing classroom audio as well. For example, a Multi-Scale Audio Spectrogram Transformer (MAST) was developed to detect interactions between teachers and students during classroom activities [7]. Nonetheless, this research mainly focused on verbal exchanges between teachers and students, neglecting other important classes essential for understanding significant learning activities.

Environmental sound recognition aims to categorize different types of sound events within a recording. This fundamental task in machine listening has numerous practical applications, including smart cities [8, 9] and bioacoustics [10]. Although recent studies have made significant progress in recognizing sound events using extensive labeled datasets [11, 12], these approaches often fall short in real-world situations. This is due to the substantial effort required to collect enough annotated data for each category during the inference stage. Recently, several studies have suggested employing few-shot learning for environmental sound classification [13-15]. These classifiers can quickly learn new acoustic patterns from a limited number of labeled examples, mainly because of their unique training objective. Nevertheless, they remain constrained to using ground truth as a binary attribute also they are not implemented on long audio streams.

This study involves first extracting LaSO [16] features from audio samples using a pre-trained base model. These features are then used as input to a Prototypical Network, which is trained to classify sound events with minimal labeled examples by calculating distances further details in section II. We apply this framework to the domain of classroom interaction detection, focusing on identifying specific interactions and sounds within a classroom setting yield an even better understanding of classroom. The contribution of the study described in three parts:

i. We utilized scalogram audio features to train the few-shot learning model

ii. Extracted LaSO features from audio samples using a pre-trained based model

iii. LaSO features are then incorporated in prototypical network to generate protypes and classify long classroom audio.

Section II outlines the related work on few shot sound classification and prototypical network. Section III encompasses an in-depth methodology. Section IV describes the experimental setup and data collection method. Section V discussed the results and detailed analysis of the and the performance comparisons across different models.

II. Related work

The issue of few-shot learning has garnered significant interest in the computer vision field as well as sound classification recently. In the Meta-Learning, or learning-to-learn, approach [17-20], models are trained on instances of few-shot learning tasks instead of individual labeled samples.

A. *Few shot Sound Categorization*

There are several studies that apply few-shot learning to everyday sound recognition [13, 15, 19, 21, 22]. Heggan et al. experimented with various few-shot algorithms on everyday sound datasets for single-label classification [15]. Wang et al. addressed the multi-label few-shot problem by creating synthesized datasets, FSD-MIX-CLIPS and FSD-MIX-SED, and compared model performances by controlling generative factors in FSD-MIX-SED [13]. Cheng et al. adapted existing single-label few-shot algorithms to multi-label classification using a One-vs.-Rest strategy [22], and tested their methods on the AudioSet [23] dataset. Shi et al. applied meta-learning algorithms and linear regression on AudioSet, finding that meta-learning outperformed other few-shot methods [19]. It is important to note that while both Cheng and Shi used AudioSet for their experiments, their results cannot be directly compared as AudioSet is not publicly released, and neither study detailed how the database was adapted for few-shot learning.

B. *Prototypical network*

A training set $A = \{(M_i, O_i)\}_{i-1}^{A}$ where M_i represents the feature vector, $O_i \epsilon O$ signifies the discrete label of the i-th example, and O is the label set comprising O classes, $O = \{1, \dots, O\}$. Prototypical networks are trained using a series of "N-way K-shot" classification tasks created from the training set A.

In an "N-way K-shot" problem, the classification task consists of three components: (a) a subset O_s of N classes sampled from the set O, (b) G examples (support examples) drawn from A for each class in O_s, and (c) P examples (query examples) also drawn from A for each class in O_s. For any class $n \in O_s$, let Sn be the set of support examples for that class, with $|S_n| = G$. The prototype b_n (1) for class n is calculated as the mean of the embedding vectors of the support examples in S_n. Formally:

$$b_n = \frac{1}{G} \sum\nolimits_{(M,O)\in S_n} g_\phi(M) \qquad (1)$$

where g denotes the embedding mapping realized by the model whose parameters are denoted collectively as ϕ.

For a given query example M_p, the model classifies by generating a probability distribution over N classes in O_s using a softmax function applied to the distances between M_p and the N prototypes within the embedding space. Specifically, the likelihood that M_p belongs to class $n \in O_s$ is computed as follows (2):

$$r_\phi(\hat{z}_p = n | M_p) = \frac{\exp\left(-a\left(g_\phi(M_p), b_n\right)\right)}{\sum_{l \epsilon O_s} \exp\left(-a\left(g_\phi(M_p), b_l\right)\right)} \qquad (2)$$

Where $\hat{z}_p$ represents the predicted label forM_p, and a is a distance metric, such as ℓ_2 or cosine distance. The network is optimized to reduce the negative log-probability of the correct class across the $N \times P$ query examples in (3):

$$\Gamma(\phi) = \sum\nolimits_{(M,O)\epsilon P} -log r_\phi(\hat{z} = O|M) \qquad (3)$$

where P is the set of query examples, $|P| = N \times P$.

While prototypical networks excel in various applications [24, 25], they are not directly applicable to multi-label few-shot classification. This is because formulating "N-way K-shot" problems becomes challenging when labels frequently co-occur in a multi-label context.

III. Proposed method

This study presents an innovative method for few-shot sound classification, focusing on scalogram spectral utilization for detecting classroom interactions by combining Label Set Operation (LaSO) features with Prototypical Networks. Scalogram features extracted from the raw audio files and used as input to the label set operation features where a pretrained model utilized and backbone of the feature extraction.

A. *Scalogram conversion*

A scalogram, (4) similar to a spectrogram, is formed from the absolute values of the Continuous Wavelet Transform (CWT) coefficients across time and scale in a two-dimensional format [23]. This representation has proven to be more effective than other time-frequency features in neural network-based audio classification [17].

$$CWT_c(s,t) = \int_{-\infty}^{\infty} x(u) \frac{1}{\sqrt{s}} \psi^* \left(\frac{u-t}{s}\right) du \qquad (4)$$

In this context (1), $x(u)$ is the input signal, is s the scale parameter, which is related to frequency, ψ^* is the conjugate of the mother wavelet, t is the translation parameter that shifts the wavelet function along the time axis, and u denotes the segment of the signal.

B. Label Set Operation (LaSO)

The method is schematically depicted in Fig. 1 The input scalograms images U and V , each associated with a corresponding set of multiple labels $L(U)$ and $L(V)$, respectively, are mapped into the joint feature space $\mathcal{H}$ as H_U and H_V . This feature space $\mathcal{H}$ is implemented using a backbone feature extraction network $\mathcal{B}$; for our experiments, we utilized InceptionV3[26] and ResNet-50 [27] backbones. Three LaSO networks, named T_{int}, T_{uni}, and T_{sub}, process the concatenated H_U and H_V to generate synthesized feature vectors within the same feature space $\mathcal{H}$. As indicated by its name $int = intersection$, T_{int} aims to synthesize a feature vector in (5).

$$T_{int}(H_U, H_V) = W_{int} \epsilon\, \mathcal{H} \qquad (5)$$

This corresponds to a theoretical scalogram image E for which $\beta(E) = W_{int}$ and $L(E) = L(U) \cap L(V)$. Essentially, this means that if a human were to observe and label I, the label set would be $L(U) \cap L(V)$. Similarly, T_{uni} and T_{sub} generate outputs W_{uni} and W_{sub} in $\mathcal{H}$, which are anticipated to correspond to the union of the label sets $L(U) \cup L(V)$ and the subtraction of the label sets $L(U) \cup L(V)$, respectively [16].

C. LaSO based Prototypical network

InceptionV3 is used as the base model, pre-trained on ImageNet. Then we exclude the final classification layer. The global average pooling layer converts the feature maps from the base model to a fixed-size vector by taking the average of each feature map. The pooled feature vectors the fed into LaSO networks are designed to learn features based on the intersection (T_{int}), union (T_{uni}), and subtraction (T_{sub}) of label sets. Scalogram spectral images are concatenated then passed through multiple dense layers with ReLU activation and dropout layers.

For each pair of support and query features, we apply the LaSO T_{int}, T_{uni}, T_{sub} networks to compute intermediate, universal, and subset-specific feature in (6-8):

$$W_{int} = T_{int}(H_u, H_v) \qquad (6)$$
$$W_{uni} = T_{uni}(H_u, H_v) \qquad (7)$$
$$W_{sub} = T_{sub}(H_u, H_v) \qquad (8)$$

Where W_{int}, W_{uni}, W_{sub} ϵ $\mathbb{R}^k$.

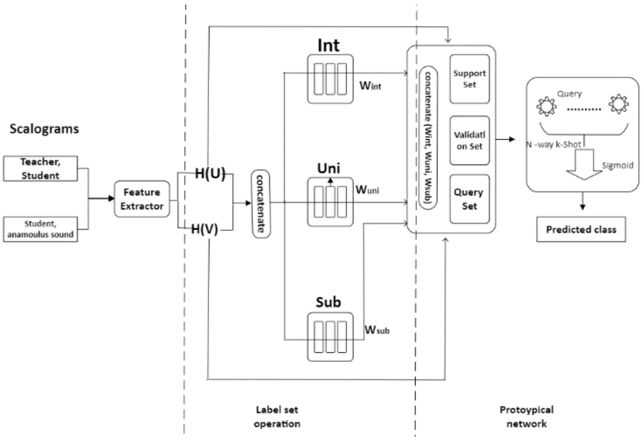

Fig. 1 LaSo based prototypical network process flow

Then Concatenated the LaSO features to form a combined feature vector (9):

$$W = concat(W_{int}, W_{uni}, W_{sub}) \in \mathbb{R}^{3k} \qquad (9)$$

In each episode of a prototypical network, we have a support set G and a query set P. For N-way K-shot learning, G contains N classes with K examples per class, and P contains query examples.

The prototype b_n (10) for each class n is found as the mean of its support examples

$$b_n = \frac{1}{K} \sum_{i=1}^{K} W_i^n \qquad (10)$$

Where W_i^n represents the combined LaSO feature for the i-th support examples of class n. Distance (11) between the query example W_P and each class prototype b_n:

$$a(W_p, b_n) = \left\|W_p - b_n\right\|_2^2 \qquad (11)$$

Then assign the query example W_p to the class with the nearest prototype $\hat{z}_p$ in (12):

$$\hat{z}_p = \arg min_m\, a(W_p, b_n) \qquad (12)$$

IV. Experimental setup

We employed prototypical networks for our study. The proposed LaSO based ProtoNet models were trained using 3-way classification tasks to align with the evaluation scenario, where only 3 classes (Teacher, Student, Anomalous sound) are considered. We excluded irrelevant labels for classes and were not included in the "N-way K-shot" setup.

The audio in a ninth-grade science classroom at an urban Australian high school was captured. This study, which extends a previous project, received approval from the University of Sydney's ethics committee and the New South Wales (NSW) Department of Education. Consent was obtained from the students, their parents, and the teacher. The classroom followed a Bring Your Own Device (BYOD) policy, using Microsoft OneNote for learning and note-taking. It was equipped with four cameras and audio recorders. The class consisted of 25 students and one teacher, meeting for 80 minutes four times a week.

In all experiments, audio recordings were sampled at 16 kHz, and the input was the Scalogram. The scalograms were generated from audio recordings using a 5 s window length. We applied inception V3 feature extractor for the audio then constructed LaSO based prototypical network and validated with validation set. Additionally, the network was tested on a long audio stream to evaluate the real-world implementation.

V. Results and Analysis

This section presents results from classifying classroom sounds using prototypical networks for an initial training and test database and for a separate unseen test database. Results are also presented when training these networks for classifying environmental sounds using two alternative well-known environmental sound databases (ESC [28] and Urban Sound [29]).

A. Training and initial Testing Performance

A database of 40 example audio segments across all 3 classes (13 teacher, 13 student and 14 anamolous classes) was prepared. From this database, a training set of 21 examples was selected and training is performed by randomly selecting 1 example of each class from this database, with the remaining 18 examples (6 per class) used for validation. This was repeated 3 times using a different example for each class each time (hence, three different models were trained). A separate test database was formed from the remaining 19 examples.

The validation accuracy in table 2 Table 2 shows the validation accuracy for one of the trained models (similar results were achieved for the other two models). The LaSO+Prototypical Network consistently outperforms the Prototypical Network across all scenarios. In the 3-way 1-shot setting, it achieves a validation accuracy of 62.05%, which is significantly higher than the Prototypical Network's performance. This trend continues with 82.90% accuracy in the 3-way 3-shot scenario and 95.95% in the 3-way 5-shot scenario. The LaSO+Prototypical Network's superior performance can be attributed to the integration of LaSO features, which enhances the model's ability to capture and utilize important relational and contextual information within the data.

Table 1. Validation accuracy of the trained model

Model	3-way 1-shot	3-way 3-shot	3-way 5-shot
Prototypical net	54.45%	74.18%	87.75%
LaSO+Prototypical net	62.05%	82.90%	95.95%

Table 2 shows the average results for the test dataset across all 3 models. The Prototypical Network shows reasonable performance, with accuracy improving from 55.35% in the 1-shot scenario to 88.75% in the 5-shot scenario. However, the LaSO+Prototypical Network significantly outperforms the Prototypical Network alone in all scenarios, achieving 63.65%, 84.60%, and 99.55% accuracy in the 1-shot, 3-shot, and 5-shot settings.

Table 2. Test accuracy of the proposed model

Model	3-way 1-shot	3-way 3-shot	3-way 5-shot
Prototypical net	55.35%	75.18%	88.75%
LaSO+Prototypical net	63.65%	84.60%	99.55%

Table 3 shows average results using different performance measures across the 3 models when used to classify the test set. The LaSO Prototypical Network across different few-shot learning scenarios exhibit a notable performance. For the 3-way 1-shot scenario, the mean Average Precision (mAP) is 62.05%, with precision at 62.00%, recall at 61.89%, and an F1 score of 60.05%. As the number of shots increases, the model's performance improves significantly. In the 3-way 3-shot scenario, the mAP rises to 82.90%, precision to 82.10%, recall to 81.50%, and the F1 score to 82.90%. The 3-way 5-shot scenario shows the highest performance with a mAP of 95.76%, precision at 95.50%, recall at 95.23%, and an F1 score of 95.50%. These results highlight the LaSO Prototypical Network's strong capability to learn and generalize from limited data, with consistent improvements in precision, recall, and F1 score as more examples per class are provided.

Table 1. Various accuracy matrices of the proposed model

Model	3-way 1-shot	3-way 3-shot	3-way 5-shot
mAP	62.05%	82.00%	95.76%
precision	62.00%	82.10%	95.50%
Recall	61.89%	81.50%	95.23%
F1 Score	62.05	82.90%	95.50%

B. Results from evaluation on a second test database

A second test database was formed from the classroom audio recordings and consisting of 40 examples (13 teacher, 13 student, and 14 anomalous examples). Table 4 shows the accuracy results for this test database.

Table 2. Testing results on unseen data

Model	3-way 1-shot	3-way 3-shot	3-way 5-shot
Prototypical net	25.35%	61.25%	70.75%
LaSO+Prototypical net	60.65%	80.60%	95.00%

The Prototypical Network achieved an accuracy of 25.35%, while the combined model significantly outperformed it with 60.65%. In the 3-way 3-shot scenario, with three examples per class, the Prototypical Network's accuracy improved to 61.25%, but the combined model still showed superior performance with 80.60%. Finally, in the 3-way 5-shot scenario, which involves five classes with five examples per class, the Prototypical Network reached 70.75% accuracy, whereas the LaSO+Prototypical Network achieved a notably higher accuracy of 95.00%.

As an example, Fig. 3 shows a 1-minute portion of the recorded classroom sounds with labeled sound class segments. The first image shows the predicted labels: '1 (Student)', '2 (Teacher)', and '3 (Anomalous Sound)', while the second image shows the ground truth graph with corresponding Letters: 'S (Student)', 'T (Teacher)', and 'A (Anomalous Sound)'.

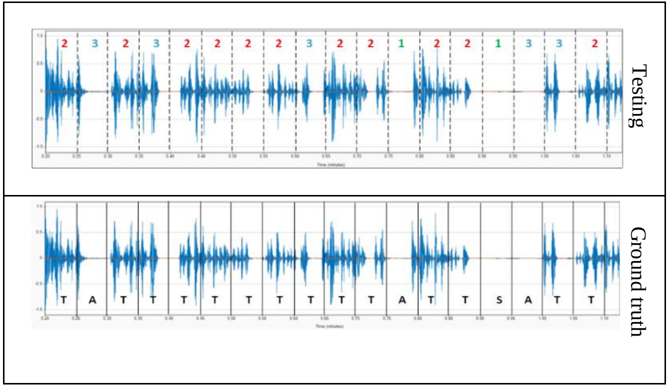

Fig 3. Testing and ground truth plot on long audio stream

There are a total of 36 instances which includes 27 Teachers, 5 anomalous sound and 4 students in the ground truth. While tested through our proposed model, the prediction shows that 24 instances are predicted as Teacher, the Student segment is predicted correctly, and anomalous sound predicted 4 times where actual ground truth is 5 and anomalous sound predicted 3 times over 4.

C. *Testing over ESC and Urban Sound dataset*

The proposed LaSO+Prototypical models trained on scalograms was further evaluated using two existing sound datasets, ESC [28] and Urban Sound [29], and three chosen classes from each database. The classes considered for ESC are Rain; Rooster; and person sneezing while for the Urban Sound dataset are: Gunshot; Dog Barking; and Siren. From each database, a set of 40 examples across all classes was created. Training proceeded using one randomly selected example for each class, leaving 37 examples that were used for testing (12 examples each for the first two classes and 13 examples for the third class). Accuracy results are shown in Table 5, including results for the classroom dataset using the LaSO+Prototypical model shown in Table 4.

Table 3. Testing the Proposed model on ESC and US dataset

Model	3-way1-shot	3-way 3-shot	3-way 5-shot
ESC	52.45%	72.18%	79.75%
Urban sound	56.05%	82.90%	92.95%
Classroom	60.65%	80.60%	95.00%

In the 3-way 1-shot scenario, where the model classifies among three classes with only one example per class, it achieved 52.45% accuracy on the ESC dataset, 56.05% on the Urban Sound dataset, and 60.65% on the classroom dataset. In the 3-way 3-shot scenario, with three examples per class, the model's performance improved, reaching 72.18% accuracy on the ESC dataset, 82.90% on the Urban Sound dataset, and 80.60% on the classroom dataset. In the 3-way 5-shot scenario, with five examples per class, the model achieved 79.75% accuracy on the ESC dataset, 92.95% on the Urban Sound dataset, and the highest accuracy of 95.00% on the classroom dataset

VI. Conclusion

This study presents an innovative approach for few-shot sound classification by leveraging scalogram spectral features to detect classroom interactions, combined with Label Set Operation (LaSO) features and Prototypical Networks. By converting raw audio files into scalograms and using these features as inputs for LaSO-based feature extraction, the study demonstrates the efficacy of scalograms over other time-frequency features in neural network-based audio classification. Experimental results reveal that the LaSO-enhanced Prototypical Network significantly outperforms the traditional Prototypical Network across various few-shot learning scenarios, achieving higher accuracy, mean Average Precision (mAP), precision, recall, and F1 scores. Specifically, the LaSO+Prototypical Network showed substantial improvements in 1-shot, 3-shot, and 5-shot settings. Validation and testing on a separate test dataset as well as two publicly available sound class datasets further confirmed the model's effectiveness. This approach demonstrates a robust capability to generalize from limited data, making it a valuable method for sound classification in educational settings and other domains requiring precise classification from minimal data.

VII. Acknowledgment

This research data was obtained from the ARC Discovery Project DP130100481.

References

[1] R. Cosbey, A. Wusterbarth, and B. Hutchinson, "Deep learning for classroom activity detection from audio," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019: IEEE, pp. 3727-3731.

[2] A. Mou, M. Milanova, and M. Baillie, "Active Learning Monitoring in Classroom Using Deep Learning Frameworks," in International Conference on Pattern Recognition, 2022: Springer, pp. 384-393.

[3] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206-219, 2019.

[4] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," Applied Sciences, vol. 10, no. 6, p. 2020, 2020.

[5] M. T. Owens et al., "Classroom sound can be used to classify teaching practices in college science courses," Proceedings of the National Academy of Sciences, vol. 114, no. 12, pp. 3085-3090, 2017.

[6] S. K. Howard, J. Yang, J. Ma, C. Ritz, J. Zhao, and K. Wynne, "Using data mining and machine learning approaches to observe technology-enhanced learning," in 2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE), 2018: IEEE, pp. 788-793.

[7] F. Liu and J. Fang, "Multi-Scale Audio Spectrogram Transformer for Classroom Teaching Interaction Recognition," Future Internet, vol. 15, no. 2, p. 65, 2023.

[8] G. Dove, C. Mydlarz, J. P. Bello, and O. Nov, "Sounds of New York city," Interactions, vol. 29, no. 3, pp. 32-35, 2022.

[9] A. Mitchell et al., "Deep learning techniques for noise annoyance detection: Results from an intensive workshop at the Alan Turing Institute," The Journal of the Acoustical Society of America, vol. 153, no. 3_supplement, pp. A262-A262, 2023.

[10] V. Morfi et al., "Few-Shot Bioacoustic Event Detection: A New Task at the DCASE 2021 Challenge," in DCASE, 2021, pp. 145-149.

[11] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880-2894, 2020.

[12] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," arXiv preprint arXiv:2104.01778, 2021.

[13] Y. Wang, N. J. Bryan, J. Salamon, M. Cartwright, and J. P. Bello, "Who calls the shots? Rethinking few-shot learning for audio," in 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2021: IEEE, pp. 36-40.

[14] J. Liang, Q. Phan, and E. Benetos, "Leveraging label hierarchies for few-shot everyday sound recognition," 2022.

[15] C. Heggan, S. Budgett, T. Hospedales, and M. Yaghoobi, "Metaaudio: A few-shot audio classification benchmark," in International Conference on Artificial Neural Networks, 2022: Springer, pp. 219-230.

[16] A. Alfassy et al., "Laso: Label-set operations networks for multi-label few-shot learning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6548-6557.

[17] C. So, "Exploring meta learning: parameterizing the learning-to-learn process for image classification," in 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), 2021: IEEE, pp. 199-202.

[18] X. Zhong, C. Gu, W. Huang, L. Li, S. Chen, and C.-W. Lin, "Complementing representation deficiency in few-shot image classification: A meta-learning approach," in 2020 25th international conference on pattern recognition (ICPR), 2021: IEEE, pp. 2677-2684.

[19] B. Shi, M. Sun, K. C. Puvvada, C.-C. Kao, S. Matsoukas, and C. Wang, "Few-shot acoustic event detection via meta learning," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: IEEE, pp. 76-80.

[20] T. Zhang, L. Yang, X. Gut, and Y. Wang, "A Task-Specific Meta-Learning Framework for Few-Shot Sound Event Detection," in 2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP), 2022: IEEE, pp. 1-6.

[21] J. Liang et al., "Adapting language-audio models as few-shot audio learners," arXiv preprint arXiv:2305.17719, 2023.

[22] K.-H. Cheng, S.-Y. Chou, and Y.-H. Yang, "Multi-label few-shot learning for sound event recognition," in 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), 2019: IEEE, pp. 1-5.

[23] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), 2017: IEEE, pp. 776-780.

[24] J. Huang, F. Chen, K. Wang, L. Lin, and D. Zhang, "Enhancing prototypical few-shot learning by leveraging the local-level strategy," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022: IEEE, pp. 1660-1664.

[25] R. Li, J. Liang, and Q. Phan, "Few-shot bioacoustic event detection: Enhanced classifiers for prototypical networks," 2022.

[26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818-2826.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

[28] K. J. Piczak, "ESC: Dataset for environmental sound classification," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015-1018.

[29] 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, 2014, pp. 1041-1044.