LEVERAGING LABEL HIERARCHIES FOR FEW-SHOT EVERYDAY SOUND RECOGNITION

Jinhua Liang, Huy Phan, Emmanouil Benetos

Centre for Digital Music, Queen Mary University of London, United Kingdom
{jinhua.liang, h.phan, emmanouil.benetos}@qmul.ac.uk

ABSTRACT

Everyday sounds cover a considerable range of sound categories in our daily life, yet for certain sound categories it is hard to collect sufficient data. Although existing works have applied few-shot learning paradigms to sound recognition successfully, most of them have not exploited the relationship between labels in audio taxonomies. This work adopts a hierarchical prototypical network to leverage the knowledge rooted in audio taxonomies. Specifically, a VGG-like convolutional neural network is used to extract acoustic features. Prototypical nodes are then calculated in each level of the tree structure. A multi-level loss is obtained by multiplying a weight decay with multiple losses. Experimental results demonstrate our hierarchical prototypical networks not only outperform prototypical networks with no hierarchy information but yield a better result than other state-of-the-art algorithms. Our code is available in: `https://github.com/JinhuaLiang/HPNs_tagging`

***Index Terms*—** Everyday sound recognition, few shot learning, hierarchical prototypical network

1. INTRODUCTION

Everyday sound recognition (or audio tagging) is to classify the types of environmental sound events in a recording or online stream, which involves many potential scenarios such as hearing aids [1], smart cities [2], and advanced healthcare [3]. In the past decades, a great amount of deep learning methods have emerged [4, 5] exploring how to boost audio networks' performance using large-scale datasets [6, 7]. While many works turned to focus on some more practical scenarios, such as mismatched domains [8], weakly supervised learning [9], and noisy labels [10], most of these methods are still restricted by the size of available datasets. This is a practical problem in the field of everyday sound recognition as it usually takes annotators more effort to mark the categories in a recording. In addition, everyday sounds cover thousands of categories, which makes it impossible to collect sufficient instances per class for supervised learning. This is thus how few-shot learning comes into the picture.

Inspired by the human ability to learn novel items with just a few examples, few-shot learning aims to capture the pattern of an unseen category using a handful of instances [11]. A typical few-shot learning framework is depicted as an N-way K-shot problem where there are N classes in a task and each class contains K instances for training. Currently only a few studies have attempted to apply few-shot learning to environmental sound recognition tasks. Although these works pioneered few-shot audio recognition, most of them were restricted to implementing off-the-shelf few-shot learning methods from other fields explicitly, which ignores exploiting the relationship between labels in audio taxonomy.

This work is motivated by the fact that we humans learn unseen concepts not only by observing their own features, but also by connecting them to existing knowledge. We thus assume that leveraging a priori knowledge helps a model to learn an unseen category with a few examples. Based on this assumption, this paper applies hierarchical prototypical networks (HPNs) to leverage the audio taxonomy knowledge drawn from the taxonomy of the dataset. Specifically, a few-shot classification problem is considered as a multi-task classification problem where both ancestor classes and descendant classes are used in separate classification tasks. Furthermore, prototypical networks are adopted as classifiers by measuring distance between query points and prototypes in the embedding space. Experimental results on the ESC-50 dataset [12] show that our HPNs yield a superior performance over prototypical networks with no hierarchy knowledge and outperform other state-of-the-art models.

The contributions of our work are three-fold:

i) Several state-of-the-art few-shot learning algorithms are benchmarked for generic everyday sound recognition. Different experimental setups are carried out to investigate the impact of data splits on model evaluation.
ii) A hierarchical prototypical network is proposed and applied to leverage a priori knowledge of sound event taxonomy by taking samples' ancestor classes into consideration.
iii) The impact of data splits on overall performance is investigated. The code is also released to benchmark state-of-the-art few-shot algorithms and to set up an evaluation environment on the ESC-50 dataset.

The remainder of this paper is organised as follows. Section 2 briefly summarises work related to our research and Section 3 introduces our proposed hierarchical prototypical network and the implementation details. In the Section 4, experimental results are discussed to demonstrate the superior performance of our network compared with other few-shot methods. Discrepancy in performance is then discussed among different data splits. Section 5 concludes the work and points out directions for future work.

2. RELATED WORK

2.1. Few-shot learning for everyday sound recognition

Few shot learning aims to use a limited amount of labeled examples to train a model that can be generalised to unseen categories easily. Suppose C_{base} and C_{novel} are two non-overlapping label sets (or splits) drawn from the whole label set C. The task is to train a classifier f with labelled samples of classes from C_{base} and to evaluate f on samples of classes belonging to C_{novel}. Transfer learning [13] and meta learning [14, 15] are two of the most frequently used techniques. On the one hand, transfer learning strategies train f

with abundant data from C_{base} and then fine-tune f through some iterations using a limited amount of data from C_{novel}. On the other hand, meta learning (or *episodic learning*) strategies [16, 15, 14] update model parameters through a series of independent tasks in the training process. A task herein is formed by drawing N classes from C_{base} each of which contains K "training" data and Q "test" data. To differentiate the above "training" and "test" data with the conventional terms, we refer to them as *support* and *query* data instead. The model is then trained to predict the categories of query data out of N classes by offering support data. The underlying assumption of episodic learning is aligning the training process with the evaluation benefits a model's ability to generalise to novel categories. However, there exist some works [13] stating that transfer learning can have a competitive performance as well. Therefore, it still remains an open problem to work out the best practice for few-shot learning.

One of the biggest challenges for everyday sound recognition nowadays is that it covers thousands of sound classes, while only hundreds of them are available in the existing large-scale datasets [6]. There are some attempts regarding how to apply few-shot learning in the everyday sound domain. Shi et al. [17] implemented several few-shot learning algorithms on a subset of the AudioSet dataset and demonstrated the advantage of meta-learning methods in the audio domain. Wang et al. [18] curated a synthesized audio dataset for few-shot audio recognition based on FSD50K [7] and compared the state-of-the-art by controlling annotated samples, polyphony levels, and signal-to-noise ratio (SNR). Heggan et al. [19] attempted to setup a benchmark for few-shot learning techniques in audio domains. In addition to generic everyday sound recognition, the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge [1] has been holding a task on fine-grained few-shot learning in the past two years which attracts increasing amount of attention. Although the previous works introduced few-shot learning paradigms to everyday sounds successfully, most of them ignored leveraging the relationship between labels in the audio taxonomy. Different from those works, this paper integrates audio taxonomy knowledge within existing few-shot algorithms.

2.2. Knowledge-based few-shot learning

Knowledge-based learning incorporates label taxonomy knowledge into the supervised learning [20]. There are some knowledge-based few-shot methods in few-shot scenarios [21, 22]. Peng et al. proposed a Knowledge Transfer Network (KTN) to incorporate visual features and semantic information for image recognition [21]. They used two independent classifiers to capture visual patterns and to conduct knowledge inference, followed by an integration network to merge them together. Due to the difference between sound and image, however, the definition of sound classes is more abstract and obscure for annotators compared with images. Garcia et al. designed hierarchical prototypical networks for music instrument recognition [22]. They aggregated classes according to a predefined instrument hierarchy and calculated the hierarchical loss by adding a weight decay to each level in the tree structure. Compared with music instrument classification, everyday sound recognition cannot be sorted into a tree structure by their physical properties merely. Some intermediate classes, such as "domestic sound" and "wild animal", are connected with psycho-acoustics directly, which makes the classification task even more complicated. Inspired by [22], this paper applies hierarchical prototypical networks to leverage the audio taxonomy knowledge derived from the dataset. Compared with their original implementation, our proposed models lower the limit on the number of prototypes for ancient prototypes generation, which encourages the model to use the audio taxonomy knowledge in more circumstances.

[1] https://dcase.community/

3. HIERARCHICAL PROTOTYPICAL NETWORKS

3.1. Prototypical networks

Prototypical networks were proposed in [16] to train a few-shot model using a series of independent tasks. Prototypical networks learn an embedding space where classification is performed by computing distances between each query sample and prototypes. Suppose a support set of $N \times K$ examples be $S = \{(x_1, y_2), ..., (x_{N\times K}, y_{N\times K})\}$ where each $x_i \in \mathbb{R}^D$ is the D-dimensional feature vector of an example obtained from an encoder f and $y_i \in \{1, ..., K\}$ be the corresponding label. A prototype could be calculated by averaging the embeddings of support samples belonging to its class:

$$c_k = \frac{1}{|S_k|} \sum_{(x_i,y_i)\in S_k} f_\phi(x_i) \tag{1}$$

where S_k denotes the set of examples whose ground truth is class k and ϕ are the learnable parameters in the encoder f.

Prototypical networks then produce a distribution over classes for a query embedding x using a softmax function over distances to the prototypes:

$$q_\phi(y = k|x) = \frac{\exp(-dist(f_\phi(x), c_k))}{\sum_m \exp(-dist(f_\phi(x), c_m))} \tag{2}$$

We applied the Euclidean distance as the distance function in our experiments. The probability distribution over classes is used to calculate cross-entropy loss.

$$L_{CE} = -\sum p(x) \log q(x) \tag{3}$$

where p, q are distributions of ground truth and predictions.

3.2. Hierarchical prototypical networks

Based on prototypical networks in Section 3.1, we devise hierarchical prototypical networks (HPNs) to incorporate the audio taxonomy knowledge into the training process. As shown in Figure 1, each level of the tree structure is treated as an independent multi-class classification task in the training stage. We thus build a HPN to extract the acoustic features with a shared encoder f and predict multiple labels corresponding to each level. Similar to (1), the acoustic features of support samples are used to calculate prototypes of the bottom level as:

$$c_k^{(0)} = \frac{1}{|S_k|} \sum_{(x_i,y_i)\in S_k} f_\phi(x_i) \tag{4}$$

To generate prototypical nodes of a higher level in the HPN, prototypes of the lower level are clustered together as per their parent level and aggregated to obtain the prototypes of a higher level:

$$c_j^{(h)} = \frac{1}{|C_k^{(h)}|} \sum_{c_j^{(h)}\in C_j^{(h)}} c_j^{(h-1)} \tag{5}$$

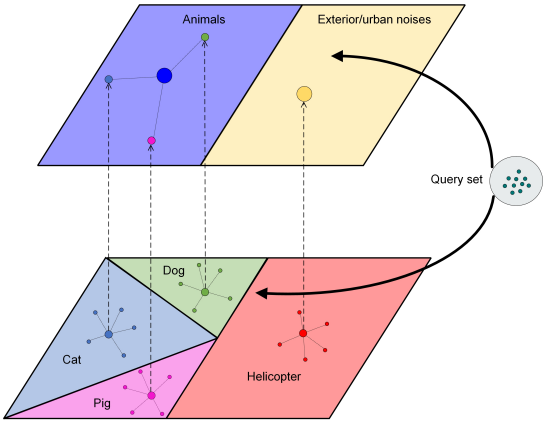

Figure 1: Illustration of training a hierarchical prototypical network in a multi-task scenario. The model learns to solve a 4-way 5-shot problem on the bottom level while trying to conduct a 2-way classification on a higher level.

Table 1: The architecture of the designed encoder.

Block name	Filter size	Output shape
Conv block 1	3×3@64	(64, 32, 215)
Conv block 2	3×3@128	(128, 16, 107)
Conv block 3	3×3@256	(256, 8, 53)
Conv block 4	3×3@512	(512,)
Fully connected layer	256	(256,)

where $C_j^{(h)}$ is a set of prototypes belonging to the class j at the h-th level in the taxonomy. The HPN iterates the clustering and aggregation process until the prototypes of the highest level are calculated.

Different from the original hierarchical network [22] which ignores an ancestor node if the number of its child nodes is less than two, HPN takes this ancestor class into consideration to fully exploit the hierarchical information. We believe this can leverage the audio taxonomy knowledge even better. The Euclidean distance between the query embedding and prototypes in each level is then calculated and used for classification. The evaluation process is similar to the training except only the classification task of the bottom level in HPNs will be taken into consideration for a fair comparison.

Following the architecture of VGGNet [23], we design a convolutional neural network containing 8 convolutional layers as the backbone of our HPNs, as shown in Table 1. Each block consists of two identical convolutional filters. Except for the last block, a max pooling layer with strides equal to 2 is appended to the block. A global pooling operation is used in the last block. Finally, the network outputs audio embedding sized 256.

3.3. Structural loss

Let h be the hierarchical level of a node in the tree structure, we can calculate the hierarchical loss [22] by using the cross-entropy loss function as follows:

$$L_{hierarchical} = \sum_{h=0}^{H} e^{\alpha h} L_{CE}^{(h)} \quad (6)$$

where $L_{CE}^{(h)}$ is the cross-entropy loss in the level h, H is the height of the taxonomy, and α is a hyper-parameter to control the loss decay of each level with respect to their height in the hierarchy. We set α equal to 1 in all experiments.

4. EVALUATION

4.1. Dataset

Table 2: Parent classes and examples of the child classes in ESC-50

Parent class	**Examples of child classes**
Animals	Dogs, Rooster, Pig, Cow, ...
Natural & water soundscapes	Rain, Sea waves, Crickets, ...
Human, non-speech sounds	Crying baby, Sneezing, ...
Interior/domestic sounds	Door knock, Clock tick , ...
Exterior/urban noises	Engine, Chainsaw, Siren, ...

The few-shot learning methods are evaluated on the ESC-50 dataset [12]. ESC-50[2] is a collection of sound events which consists of 2000 5-second recordings in total. Theses recordings are assigned one label out of 50 child classes. Table 2 shows the tree structure of ESC-50. It can be observed that these child classes are loosely arranged into 5 parent categories: "Animals", "Natural & water soundscapes", "Human, non-speech sounds", "Human, non-speech sounds", "Interior/domestic sounds", "Exterior/urban noises". It should be noted that neither the original nor our experimental setting uses the parent classes in the evaluation stage.

4.2. Comparative methods

In addition to prototypical networks, we also use a selection of few-shot algorithms for comparison [13, 14]. The transfer learning method in [13] trains an encoder from scratch using the base split. It then applies this trained model with the best validation performance for few-shot evaluation using the novel split. Another work in [14] used matching networks to calculate an attention matrix between query samples and support samples. The attention matrix is multiplied by matrix of support labels to obtain the logits of the query ones.

Except matching networks, we apply the identical encoder as described in Table 1. For the matching network we build a model following the implementation in [14]. As all of the comparative methods are metric-based, we apply the Euclidean distance to assess the similarity between two samples.

4.3. Evaluation metrics

As with Sect.6.5 in [24], this work uses *accuracy* to measure the ability of a classifier to make the correct decisions. The F_1 score is calculated to trade off between the ability to make correct decisions and to retrieve positive samples.

4.4. Experiment setup

Inputs for the few-shot methods are log Mel spectrograms. We set the sampling rate to 44100 Hz. The window length is 1024 sample points (roughly 20ms) with 50% overlap, and the number of Mel bank filters is 64. We finally get spectrograms sized 431×64. Before forwarding the extracted feature into network, frequency normalisation is applied along each Mel bin.

[2]Dataset available in https://github.com/karolpiczak/ESC-50

Following [22], in addition to a typical 5-way 5-shot problem, all few-shot learning methods were evaluated in a 12-way classification as well. For both experiments, each task (or episode) contains 5 support samples and 5 query samples. i.e. $K = 5$ and $Q = 5$. There are 32 episodes for each epoch. For a fair comparison, all models are trained through 100 epochs using the Adam optimiser with learning rate equal to 0.0001. 5-fold cross-validation was used throughout experiments. We randomly split the label set into two non-overlapping datasets, train and evaluation sets, as per the ratio 7:3. Following [22] we ensure ratios of children classes under same parent classes are identical between the train and evaluation process. We refer to this data split method as *uniform split* hereafter. It should be noted that the split label sets are changed with respect to different folds in the cross-validation, but all methods in the same folds share the same sets of train and evaluation.

4.5. Experimental results

Table 3: The performance of 12-way 5-shot learning methods on the ESC-50 dataset. The best one is highlighted in **bold**.

	$Accuracy$	F_1
Transfer Learning [13]	72.90%	72.87%
Proto [16]	77.70%	77.52%
Matching [14]	71.81%	71.75%
HPN (ours)	**78.65%**	**78.51%**

Table 4: The performance of three episodic learning methods for 5-way 5-shot problems on the ESC-50 dataset. The best one is highlighted in **bold**.

	$Accuracy$	F_1
Proto [16]	88.18%	88.18%
Matching [14]	86.83%	86.83%
HPN (ours)	**88.90%**	**88.88%**

Table 3 compares four few-shot learning methods in terms of accuracy and F1-score. Our hierarchical prototypical network yields the best performance among the transfer learning method, prototypical network (Proto), and matching network. The hierarchical prototypical network outperforms the best baseline (i.e. the prototypical network) by 0.95% and 0.99% in terms of accuracy and F1-score, respectively. This indicates that audio taxonomy knowledge can help an encoder to learn a better embedding space. Table 4 compares three episodic learning methods for 5-way 5-shot problems. Our HPNs can still yield a superior performance than prototypical networks and matching networks.

4.6. Impact of data split

In order to investigate the impact of data splits on overall performance, we compare different data split methods using the same few-shot model. Table 5 shows results of baseline prototypical networks with three data split methods. The *random split* method herein is to select 15 classes as novel classes randomly, so that the ratios of children classes under the same parent ones are not fixed. The *parent split* method is to select all child classes under the same parent category and fill this selected class set with the classes belonging to the rest parent categories. In this way, we make the number of classes in the selected set obtained by parent split method identical to the one by other methods. The descending order of the performance of three data split methods is "Uniform" > "Random" > "Parent". This observation adheres to our intuition that the performance on evaluation drops as the gap between the distribution of a base split and a novel split gets bigger. It also demonstrates that the characteristics between child classes from the same ancestor class are more similar than those from different classes, suggesting the importance of audio taxonomy knowledge in classification tasks.

Table 5: The performance of prototypical networks with different data splits on the ESC-50 dataset.

	acc	F_1
Random	74.59%	74.59%
Parent	73.35%	73.35%
Uniform	77.70%	77.52%

5. CONCLUSION AND FUTURE WORK

This work designs a hierarchical prototypical network for everyday sound recognition. The network extracts acoustic features and generates prototypical nodes corresponding to multiple levels in the tree structure. Distances between query samples and prototypical nodes in each level are then calculated and used for classification separately. Compared with prototypical networks with no hierarchy information, our model achieved a better performance in terms of accuracy and F_1-score.

Although hierarchical prototypical networks proves that it is promising to incorporate the audio taxonomy knowledge in few-shot everyday sound recognition, it still suffers from some limitations. First, HPNs cannot be applied to some datasets where an explicit taxonomy is not available. Second, some taxonomies are too complex to assign a hierarchical level to each label (e.g., a label having multiple paths to the root). In the future we plan to extend HPNs to a more complicated large-scale dataset. In addition, it is also intriguing to explore knowledge-based methods without an explicit taxonomy.

6. ACKNOWLEDGEMENT

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/T518086/1]. The research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT, http://doi.org/10.5281/zenodo.438045.

7. REFERENCES

[1] X. Fan, T. Sun, W. Chen, and Q. Fan, "Deep neural network based environment sound classification and its implementation on hearing aid app," *Measurement*, vol. 159, p. 107790, July 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0263224120303286

[2] T. Spadini, D. L. d. O. Silva, and R. Suyama, "Sound Event Recognition in a Smart City Surveillance Context," *arXiv:1910.12369 [cs, eess, stat]*, Feb. 2020, arXiv:

1910.12369. [Online]. Available: http://arxiv.org/abs/1910.12369

[3] B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng, and X. Li, "COVID-19 and Computer Audition: An Overview on What Speech & SoundAnalysis Could Contribute in theSARS-CoV-2 Corona Crisis," *Frontiers in digital health*, vol. 3, p. 14, 2021, publisher: Frontiers.

[4] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proc. Interspeech 2021*, 2021, pp. 571–575.

[5] T. Zhang, J. Liang, and B. Ding, "Acoustic scene classification using deep CNN with fine-resolution feature," *Expert Systems with Applications*, vol. 143, p. 113067, Apr. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417419307845

[6] 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: IEEE, Mar. 2017, pp. 776–780. [Online]. Available: http://ieeexplore.ieee.org/document/7952261/

[7] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022, conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[8] B. Kim, S. Yang, J. Kim, and S. Chang, "QTI Submission to DCASE 2021: Residual Normalization for Device-Imbalanced Acoustic Scene Classification with Efficient Design," DCASE2021 Challenge, Tech. Rep., June 2021.

[9] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-Scale Weakly Supervised Audio Classification Using Gated Convolutional Neural Network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB: IEEE, Apr. 2018, pp. 121–125. [Online]. Available: https://ieeexplore.ieee.org/document/8461975/

[10] T. Iqbal, Y. Cao, A. Bailey, M. D. Plumbley, and W. Wang, "ARCA23K: An Audio Dataset for Investigating Open-Set Label Noise," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, Nov. 2021, pp. 201–205.

[11] X. Li, Z. Sun, J.-H. Xue, and Z. Ma, "A concise review of recent few-shot meta-learning methods," *Neurocomputing*, vol. 456, pp. 463–468, Oct. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231220316222

[12] K. J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. Brisbane Australia: ACM, Oct. 2015, pp. 1015–1018. [Online]. Available: https://dl.acm.org/doi/10.1145/2733373.2806390

[13] Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, and P. Isola, "Rethinking Few-Shot Image Classification: A Good Embedding is All You Need?" in *Computer Vision – ECCV 2020*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 266–282.

[14] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, "Matching Networks for One Shot Learning," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/hash/90e1357833654983612fb05e3ec9148c-Abstract.html

[15] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[16] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-shot Learning," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4077–4087, 2017.

[17] 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, pp. 76–80, iSSN: 2379-190X.

[18] 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)*. IEEE, 2021, pp. 36–40.

[19] C. Heggan, S. Budgett, T. Hospedales, and M. Yaghoobi, "MetaAudio: A Few-Shot Audio Classification Benchmark," *arXiv:2204.02121 [cs, eess]*, Apr. 2022, arXiv: 2204.02121. [Online]. Available: http://arxiv.org/abs/2204.02121

[20] H. Phan, L. Hertel, M. Maass, P. Koch, R. Mazur, and A. Mertins, "Improved Audio Scene Classification Based on Label-Tree Embeddings and Convolutional Neural Networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1278–1290, June 2017, conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[21] Z. Peng, Z. Li, J. Zhang, Y. Li, G.-J. Qi, and J. Tang, "Few-Shot Image Recognition With Knowledge Transfer," 2019, pp. 441–449. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Peng_Few-Shot_Image_Recognition_With_Knowledge_Transfer_ICCV_2019_paper.html

[22] H. F. Garcia, A. Aguilar, E. Manilow, and B. Pardo, "Leveraging Hierarchical Structures for Few-Shot Musical Instrument Recognition," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021, Online, November 7-12, 2021*, J. H. Lee 0001, A. Lerch 0001, Z. Duan, J. Nam, P. Rao, P. v. Kranenburg, and A. Srinivasamurthy, Eds., 2021, pp. 220–228. [Online]. Available: https://archives.ismir.net/ismir2021/paper/000027.pdf

[23] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations*, 2015.

[24] T. Virtanen, M. D. Plumbley, and D. Ellis, Eds., *Computational Analysis of Sound Scenes and Events*. Cham: Springer International Publishing, 2018. [Online]. Available: http://link.springer.com/10.1007/978-3-319-63450-0