INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

A multimodal prototypical approach for unsupervised sound classification

Saksham Singh Kushwaha[1,2], *Magdalena Fuentes*[2,3]

[1]Courant Institute of Mathematical Sciences, New York University, NY, USA
[2]MARL, New York University, NY, USA
[3]IDM, New York University, NY, USA
sk8974@nyu.edu, mf3734@nyu.edu

Abstract

In the context of environmental sound classification, the adaptability of systems is key: which sound classes are interesting depends on the context and the user's needs. Recent advances in text-to-audio retrieval allow for zero-shot audio classification, but performance compared to supervised models remains limited. This work proposes a multimodal prototypical approach that exploits local audio-text embeddings to provide more relevant answers to audio queries, augmenting the adaptability of sound detection in the wild. We do this by first using text to query a nearby community of audio embeddings that best characterize each query sound, and select the group's centroids as our prototypes. Second, we compare unseen audio to these prototypes for classification. We perform multiple ablation studies to understand the impact of the embedding models and prompts. Our unsupervised approach improves upon the zero-shot state-of-the-art in three sound recognition benchmarks by an average of 12%.

Index Terms: zero-shot prototypical learning, text-to-audio retrieval, environmental sound classification, sound recognition.

1. Introduction

Environmental sound event classification has several applications of interest to public health and industry such as assistive devices [1], autonomous navigation [2], home assistants [3], noise mitigation [4], among others. Typical sound recognition systems consist of deep-learning-based supervised models, where human annotations are needed to train a model to recognize sounds from a predefined set of classes. The main disadvantage of such models in practice is that they are very inflexible to work with out-of-domain sounds. Recent work has highlighted the importance of adaptability of sound recognition systems in the context of assistive devices [1], but this also holds in general: the vocabulary of sounds that such a system should recognize will vary from home to home, city to city, and application to application. It is troublesome to re-train such models each time.

Many efforts have been made in making sound recognition models more adaptable, e.g. using few-shot learning [5, 6] where only a few curated examples from each sound class are needed to train a competent model for environmental sound recognition and music classification [5]. In the context of assistive devices, prototypical approaches have shown promise [1], also needing a few inputs from the user to select audios from a database or record them themselves. Although both approaches propose a useful and flexible change of paradigm with respect to previous supervised approaches, there is still the need for human supervision, and in cases where several sound classes want to be recognized, the time spent curating or selecting audio examples can be considerable.

Recently, with the introduction of multimodal deep learning text and audio self-supervised models, the prospect of successfully doing zero-shot classification (classifying instances of unseen data without any training or fine-tuning) has improved considerably [7, 8]. This opens the possibility for domain adaptation of sound recognition systems by exploiting the correspondence of text and audio without any human intervention.

In this work, we propose an unsupervised multimodal prototypical approach that leverages zero-shot text-to-audio retrieval capabilities of large multimodal models. To do so, unlike previous approaches, we use text embeddings to find representative audio clusters in the joint audio-text embedding space without any human supervision and compute the cluster's centroid as the prototype. At classification time, we use these audio prototypes to compare the unseen audio query and classify it. Our approach improves upon the zero-shot state-of-the-art in three well-known environmental sound classification benchmarks, namely ESC-50, UrbanSound8K, and FSD50k, and performs competitively to supervised approaches in a challenging multi-class scenario. Our contributions are as follows: 1) we propose an unsupervised multimodal strategy to select audio prototypes using text for sound classification; 2) we evaluate the effectiveness of this approach using different datasets (single-label and multi-label) and different pre-trained text-audio models; 3) and we investigate the impact of prompting as well as cluster's size in the accuracy of our approach. Our code is open-source and available for research. [1]

2. Related work

Supervised models. Supervised models for environmental sound classification have recently shifted to rely heavily on transfer learning, the most popular approach being to pre-train audio-visual deep learning models using self-supervision [9, 10, 11] on large amounts of data so it learns meaningful features from audio and images, and then using its audio encoder as input to a shallow classifier to work on new unseen audio data (see Figure 1a). This is typically done by exploiting the semantically related information between the two modalities, typically audio and image, to algorithmically generate labels for large amounts of data and pre-train a large model without any human supervision. The fact that labels are generated automatically allows for exposing these self-supervised models to large amounts of data which would be impossible otherwise, and thus the superiority of these models with respect to supervised ones trained on small datasets. This transfer learning approach has proved to be effective in environmental sound

[1] https://github.com/sakshamsingh1/audio_text_proto

10.21437/Interspeech.2023-982

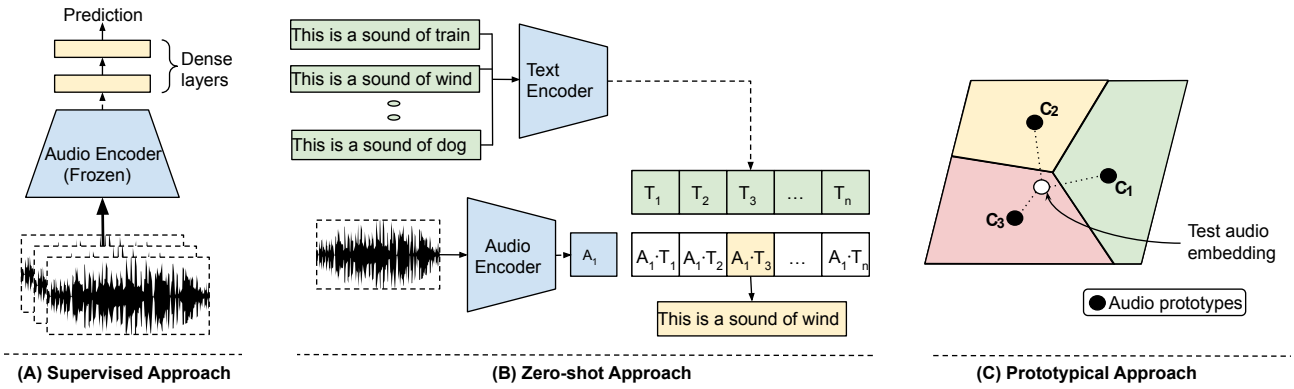

Figure 1: *Typical classification approaches in the literature of environmental sound classification.*

classification [9, 12], domestic sound classification [13], among others. However, still relies on annotated labels for the fine-tuning stage, so the human intervention bottleneck remains. We include supervised models that leverage transfer learning from audio and text in our experiments for comparison to our prototypical approach, as explained in Section 3.

Text-audio deep learning models. After the introduction and open release of CLIP [14], a large text-image multimodal deep learning model which showed impressive results for text-image retrieval, zero-shot classification, image captioning[15], and text-to-image generation [16], many approaches have been proposed to create models that have equivalent text-audio capabilities. Some of those approaches directly build on CLIP, e.g. by using its frozen image encoder to guide the training of an audio encoder that would learn embeddings in the pre-existing text-image joint embedding space [17]. Other approaches fine-tune the text and image encoders along with an audio encoder in datasets that contain text, audio and image samples [7]. Other approaches train audio and text encoders from scratch using contrastive loss from audio and text pairs [8], and even explore text-augmentation techniques to make the model more flexible to natural language inputs [18]. As mentioned before, these text-audio models have shown great potential for zero-shot classification in new, unseen datasets, which is achieved by embedding audio samples and text labels into the same space and computing the similarity between them (see Figure 1b). The biggest advantage of this approach is that it is completely unsupervised, but its main disadvantages are that it is sensitive to the "quality" of the prompt and its performance is still considerably lower than supervised approaches. In this work, we leverage the potential of these models for zero-shot classification within a different approach: prototypical classification. For this, we explore the effectiveness of different pre-computed text-audio embeddings, in particular [7, 18], as they are the state-of-the-art in zero-shot environmental sound classification.

Prototypical approaches. Prototypical approaches have been successfully used in the context of computer vision [19] and sound recognition [1]. These approaches typically consist of a first stage of selecting a small set of examples that are characteristic of a class, and then obtain the centroid of each class group as the prototype. Either if the embeddings are pre-trained [1] or learned in the process [19] of computing clusters and centroids, similarly to few-shot learning, prototypical approaches typically rely on few annotated data (the examples). We propose to select the examples without any human intervention, by leveraging the text-to-audio zero-shot capabilities of multimodal deep learning models. We do this by converting both text and audio into embeddings and using the proximity of those embeddings to query audio using text. What this means in practice is that the user would input a text prompt as query, and the model would internally retrieve a relevant audio prototype to represent the user query or label, without the need of further recording or choosing between examples on the users' side. Our method is explained in detail below.

3. Method

3.1. Datasets and metrics

We use three audio classification datasets for our experiments that differ in size, number and type of labels. These datasets are described below.

ESC-50[20]: The ESC-50 dataset compromises of 2000 environmental audio recordings, with each clip of 5 seconds. The audio clips belong to 50 class labels that can be divided into 5 major categories such as animals and urban noises. The dataset is divided into 5 non-overlapping folds by the authors for cross-validation. The models are evaluated using 5-fold multiclass classification accuracy.

UrbanSound8K(US8K)[21]: This dataset consists of 8732 recordings (each track $\leq$ 4s) which belong to 10 categories (eg. car horn, children playing). Similar to ESC-50, this dataset is also divided into 10 non-overlapping folds and is evaluated using 10-fold multiclass classification accuracy.

FSD50K[22]: This dataset consists of 51,197 Freesound[23] that span over 200 classes. The clips have varying lengths ranging from 0.3s to 30s and are organized hierarchically (144 leaf nodes and 56 intermediate nodes) with a subset of the AudioSet Ontology. The dataset is a multi-label dataset and has been divided into train, validation, and test split. To evaluate the performance of models trained on this dataset, the mean average precision(mAP) metric has been adopted.

3.2. Multimodal prototypical approach

Our approach is illustrated in Figure 2. It consists of two main steps: 1) retrieving audio prototypes from text, and 2) using these prototypes for classification. Additionally, we explore the impact of prototype selection, as explained below.

Prompt selection. The performance of text-audio models for zero-shot classification is sensitive to the particular text prompts

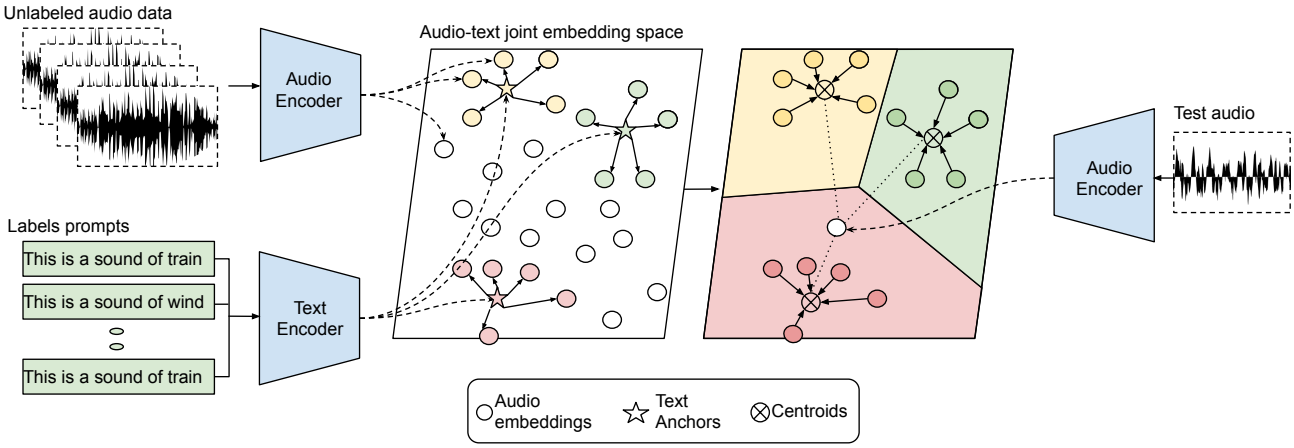

Figure 2: *Our unsupervised multimodal prototypical approach. We use text queries as anchors to retrieve clusters of audio embeddings that best represent the text. The centroids of these clusters serve as prototypes. During inference, the new audio is compared to the prototypes and assigned the corresponding text label.*

used to query [8], given the data that was used to train them [18]. To analyze that and mitigate its impact in our study, we use different formulations of prompts that include the labels of each dataset and compare the performance of the different models for the ESC-50 dataset. Based on the accuracy performance of each configuration, we select the best prompt for each model and use it for the remaining experiments. In practice, this can be done in an annotated dataset different from the target data.

Embedding models. Our prototypical approach leverages pre-trained audio and text encoders from two state-of-the-art multimodal models, namely AudioClip [7] and LAION-CLAP [18]. Specifically, we refer to our approach based on AudioClip as Proto-AC. Additionally, our approach that employs the encoders of LAION-CLAP with keyword-to-caption and feature fusion is referred as Proto-LC.

Unsupervised selection of audio prototypes using text. As depicted in Figure 2, our approach uses text queries (represented by labels prompts from each dataset) as anchors in the joint audio-text embedding space to retrieve local neighborhoods (clusters) of audio embedding that better represent the text query. We use k-nearest neighbors for this, where k was chosen via a grid search for ESC-50 but kept the same for the rest of the datasets. The reasoning behind this is that we want to strike a reasonable number for k but we do not want to tune it for each dataset since this would require using the labels in practice. The centroids of those clusters become the prototypes for the different sound classes of interest. As stated before, our interest in exploiting the retrieval capabilities of these multimodal models is to have a completely unsupervised approach, that intuitively will have better audio-text matches than zero-shot learning given that the matching process is done considering multiple examples (the clusters) instead of one, and similarities are measured between embeddings of the same modality to make the final assignments.

Classifying unseen audio samples. Given a new audio sample to be classified, we first extract its embedding using the same audio encoder that is used for computing the clusters and prototypes. Then we proceed in two different ways depending if the dataset is single-label or multi-label. For single-label datasets like ESC-50 and US8K, we choose the predicted label to be the one whose prototype is closest (via cosine similarity) to the embedding of the unseen audio. In the case of multi-label datasets like FSD50K, we compute a vector with the different classes' likelihoods by computing the sigmoid of the cosine-similarity between the embedding of the unlabeled audio and the embedding of each of the prototypes. We then calculate the mAP with multi-label targets as one-hot vectors.

3.3. Comparison to other approaches

Zero-shot classification. We use the same pre-trained embedding models as our prototypical approach to do zero-shot classification as a baseline. To do so, we first compute the embeddings for all the test audio and prompted text labels using the pretrained encoders. Because text and audio share the same embedding space we compute cosine similarity between their embeddings. We then use softmax and sigmoid over this distribution for single-label and multi-label classification respectively.

Supervised classification with pre-trained embeddings. Using the same embedding models as before (LAION-CLAP and AudioCLIP), we follow the supervised approaches explained in Section 2, pre-compute the audio embeddings and use them as input to a shallow classifier. The classifier consists of 3 fully connected layers with relu activations in all of them except the last. For the last layer, in the case of single-label classification we use a softmax activation, and for multi-label classification we use a sigmoid activation. We train this network using Adam with learning rate of 1e-4, $\beta_1 = 0.9$ and $\beta_2 = 0.999$.

Supervised prototypical networks. To understand how good are the embeddings to characterize each sound class within each dataset, we include two baselines (one with LAION-CLAP and another with AudioClip) in which we select the audio clusters using their labels directly. We then compute the centroids as prototypes and perform inference exactly as the prototypical approaches explained before.

Our results are presented in Table 1. We group results as *zero-shot* when methods do not use any label, and *supervised* when the labels are used as explained in Section 3.3. A first observation is that the prototypical approach performs better in most cases, for all datasets, with the best configuration being Proto-LC. In the following, we break down the discussion in the different aspects of this study.

	ESC-50 (acc)		US8k (acc)		FSD50K (mAP)	
	Zero Shot	Supervised	Zero Shot	Supervised	Zero Shot	Supervised
Wav2Clip	0.41	0.86	0.40	0.81	0.03	0.43
AudioClip[*,†]	0.68	0.88	0.62	0.86	0.20	0.50
CLAP	0.83	**0.97**	**0.73**	0.88	0.30	0.59
LAION-CLAP[*,†]	0.91	0.96	0.72	**0.89**	0.22	0.61
Proto-AC[†]	0.78	0.82	0.71	0.77	0.40	0.48
Proto-LC[†]	**0.96**	**0.97**	**0.73**	0.83	**0.52**	**0.65**

Table 1: *Classification results for the different approaches and configurations.* [*]*Results reproduced using author's code* [†]*Best label prompts were selected based on the performance on ESC-50*

4. Results and discussion

What is the effect of the prompt? We analyze the effect of the prompt by trying different prompting variations as text anchors, as shown in Table 2. We examine the performance of the embedding models as well as the prototypical models in zero-shot in the ESC-50 dataset. We selected 5 prompts that show promise in previous works [18]. A first observation is that the different embedding models have different robustness to changing the prompts. AudioClip's performance variation is relatively small (2% maximum), while LAION-CLAP's variation is larger with up to 9% difference in performance. Surprisingly, that trend does not transfer to the prototypical models that use AudioClip and LAION-CLAP respectively: the variation for both Proto-AC and Proto-LC is relatively large (8% and 4% respectively), with AudioClip having the least variability.

	AudioClip	LAION-CLAP	Proto-AC	Proto-LC
Prompt 1	0.67	0.83	0.77	0.94
Prompt 2	0.68	0.86	0.72	0.92
Prompt 3	**0.69**	0.90	0.72	**0.96**
Prompt 4	0.68	0.88	**0.78**	0.95
Prompt 5	0.67	**0.92**	0.70	**0.96**

Table 2: *Accuracy on ESC-50 with different prompts. Prompt 1: '{Class label}', Prompt 2: 'I can hear {class label}', Prompt 3: 'This is an audio of {class label}', Prompt 4: 'This is {class label}', Prompt 5: 'This is a sound of {class label}'.*

Another surprising finding not included in Table 2 was that the models' performance changed if the prompts started with an uppercase or lowercase letter, where the lowercase configuration performed considerably worse (an average of 5%). This shows how sensitive this models are to prompting, and indicates that further augmentations beyond rephrasing are needed during training to ensure robustness in practice. Based on the results in Table 2 we chose the best prompt for each model according to this ablation study for the remaining experiments.

The impact of the number of neighbours. In order to obtain the clusters of audio examples, we need to choose how many examples we consider in each cluster. Intuitively, this hyper-parameter (k) will impact the quality of the prototypes, since too few or too many examples could lead to an inaccurate prototype choice. The selection of such a hyper-parameter is the only stage in the prototypical approach that would require the use of labels, if we want to compute the performance of the model at different values of k and choose the optimal one. As explained in Section 3.2, we perform a preliminary study in ESC-50 to understand the impact of such parameter, strike a reasonable value for k and keep the same value to the other datasets to simulate how the method would be used in practice without this information. Figure 3 shows this for the ESC-50 dataset. As shown there, the performance of the model is equally high for a large set of values of k. We choose $k = 35$, and *we keep this value for the rest of the datasets*. As shown in Table 1, despite not optimizing k for US8K or FSD50K, the prototypical approach outperforms the other methods, showing promise in robustness and low parameter tuning.

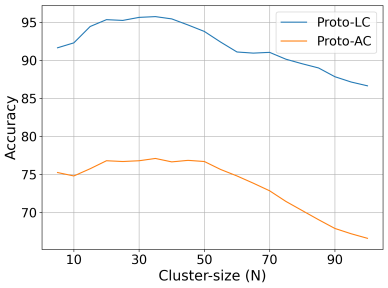

Figure 3: *Cluster size vs accuracy of prototypical methods.*

Performance in single-label vs. multi-label datasets. Table 1 shows that the prototypical approach is significantly better than the zero-shot baseline, especially in the FSD50K dataset compared to US8K or ESC-50. This implies that Proto-AC/Proto-LC are more effective in handling complex relationships in multi-label datasets than the zero-shot method. The reason for this may be that computing the prototypes as centroids of nearby audios leads to a better alignment of the prototypes with the audio embeddings than the text embeddings, which we will further investigate in future work.

5. Conclusions

Our work proposes using text-audio multimodal deep learning model capabilities to classify environmental sounds using prototypical classification, without the need for human intervention. Our method performs better than zero-shot classification and can enable user-adaptable sound recognition systems through text. For future research, we will investigate training encoders to be more robust to prompt changes, as well as compare the computational complexity of different approaches.

6. References

[1] D. Jain, K. Huynh Anh Nguyen, S. M. Goodman, R. Grossman-Kahn, H. Ngo, A. Kusupati, R. Du, A. Olwal, L. Findlater, and J. E. Froehlich, "Protosound: A personalized and scalable sound recognition system for deaf and hard-of-hearing users," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–16.

[2] Y. Furletov, V. Willert, and J. Adamy, "Auditory scene understanding for autonomous driving," in *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2021, pp. 697–702.

[3] R. Haeb-Umbach, S. Watanabe, T. Nakatani, M. Bacchiani, B. Hoffmeister, M. L. Seltzer, H. Zen, and M. Souden, "Speech processing for digital home assistants: Combining signal processing with deep-learning techniques," *IEEE Signal processing magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[4] J. P. Bello, C. Silva, O. Nov, R. L. Dubois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "Sonyc: A system for monitoring, analyzing, and mitigating urban noise pollution," *Communications of the ACM*, vol. 62, no. 2, pp. 68–77, 2019.

[5] Y. Wang, J. Salamon, N. J. Bryan, and J. P. Bello, "Few-shot sound event detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 81–85.

[6] Y. Wang and D. V. Anderson, "Hybrid attention-based prototypical networks for few-shot sound classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 651–655.

[7] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[8] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," *arXiv preprint arXiv:2206.04769*, 2022.

[9] 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)*. IEEE, 2019, pp. 3852–3856.

[10] M. Cartwright, J. Cramer, J. Salamon, and J. P. Bello, "Tricycle: Audio representation learning from sensor network data using self-supervision," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 278–282.

[11] R. Arandjelovic and A. Zisserman, "Objects that sound," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[12] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[13] A. Politis, A. Mesaros, S. Adavanne, T. Heittola, and T. Virtanen, "Overview and evaluation of sound event localization and detection in dcase 2019," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 684–698, 2020.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[15] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[16] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[17] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4563–4567.

[18] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," *arXiv preprint arXiv:2211.06687*, 2022.

[19] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

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

[21] 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.

[22] 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, 2021.

[23] E. Fonseca, J. Pons Puig, X. Favory, F. Font Corbera, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: a platform for the creation of open audio datasets," in *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.