On the Role of Training Class Distribution in Zero-Shot Audio Classification

Duygu Dogan, Huang Xie, Toni Heittola, Tuomas Virtanen

Signal Processing Research Centre, Tampere University, Finland

***Abstract*—Zero-shot learning (ZSL) enables the classification of audio samples into classes that are not seen during training by transferring semantic information learned from seen classes to unseen ones. Thus, the ability of zero-shot models to generalize to unseen classes is inherently affected by the training data. While most audio ZSL studies focus on improving model architectures, the effect of training class distribution in the audio embedding space has not been well explored. In this work, we investigate how the distribution of training classes in audio embedding space, both internally and in relation to unseen classes, affects zero-shot classification performance. We design two controlled experimental setups to understand the impact of training classes: (i) a similarity-based configuration, where we experiment with varying acoustic similarity between training classes and unseen test classes, and (ii) a diversity-based configuration, where the training sets are constructed with different levels of coverage in the audio embedding space. We conduct our experiments on a subset of AudioSet, evaluating zero-shot classification performance under different training class configurations. Our experiments demonstrate that both higher acoustic similarity between training and test classes and higher acoustic diversity among training classes improve zero-shot classification accuracy.**

***Index Terms*—zero-shot learning, audio classification**

1. INTRODUCTION

Zero-shot learning (ZSL) refers to the ability of a model to recognize classes that were not introduced during training by utilizing semantic information such as textual class descriptions or attribute embeddings [1]. By transferring knowledge from seen to unseen classes, zero-shot models can perform classification tasks without requiring labeled examples from every class. This ability is especially useful where data collection is difficult, such as for rare classes.

In the audio domain, ZSL is particularly important as the temporal nature of audio makes data annotation expensive. Instead of requiring labeled examples for every target class, zero-shot learning enables classification of unseen classes by learning to map audio and semantic embeddings into a shared space using only seen classes. However, it makes two strong assumptions: (i) audio and semantic embeddings of the same class can be mapped to nearby points in the shared space, and (ii) the shared space learned from seen classes is structured to generalize to unseen class embeddings. In practice, domain mismatch between audio and text embeddings often limits this alignment, and learned projection functions may fail to generalize to the regions of the shared space where unseen classes fall into, leading to poor zero-shot performance.

Most prior work in audio ZSL has focused on improving the model architectures and modality alignment through objective functions. Perhaps the most influential models in ZSL like CLAP [2], [3] and AudioCLIP [4] demonstrate the power of contrastive learning for audio–text embedding. More recent studies utilize large language models (LLMs) to enhance class representations by generating rich textual descriptions or attribute-based prompts to improve the alignment between audio and semantic embeddings in zero-shot settings [5]–[7].

In computer vision, recent works have studied the topological structure of the embedding space and have shown that preserving or modeling relationship between classes can improve ZSL performance, for example, by preserving the class topology in the embedding space in [8], aligning semantic and visual graphs in [9], [10], and modeling the global structure for compositional ZSL models in [11], [12]. However, these approaches are mainly limited to the vision domain. In the audio domain, while projection quality has been widely explored, the impact of training class structure in ZSL is underexplored.

Furthermore, recent studies have emphasized the importance of the distribution of training classes in zero-shot learning. They show that class imbalance or distribution shifts can harm generalization, and propose solutions such as synthetic data generation [13] or out-of-distribution detection [14]. Motivated by these, we investigate the effect of the distribution of training data in the audio embedding space on zero-shot performance.

In this work, we study how the relative position of training classes, both with respect to each other and to those of test classes, in the audio embedding space affects zero-shot classification performance. Our hypothesis is that having training classes that are acoustically more similar to test classes can support better alignment, while a diverse set of training classes that spans a larger region of the space may provide better coverage and improve generalization compared to training classes clustered in a narrow region.

We address two questions: (i) Does greater acoustic similarity between training and test classes improve zero-shot performance? (ii) Does increasing diversity among training classes improve generalization? To answer these questions, we design two controlled experimental setups:

- Similarity-based analysis, where we investigate how the acoustic proximity between training and test classes affects performance by constructing training sets with varying levels of similarity to the test classes.
- Diversity-based analysis, where we vary the internal pairwise similarity of training classes to examine the effect of training set coverage in the audio embedding space.

By evaluating performance across a wide range of configurations, we analyze how structural factors influence zero-shot accuracy. Our results show that both high train–test similarity and moderate training diversity are essential for robust zero-shot performance.

2. ZERO-SHOT AUDIO CLASSIFICATION FRAMEWORK

Figure 1 illustrates the pipeline of our zero-shot audio classification model, which learns to map audio clips to class descriptions only using seen training classes. Our approach is based on previous work on projection-based alignment between audio and semantic embeddings for zero-shot classification in [15], with an additional learnable semantic projection layer. Let X be the set of audio samples and C the set of class indices, split into seen C_{train} and unseen C_{test}, where $C_{\text{train}} \cap C_{\text{test}} = \varnothing$. The training set is defined as the labeled pairs

$$D_{\text{train}} = \{(x_n, y_n)\}_{n=1}^{N}, \quad x_n \in X,\ y_n \in C_{\text{train}}, \tag{1}$$

where N is the number of training samples, x_n is the n-th audio sample, and y_n is its corresponding class index. Each audio sample x

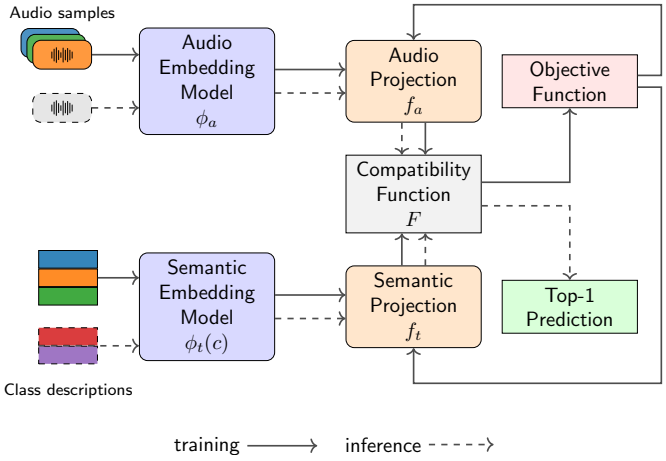


Fig. 1: Overview of the zero-shot audio classification framework. Audio and semantic embeddings are projected into a shared embedding space where their compatibility is computed. Projection layers are updated via an objective function in training. Inference selects the top-1 predicted class based on compatibility scores.

is encoded using an audio embedding function $\phi_a : X \to \mathbb{R}^{d_a}$, and each class description $c \in C$ is embedded using a language embedding model $\phi_t : C \to \mathbb{R}^{d_t}$, where d_a and d_t denote the output dimensions of the audio and language embedding models, respectively. These embeddings are then projected into a shared space using learnable functions

$$f_a : \mathbb{R}^{d_a} \to \mathbb{R}^D, \; f_t : \mathbb{R}^{d_t} \to \mathbb{R}^D, \tag{2}$$

where D is the dimension of the shared space. To measure how well an audio sample and a class description match, we compute their compatibility in the shared space using cosine similarity:

$$F(x, c) = \cos(f_a(\phi_a(x)), f_t(\phi_t(c))). \tag{3}$$

The objective is to bring audio embeddings closer to their corresponding semantic embeddings in the shared space. To achieve this, we use the Weighted Approximate Rank Pairwise (WARP) loss following [15]. The model is trained to assign a higher compatibility score $F(x_n, y_n)$ to correct pairs (x_n, y_n) than to incorrect ones.

At inference, for an unseen audio sample x, the model predicts the class with the highest compatibility score among the unseen classes $z \in C_{\text{test}}$:

$$\hat{z} = \arg\max_{z \in C_{\text{test}}} F(x, z). \tag{4}$$

In our implementation, we use pretrained and frozen VGGish [16] as the audio embedding model (d_a =128) and Sentence-BERT (SBERT) [17] as the semantic embedding model (d_t =768). The audio projection consists of two fully connected layers of sizes 256 and 512, with a tanh activation in between. The semantic projection layer is a single fully connected layer of size 512.

As the projection layers are learned only on the training pairs, the geometry of the shared embedding space is shaped by their structure. Unseen test classes may project into any region of this space, but may fall into misaligned regions to have meaningful representation for reliable prediction. This is illustrated in Figure 2, where unseen classes and test audio embeddings fall outside the training subspace. Understanding how these structures affect zero-shot classification is therefore important for designing more robust models.

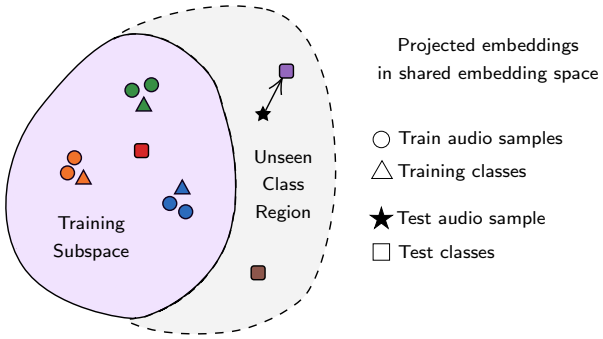


Fig. 2: Illustration of the shared embedding space. Purple area (left): Projected training and class embeddings form a learned training subspace. Gray area (right): Unseen test classes and test audio embeddings may project outside the training subspace. Test audio is predicted as the class whose embedding is the most similar.

3. TRAINING SET CONFIGURATIONS

To analyze the effect of training class distribution in the audio embedding space, we design two controlled experimental setups varying training class sets. We apply these analyses to audio embeddings (see Section 4.1 for details). For the purpose of analysis, we assume access to the full labeled dataset before defining training and test splits. To define the configurations, we first represent each class by the class-level audio embedding μ_c, defined as the centroid of the audio embeddings of all samples in a class c:

$$\mu_c = \frac{1}{N_c} \sum_{i \in I_c} \phi_a(x_i), \tag{5}$$

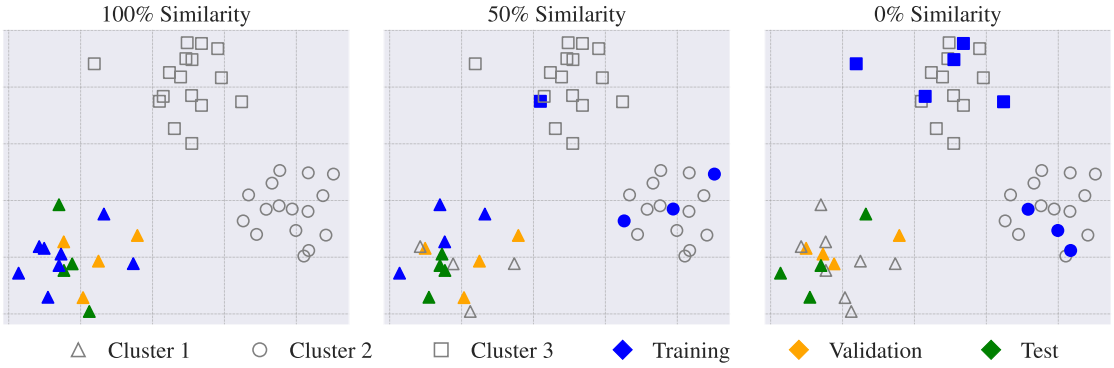
where $\phi_a(x_i)$ is the pretrained audio embedding of sample x_i, y_i refers to the label of sample x_i, $I_c = \{i \mid y_i = c\}$ is the set of indices of samples labeled as class c, and $N_c = |I_c|$.

Based on these class-level audio embeddings, we create training sets with controlled class distributions for two different analyses. In the similarity-based analysis, we construct train–test configurations where the training classes have different levels of acoustic similarity to the test classes. In the diversity-based analysis, we create training sets where the internal similarity of class-level audio embeddings among training classes are within different ranges, resulting in varying levels of coverage in the audio embedding space.

3.1. Similarity-based Analysis

To understand how the local alignment between seen and unseen classes in the audio embedding space affects zero-shot performance, we systematically vary the similarity ratio between training and test samples in the audio embedding space. We first group acoustically similar classes by clustering them. To perform clustering based on cosine similarity, we normalize all class-level audio embeddings to unit length $\hat{\mu}_c = \mu_c / \|\mu_c\|$ and then apply K-means (with $K = 3$) to the set $\{\hat{\mu}_c \mid c \in C\}$.

Each cluster is used once as the source of test and validation classes in a separate evaluation configuration. That is, in each configuration, we select a fixed number of test and validation classes exclusively from one cluster, and refer to that as the test cluster, while the remaining classes from all clusters are considered as training candidates. For each test cluster, we construct five different training sets by varying the proportion of training classes selected from the test cluster. We define the similarity ratio as the percentage of training classes that belong to the same cluster as the test set. Specifically, we use the



(a) Similarity-based configuration.

(b) Diversity-based configuration.

Fig. 3: Illustration of training set configurations with toy data in (a) similarity-based, (b) diversity-based. Each subplot shows one illustrative instance: a single test cluster in (a), and a fixed test/validation set in (b). In the actual experiments, we repeat this procedure for multiple configurations (see Section 4.2).

levels 1.0, 0.75, 0.50, 0.25, and 0.0. For example, a similarity ratio of 0.75 means that 75% of training classes are selected from the same cluster as the test set, and 25% from the remaining clusters.

We repeat this process for each cluster, each with five training sets of varying similarity levels, and report the Pearson correlation between similarity ratio and test accuracy. Figure 3a illustrates the effect of the similarity ratio on train and test sets on a toy data.

3.2. Diversity-based Analysis

To investigate whether diversity among training classes helps generalization, we define training sets with different diversity levels in the audio embedding space. We measure diversity by the average pairwise cosine similarity between class-level audio embeddings. Specifically, we define five different cosine similarity ranges as (0.5, 0.6], (0.6, 0.7], (0.7, 0.8], (0.8, 0.9], and (0.9, 1.0]. We create training sets with different diversities with an iterative algorithm as follows:

1) Randomly select three test sets and three validation sets. These sets remain fixed for every diversity range.
2) For each range
   - Initialize the training set with a randomly chosen class from the remaining set (excluding any test and validation set).
   - Iteratively add new classes whose average pairwise cosine similarity to the current training set falls within a target range (e.g., (0.5, 0.6]).
   - Repeat until the training set reaches a predefined number of classes for each target range.
3) Evaluate the zero-shot accuracy obtained using a model trained with that training set on each of the three fixed test sets.

The goal of this setup is to measure the effect of the training data diversity on performance when the unseen classes are fixed. We compare performance across ranges by measuring the correlation between the average pairwise cosine similarity of training classes and test accuracy. Figure 3b illustrates through toy examples with different diversity levels: high, moderate, and low diversity. Higher diversity means lower internal cosine similarity and high-diverse sets provide larger coverage in the audio embedding space.

4. EXPERIMENTS

This section presents the experimental setup, including the dataset and training procedure for both analyses.

4.1. Dataset

We conduct our experiments on a subset of AudioSet [18], focusing on single-label samples. The AudioSet ontology provides a tree-like class hierarchy. To have enough data per class, we construct the subset by traversing the AudioSet ontology upward and selecting the deepest node with more than 100 samples, or its parent if none qualify. With this strategy, we ensure to choose the most specific classes with sufficient data. Applying this algorithm results in 143 classes, each with at least 100 samples.

We extract audio embeddings using a pretrained VGGish model. Each 10-second audio clip is divided into 10 segments, where each segment is mapped to a 128-dimensional vector. These segment-level embeddings are averaged to produce a single 128-dimensional

clip-level audio embedding. These embeddings are used both for training and for constructing class-level audio representations (obtained by averaging over all samples of a class) used in class selection strategies. For semantic embeddings, we apply the Sentence-BERT (SBERT) model [17] to the sentence-level class descriptions provided by AudioSet, resulting in a 768-dimensional vector for each class.

4.2. Training Setup

In the similarity-based analysis, we define 15 configurations with 5 similarity ratios (1.0, 0.75, 0.50, 0.25, and 0.0) and 3 test clusters. Each configuration corresponds to a distinct pairing of a test cluster and a similarity ratio. In the diversity-based analysis, we also define 15 configurations with 5 cosine similarity ranges ((0.5, 0.6], (0.6, 0.7], (0.7, 0.8], (0.8, 0.9], and (0.9, 1.0]) and 3 test sets. We use multiple test clusters (similarity-based) and test sets (diversity-based) to reduce the impact of any single test configuration. Each configuration is repeated 6 times with different random seeds, resulting in different combinations of train, validation, and test classes with the same analysis conditions. This leads to a total of 90 models for each analysis.

We set the number of clusters in similarity-based analysis to $K = 3$, selected using the elbow method. In the diversity-based analysis, the lower bound of 0.5 is chosen since the class pairs with pairwise similarity below 0.5 are extremely rare to construct complete training configurations with the given audio embeddings. Each configuration has 20 training, 5 validation, and 5 test classes. All models are trained from scratch and optimized using stochastic gradient descent (SGD) with a learning rate of 0.001, batch size of 64, and for 100 epochs.

5. RESULTS AND ANALYSIS

We report our findings separately for the similarity-based and diversity-based setups. In both cases, we evaluate the performance using top-1 accuracy averaged across random seeds per configuration. We analyze the relationship between class structuring metrics (similarity ratio and pairwise cosine similarity) and test accuracy via Pearson correlation.

5.1. Similarity-based Results

To measure the impact of train–test similarity, we defined a similarity ratio as the proportion of training classes from the same cluster as the test set. Figure 4 shows that accuracy increases with higher similarity, rising from 0.25 at similarity ratio 0.0 to 0.39 at similarity ratio 1.0. The trend holds across all test clusters despite minor accuracy variations.

Similarity ratio and accuracy show a significant positive correlation with $r = 0.374$ and $p < 0.001$ as shown in Table 1, indicating that increasing the acoustic similarity between test and training classes improves the model's ability to generalize to test classes.

5.2. Diversity-based Results

In the diversity-based setup, we varied the acoustic diversity of training classes by controlling their average pairwise cosine similarity. Results in Figure 5 show that accuracy peaks in the (0.6, 0.7] similarity range, where training classes are moderately diverse, with a mean of 0.50. Accuracy drops at both low (0.29 at (0.9, 1.0]) and high diversity (0.43 at (0.5, 0.6]), suggesting that moderate diversity provides the best balance between coverage and coherence. In addition, the standard deviation in accuracy increases with lower diversity, indicating less stable generalization when the model is trained on narrowly distributed classes. For example, the standard deviation was 0.07 in the (0.5, 0.6] range, compared to 0.13 in the (0.9, 1.0] range.

The Pearson correlation between mean pairwise similarity and test accuracy confirms this relationship with $r = -0.434$, $p < 0.001$ as shown in Table 1, indicating that greater training data diversity (i.e., lower internal similarity) is beneficial for model performance.

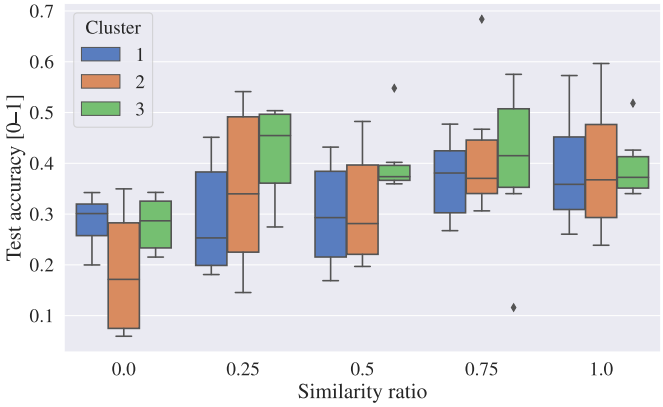


Fig. 4: Similarity-based results: Zero-shot classification accuracy across five train–test similarity ratios per cluster.

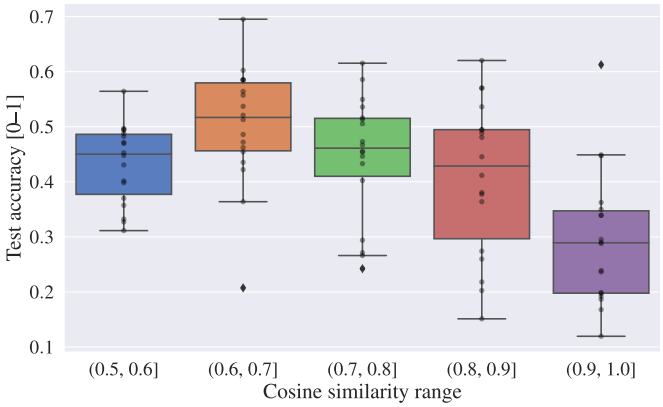


Fig. 5: Diversity-based results: Zero-shot classification accuracy grouped by the average pairwise cosine similarity between class-level audio embeddings.

Table 1: Pearson correlation coefficients between class structuring metrics and zero-shot accuracy, computed over 90 models per setup.

Strategy	Correlation (r)	p-value
Similarity-based	0.374	<0.001
Diversity-based	–0.434	<0.001

6. CONCLUSION

In this work, we investigated how the structure of training classes in the audio embedding space affects generalization in zero-shot audio classification. We proposed two experimental setups: one that changes the similarity ratio between training and test classes, and another that varies the diversity level among training classes. The similarity-based analysis demonstrates that increasing the acoustic similarity between training and test classes improves zero-shot performance. The diversity-based results show that when training classes are overly similar, the model may overfit and fail to capture patterns outside of the training class regions, whereas training sets that span a larger region of the space lead to better generalization. Based on these findings, future zero-shot audio classification models may benefit from explicitly leveraging structural information, such as graph-based or geometry-aware techniques to model the topologies of audio and semantic spaces.

REFERENCES

[1] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 9, pp. 2251–2265, 2019.

[2] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[3] B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 336–340, 2023.

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

[5] S. Ghosh, S. Kumar, C. K. R. Evuru, O. Nieto, R. Duraiswami, and D. Manocha, "Reclap: Improving zero shot audio classification by describing sounds," in *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[6] X. Xu, P. Zhang, M. Yan, J. Zhang, and M. Wu, "Enhancing zero-shot audio classification using sound attribute knowledge from large language models," in *Interspeech 2024*, 2024, pp. 4808–4812.

[7] N. Anand, A. Seth, R. Duraiswami, and D. Manocha, "Tspe: Task-specific prompt ensemble for improved zero-shot audio classification," in *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2025, pp. 1–5.

[8] H. Chen, Y. Liu, Y. Ma, N. Zheng, and X. Yu, "Tpr: Topology-preserving reservoirs for generalized zero-shot learning," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 102 229–102 254.

[9] Y. Hu, G. Wen, A. Chapman, P. Yang, M. Luo, Y. Xu, D. Dai, and W. Hall, "Graph-based visual-semantic entanglement network for zero-shot image recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 2473–2487, 2022.

[10] Z. Zhang and W. Cao, "Visual-semantic consistency matching network for generalized zero-shot learning," *Neurocomputing*, vol. 536, pp. 30–39, 2023.

[11] M. Naeem, Y. Xian, F. Tombari, and Z. Akata, "Learning graph embeddings for compositional zero-shot learning," in *34th IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[12] M. Mancini, M. Naeem, Y. Xian, and Z. Akata, "Open world compositional zero-shot learning," in *34th IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[13] Y. Slavutsky and Y. Benjamini, "Class distribution shifts in zero-shot learning: Learning robust representations," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, pp. 89 213–89 248.

[14] L. Wen, "Out-of-distribution detection for audio-visual generalized zero-shot learning: A general framework," in *Proceedings of the 35th British Machine Vision Conference (BMVC)*, 2024.

[15] H. Xie and T. Virtanen, "Zero-shot audio classification via semantic embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1233–1242, 2021.

[16] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135.

[17] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 11 2019.

[18] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.