Latent Multi-view Learning for Robust Environmental Sound Representations

Sivan Ding[1], *Julia Wilkins*[1], *Magdalena Fuentes*[1], *Juan Pablo Bello*[1]

[1] Music and Audio Research Laboratory, New York University, New York, NY, USA

***Abstract*—Self-supervised learning (SSL) approaches, such as contrastive and generative methods, have advanced environmental sound representation learning using unlabeled data. However, how these approaches can complement each other within a unified framework remains relatively underexplored. In this work, we propose a multi-view learning framework that integrates contrastive principles into a generative pipeline to capture sound source and device information. Our method encodes compressed audio latents into view-specific and view-common subspaces, guided by two self-supervised objectives: contrastive learning for targeted information flow between subspaces, and reconstruction for overall information preservation. We evaluate our method on an urban sound sensor network dataset for sound source and sensor classification, demonstrating improved downstream performance over traditional SSL techniques. Additionally, we investigate the model's potential to disentangle environmental sound attributes within the structured latent space under varied training configurations.**

***Index Terms*—Self-supervised Learning, Urban Sound, Environmental Sound Classification, Sensor Classification**

1. INTRODUCTION

Environmental sound representation learning is a crucial field in machine listening, serving as a foundation for tasks like environmental sound classification and acoustic scene analysis. To address the scarcity of annotations in environmental sound data, self-supervised learning (SSL) aims to approach and even sometimes surpass the performance of supervised methods through self-supervision techniques, such as contrastive and generative objectives.

Contrastive learning (CL) for audio representations exploits the assumption of commonalities across different views of data with the same semantics to create positive pairs, such as using augmented "views" of the same audio clip [1]–[3]. These principles have been widely employed for environmental sound [4]–[6]. Generative learning focuses on reconstruction-based objectives to learn an embedding space for a single "view" of complete or masked input [7]–[12]. In acoustic scene classification and sound event detection [13]–[16], these methods have proven effective at capturing a diverse acoustic content by encoding information across both temporal and feature dimensions. At the same time, in other domains, SSL research has begun to explore the complementary effects of combining reconstructive and contrastive learning frameworks in various ways. For example, [17] and [18] propose multi-view learning frameworks for visual clustering or classification. These methods simultaneously learn a "shared" latent subspace that is invariant across views and a "private" latent subspace that varies across views, by optimizing view reconstruction from both subspaces within an autoencoder architecture. This approach employs a view-contrastive strategy by learning both view-common and view-specific subspaces, without relying on explicit contrastive objectives. In [19], masked reconstruction is combined with CL objectives in an audio-visual context, but with one shared latent space for each modality view.

In the music domain, for pitch-timbre disentanglement, [20] leverages random perturbations to form view-contrasting training strategies within a reconstruction-based pipeline. Similarly, [21] studies how pitch-shifting can be used to create auxiliary objectives such as a contrastive loss in addition to single-view reconstruction. Furthermore, our previous work [22] uses a multi-view learning framework to learn disentangled latent subspaces for pitch and timbre without explicit contrastive objectives. These methods provide insights on how we can combine different SSL strategies and potentially disentangle inherent factors in audio. However, such methods haven't been explored in a more challenging context such as real-world environmental sound recordings, and there lacks a thorough investigation into how objective design choices affect latent subspace structures.

In this work, we combine contrastive and generative objectives within a multi-view learning framework to explore their combined impact on environmental sound representations. Using DAC [7] as a latent feature extractor, the framework encodes the audio feature into separate private and shared subspaces based on a metadata-driven data pairing strategy, eliminating the need for explicit class annotations. To strengthen the training signal of contrasting views in our multi-view backbone, we investigate similarity or separation-based contrastive objectives on the subspaces between views. Our evaluation is conducted on an urban sound sensor network dataset (SONYC-UST-V2 [23]), with performance measured on sound source and sensor classification tasks. In addition, we examine the model's ability to disentangle sound attributes, offering insights into the structure of the learned latent space. Our contributions are summarized as follows:

- We propose a novel latent multi-view contrastive learning framework for environmental sound representation learning.
- By simply creating pairs of data with available metadata and training a lightweight autoencoder in a self-supervised manner, we demonstrate a downstream performance boost on sound and sensor classification compared to traditional SSL baselines, achieving results comparable to supervised baselines.
- We investigate how different combinations of SSL training strategies influence the latent subspace structures, providing insights into environmental sound attribute disentanglement.

2. METHOD

2.1. Multi-view Representation Learning

Our proposed method is shown in Figure 1. The "views" of data used in our method are two audio clips, denoted $\{a_1, a_2\}$. Our system requires a single assumption between these audio clips: that they share a common attribute (e.g., recordings from the same sensor). We first pass a_1 and a_2 through a pretrained encoder to obtain 2D embeddings $\{x_1, x_2\}$ of dimension $(n \times d)$, where n is the number of frames and d is the feature dimension. In this work we use DAC [7] as the pretrained encoder. The DAC latent space, capable of preserving detailed information in environmental sound while reconstructing high-fidelity signals, provides us with a strong foundation for discriminative downstream tasks.

We pass embeddings x_1 and x_2 through a simple MLP encoder, denoted ϵ_ϕ. The weights of ϵ_ϕ are shared across views of data. This encoder projects each input embedding into two separate *private* and *shared* latent subspaces, as defined in traditional multi-view learning

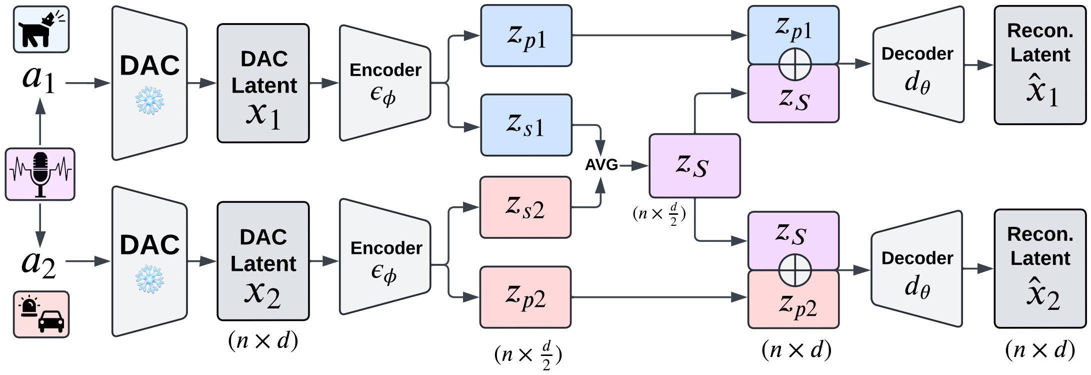

Fig. 1: Our self-supervised multi-view framework for environmental sound representation learning. Learned latent subspaces are used for downstream sound source and recording sensor classification tasks.

frameworks [18], denoted z_{pi} and z_{si} respectively, where i is the data view (for $i \in 1, 2$ in this study). The private subspace is designed to capture information specific to each individual view, while the shared subspace should capture information shared across both views. Latent subspaces z_{pi} and z_{si} of dimension $(n, \frac{d}{2})$ are then averaged to form a joint shared embedding: z_S. Additional training strategies can then be applied to these latent subspaces to incentivize desired information flow, which we introduce in Section 2.2. Lastly, an MLP decoder d_θ takes the concatenation of z_{pi} and z_S as input and projects back to the original dimensionality, reconstructing the original pretrained latents: $\{\hat{x_1}, \hat{x_2}\}$.

Importantly, note that because the weights of our projection encoder and decoder are shared across data streams, at inference time, our model can operate using only a single audio input; pairs are not needed. The model encodes the input audio into the previously learned private, shared, and combined subspaces, yielding z_p, z_s and their concatenation $(z_p \oplus z_s)$, reusable for further downstream tasks.

2.2. Self-supervised Training Strategies

To combine generative and contrastive SSL frameworks to learn robust environmental sound representations, we design a suite of training strategies to apply within the multi-view learning backbone.

Reconstruction: The base version of our model is trained using a mean squared error reconstruction loss $\mathcal{L}_{rec}$ between the original and reconstructed embeddings, x_i and $\hat{x}_i$ respectively, where i is the view index in the data pair and j is the sample in a dataset of size N samples:

$$\mathcal{L}_{rec} = \sum_i \left[\frac{1}{N} \sum_{j=1}^{N} (x_{i,j} - \hat{x}_{i,j})^2 \right] \tag{1}$$

Cosine Distance: In addition to the base reconstruction loss, we utilize contrastive learning in our framework via objectives based on cosine distance, in which similar embeddings are incentivized to be close together in the latent space and dissimilar embeddings are pushed farther apart [24]–[26]. The general form of this loss term is referred to as $\mathcal{L}_{cos}$. To encourage private latents to capture view-specific information, we introduce a loss term that enforces separation between the two private latents, z_{p1} and z_{p2}, by minimizing cosine similarity, denoted $\mathcal{L}_{cos^-}$. Along the same lines, we maximize cosine similarity between z_{s1} and z_{s2} to encourage the shared latents to contain similar information ($\mathcal{L}_{cos^+}$).

We experiment with these similarity or separation-based configurations on both a batch and sample level. The batch-level version includes inherent negatives from other samples in the batch similar to the traditional InfoNCE [27] setting, while the sample-level version treats cosine similarity as a simple binary classification, without cross-batch negatives. The sample-level similarity and separation-based objectives are defined below, where j or k refers to the index of samples within a batch of size B:

$$\mathcal{L}_{cos^+} = -\mathbb{E}_{j \in [1,B]} \big[\log sim(z^i_{s1}, z^i_{s2}) \big] \tag{2}$$

$$\mathcal{L}_{cos^-} = -\mathbb{E}_{j \in [1,B]} \big[\log(1 - sim(z^j_{p1}, z^j_{p2})) \big] \tag{3}$$

Similarly, the batch-level objectives are expressed as:

$$\mathcal{L}_{cos^+} = -\mathbb{E}_{j \in [1,B]} \left[\log \frac{\exp(sim(z^j_{s1}, z^j_{s2}))}{\sum_k^B \exp(sim(z^j_{s1}, z^k_{s2}))} \right] \tag{4}$$

$$\mathcal{L}_{cos^-} = -\mathbb{E}_{j,k \in [1,B]} [\log(1 - sim(z^j_{p1}, z^k_{p2}))] \tag{5}$$

When a cosine distance objective is used, it is added to the reconstruction objective for a final loss term of $\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{cos}$.

Masking: Drawing on works in masked acoustic token modeling from the generative-based methods [11], [19], [28], we mask a ratio (r) of random entries in the projected latent subspaces during training. By masking one latent subspace and leaving the other unchanged, the intuition is to encourage the model to rely more on a subset of the subspaces to infer necessary information. We leverage this latent subspace-based masking mechanism as a motivation to enforce the model to learn robust representations that capture the underlying structure in each subspace.

3. EXPERIMENTAL DESIGN

3.1. Dataset and Multi-View Data Pairing

We use the SONYC-UST-V2 dataset [29], which contains 10-second audio clips recorded from 56 sensors across New York City as part of the SONYC project [23]. The 56 sensor classes refer to individual recording devices placed in distinct urban locations, each capturing various city soundscapes. Sensors capture important channel effect information, namely the environmental acoustics unique to that location and microphone position. The recordings are non-overlapping in time, meaning that we can obtain multiple audio clips from the same sensor "for free" by selecting different temporal segments within a recording. Therefore, we consider our data pairing strategy as a self-supervised mechanism, as it is similar to sampling pairs of segments from the same recording in the traditional SSL setup. Each audio clip

Table 1: Downstream source and sensor classification vs. baselines. We present the results of our best-performing model, which uses reconstruction loss and sample-level cosine distance loss to separate private latents as the combined training strategy.

Method	Objective	Source $n_c = 8$	Sensor $n_c = 12$
Multi-view (Best Config.)	$\mathcal{L}_{rec}$ + $\mathcal{L}_{cos-}$	**0.633**	**0.735**
Single-view Autoencoder	$\mathcal{L}_{rec}$	0.582	0.710
Contrastive Learning	Info-NCE [27]	0.324	0.392
DAC [7]	N/A	0.583	0.684
Supervised Learning	BCE/CE	0.699	0.732

is labeled with one or more of 8 sound source categories, including engine, alarm, and human voice.

For this study, we use the recording *sensor* as the *shared* factor, pairing clips recorded by the same sensor at different times. These randomly selected pairs are assumed to naturally differ in *sound source content*, which we use as the *private* factor in this study. An example data pair could be two audio recordings from the same sensor location but recorded on different days, where one contains a dog barking sound and another contains an engine and car alarm.

We construct $39k$ training pairs, $6k$ for validation, and $11k$ for testing, containing 39/5/12 disjoint sensors respectively. We evaluate on unseen sensors to test the model's generalization ability and robustness for unseen scene variations. For downstream evaluation, we further partition this test set into train, validation, and test subsets stratified by sensor and sound source, resulting in an 8-class multi-label classification for downstream sound source classification, and 12-class classification for sensor. Importantly, we only use these labels in downstream evaluation and our core method is fully self-supervised.

3.2. Audio Preprocessing

We follow the parameters used for preprocessing audio for the Descript Audio Codec (DAC) [7]. We resample audio recordings from SONYC to 44.1KHz and normalize them to -24 dB LUFS, following DAC. We pass the full 10-seconds of audio to the pretrained DAC model[1]. For a 10-sec. audio clip, this yields an embedding of shape $(862, 1024)$, where 862 is the number of frames and 1024 is the feature dimension. We use the pre-quantized continuous latents from DAC. Pairs of these embeddings are used as input to our multi-view autoencoder.

3.3. Training Recipe

We train all of our models for 100 epochs on a single A100 GPU using a batch size of 16. We use a learning rate of $1e-3$ and AdamW optimization, and perform model selection using minimum total validation loss.

3.4. Downstream Classification

We evaluate the informativeness of our learned joint audio representations on downstream sensor and source classification tasks. After training our multi-view autoencoder model, we freeze the trained encoder and use it as a feature extractor to obtain private and shared latents (z_p and z_s) from a single input audio clip to use for downstream training. We train independent 2-layer MLP classifier heads on top of the private, shared, and concatenated latents separately for source and sensor classification. We use cross entropy (CE) as the objective for 12-class sensor classification, and binary cross entropy (BCE) as the objective for 8-class source classification in the multi-label setting.

3.5. Evaluation Metrics

We evaluate downstream performance using accuracy for 12-class sensor classification, and the Jaccard indexfor 8-class multi-label source classification. For each task, we use the following metrics:

Overall accuracy: We concatenate the private and shared subspaces ($z_p \oplus z_s$) and use this joint latent as the feature for the downstream task. This gives a measure of overall robustness and information retained in the learned embedding space.

Subspace accuracy: To better-understand the information structure in the latent subspace, we use either z_{pi} or z_{si} as features for downstream classification. This allows us to assess the information present in the separate learned subspaces.

Directional subspace classification ($DSC\Delta$): As a proxy metric to investigate the disentanglement level of the private and shared subspaces, we measure the difference in subspace performance for both source and sensor classification tasks. We define $DSC\Delta$ for the private or shared latent as:

$$\text{DSC}\Delta_{priv} = \text{clf}(z_p)_{\text{source}} - \text{clf}(z_s)_{\text{source}} \tag{6}$$

$$\text{DSC}\Delta_{shared} = \text{clf}(z_s)_{\text{sensor}} - \text{clf}(z_p)_{\text{sensor}}, \tag{7}$$

where $\text{clf}(\cdot)$ indicates the downstream classifier trained per task. An ideal scenario for disentanglement in our framework is for the private latent to capture view-specific information (source in this study), and for the shared to capture common information (sensor). If there is no disentanglement and the subspace performance using either latent is identical, this yields $DSC\Delta = 0$. Thus, we aim for a positive $DSC\Delta$ score for each latent, indicating the desired disentanglement.

3.6. Baseline Methods

We apply the framework above to a series of baseline training strategies, each built upon the pretrained DAC latents as input:

DAC [7] without training: The naive baseline is training the downstream classifiers directly on top of off-the-shelf DAC latents. This represents the baseline classification potential of the latents before any additional training or disentanglement.

Single-view Autoencoder: We apply the traditional generative learning pipeline using the same encoder and decoder architecture as our method with the reconstruction loss $\mathcal{L}_{rec}$, but with just one view.

Contrastive learning: We apply the traditional contrastive learning framework using the same encoder architecture as our method, but without (1) the notion of separated latent subspaces and (2) using a decoder for reconstruction. This yields one latent space of the same dimensionality as the joint latent in our multi-view based methods. We train the encoder on pairs of data with the same dataset used in our multi-view framework (data paired by common sensor class), utilizing only one training strategy of an InfoNCE [27] loss, where each pair is considered a positive sample and samples from different pairs are considered negative samples.

Supervised learning: using the same encoder as our method, with an additional one-layer linear classification head for either multi-label source classification (BCE objective), or multi-class sensor classification (CE objective), without reconstruction. This method provides an upper bound of task performance, not a direct comparison with our proposed methods, since our training is label-free.

4. RESULTS AND DISCUSSION

We train our model with different combinations of training strategies introduced in Sec. 2.2. Our base method uses a multi-view learning backbone with $\mathcal{L}_{rec}$. Variants of our method are trained using one of the following training strategies on top of the base method: sample

[1] We use the 44.1KHz, 8kbps bitrate pretrained DAC.

Table 2: Examining the effects of varied objective functions on downstream source and sensor classification.

Objective	**Configuration**	**Source** $n_c = 8$	**Sensor** $n_c = 12$	**Avg. Acc.**
$\mathcal{L}_{rec}$	N/A	0.610	**0.739**	0.675
$\mathcal{L}_{rec} + \mathcal{L}_{cos-}$	**Sample-level**	**0.633**	0.735	**0.684**
	Batch-level	0.610	0.737	0.674
$\mathcal{L}_{rec} + \mathcal{L}_{cos+}$	Sample-level	0.593	0.719	0.656
	Batch-level	0.603	0.728	0.666
$\mathcal{L}_{rec}$ + Mask z_p	$r = 0.4$	0.594	0.733	0.664

or batch-level cosine distance objectives applied to the private latents to encourage separation ($\mathcal{L}_{cos-}$) or the shared latents to encourage similarity ($\mathcal{L}_{cos-}$), and lastly masking z_p with masking ratio $r \in [0.2, 0.4, 0.6, 0.8]$.

4.1. Main Comparisons

The evaluation results of our best-performing model configurations and baselines are shown in Table 1. We first observe that the standard contrastive learning approach, trained using data paired by the same sensor class using Info-NCE, yields lower accuracy across the board for both tasks. This indicates that the assumption of commonality in CL is less effective in this nuanced task of sensor classification, and fails at extracting sound source information without an appropriate pairing strategy. Additionally, the more traditional single-view autoencoder provides 3.9% relative improvement on sensor classification but no improvement on source classification when compared to DAC latents without any training (denoted with “N/A” objective), In contrast, even without curating data pairs with matching sound source information, our method combining generative and contrastive principles is able to successfully capture information in both tasks.

Our best model configuration, using $\mathcal{L}_{rec}$ and sample-level $\mathcal{L}_{cos-}$, improves overall accuracy results by 8.6% and 3.5% for source and sensor classification respectively, compared to the single-view autoencoder baseline. Further, this configuration significantly outperforms the contrastive and DAC-only baselines. Such improvements indicate the complementary effects of contrastive and generative principles to produce robust environmental sound representations, while overcoming limits of individual traditional SSL methods.

We also include a fully supervised upper bound for reference. While the supervised sensor classification achieves the highest performance with the help of complete label supervision, our method significantly narrows the gap with no explicit prior knowledge about sensor and source information. At the same time, we observe that our training framework even leads to a 0.4% relative performance improvement on sensor classification versus a supervised approach.

In Table 2, we investigate how different objective strategies introduced in Sec. 2.2 affect the the learned representations on downstream performance. We find that separation-based objectives ($\mathcal{L}_{cos-}$) tend to perform better than similarity-based objectives ($\mathcal{L}_{cos+}$), for both sample and batch-level experiments. We do not observe a consistent trend between sample and batch-level cosine distance-based experiments across tasks; for $\mathcal{L}_{cos-}$, sample-level performs better (our best configuration in terms of averaged accuracy), but for $\mathcal{L}_{cos+}$, the batched version is marginally better than sampled.

4.2. Disentanglement Investigation

We also observe the potential for environmental sound attribute disentanglement using our method and investigate this under different training configurations. In Figure 2, we visualize the trend of information flow for source (left) and sensor (right) attributes grouped by three types of strategies: cosine distance-based disentanglement using only the multi-view backbone with $\mathcal{L}_{rec}$ (green), adding $\mathcal{L}_{cos+}$ or $\mathcal{L}_{cos-}$ (orange), or $\mathcal{L}_{rec}$ with masking (blue). Each point represents a specific training configuration within one type of training strategy, e.g. a specific masking ratio r within the “masking” category, or sample-based $\mathcal{L}_{cos+}$ within the $\mathcal{L}_{cos}$ grouping. The x-axis in Fig. 2 represents the overall classification accuracy obtained using the complete latent space ($z_p \oplus z_s$). The red vertical line marks the accuracy scores per task of the DAC baseline as reported in Table 1. The y-axis represents the DSCΔ to measure the level of information flow in the desired direction as a proxy for disentanglement quality.

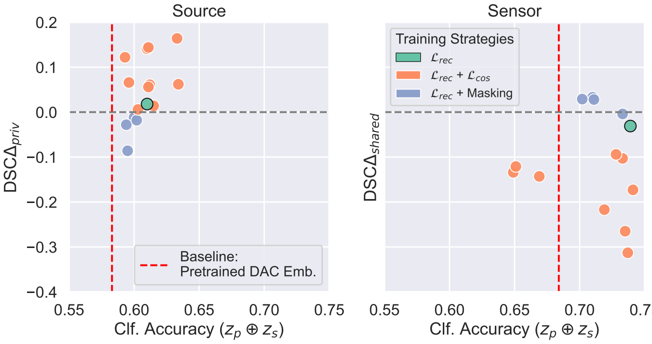

Fig. 2: Visualizing the effects of cosine distance and masking strategies on downstream tasks using the overall performance vs. the Directional Subspace Classification metric ($DSC\Delta$).

We first show that the majority our multi-view strategies infuse useful source information in the latent subspaces, outperforming the pretrained DAC baseline marked with the red dotted line. We also observe that the multi-view backbone with only $\mathcal{L}_{rec}$ has a positive $DSC\Delta$ for source, but negative $DSC\Delta$ for sensor, indicating that without explicit constraints utilizing multi-view assumptions, the model shows a tendency to steer information about both source and sensor into the private subspace. While Fig. 2 suggests that cosine distance-based strategies generally tend to push both source and sensor information into the private subspace, masking z_p shows the opposite effect; masking shifts overall information into the shared subspace. However, we did not observe obvious trend for different masking ratios. We speculate that the shared sensor information between views may be too subtle to capture in the common subspace with reconstruction alone, while masking z_p puts more weight on the shared subspace to encode all information to optimize reconstruction.

5. CONCLUSION AND FUTURE WORK

In this work, we proposed a novel self-supervised multi-view learning framework that integrates contrastive principles within a generative pipeline to improve the robustness of environmental sound representations. Our experiments on the SONYC-UST-V2 dataset demonstrate that our method improves downstream performance in both source and sensor classification on recordings in unseen sensors compared to traditional SSL methods. Beyond improved representation robustness, we investigate the effects of different training strategies on latent subspace information flow, showing the potential for environmental sound attribute disentanglement. In the future, we plan to explore using audio understanding models within our framework such as AudioMAE [11] or BEATs [12], and extend downstream tasks to out-of-distribution data.

6. ACKNOWLEDGMENTS

We thank our colleagues at Bosch for their support on this work, which is partially supported by award #24-2283 (Robert Bosch, LLC).

REFERENCES

[1] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3875–3879.

[2] E. Fonseca, D. Ortego, K. McGuinness, N. E. O'Connor, and X. Serra, "Unsupervised contrastive learning of sound event representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 371–375.

[3] H. Al-Tahan and Y. Mohsenzadeh, "Clar: Contrastive learning of auditory representations," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2530–2538.

[4] A. Nasiri and J. Hu, "Soundclr: Contrastive learning of representations for improved environmental sound classification," 2021. [Online]. Available: https://arxiv.org/abs/2103.01929

[5] M. Mahyub, L. S. Souza, B. Batalo, and K. Fukui, "Signal latent subspace: A new representation for environmental sound classification," *Applied Acoustics*, vol. 225, p. 110181, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X24003323

[6] D. Perera, S. Essid, and G. Richard, "Latent and adversarial data augmentations for sound event detection and classification," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[7] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," 2023. [Online]. Available: https://arxiv.org/abs/2306.06546

[8] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," 2022. [Online]. Available: https://arxiv.org/abs/2210.13438

[9] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "Ssast: Self-supervised audio spectrogram transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 699–10 709.

[10] D. Chong, H. Wang, P. Zhou, and Q. Zeng, "Masked spectrogram prediction for self-supervised audio pre-training," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[11] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 708–28 720, 2022.

[12] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 5178–5193. [Online]. Available: https://proceedings.mlr.press/v202/chen23ag.html

[13] J. Abeßer, S. I. Mimilakis, R. Gräfe, H. M. Lukashevich, and I. Fraunhofer, "Acoustic scene classification by combining autoencoder-based dimensionality reduction and convolutional neural networks." in *DCASE*, 2017, pp. 7–11.

[14] P. Zinemanas, M. Rocamora, E. Fonseca, F. Font, and X. Serra, "Toward interpretable polyphonic sound event detection with attention maps based on local prototypes." in *DCASE*, 2021, pp. 50–54.

[15] J. Abeßer, S. I. Mimilakis, R. Gräfe, and H. Lukashevich, "Acoustic scene classification by combining autoencoder-based dimensionality reduction and convolutional neural networks," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE2017)*, November 2017, pp. 7–11.

[16] K. Wilkinghoff and F. Kurth, "Open-set acoustic scene classification with deep convolutional autoencoders," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 258–262.

[17] J. Xu, Y. Ren, H. Tang, X. Pu, X. Zhu, M. Zeng, and L. He, "Multi-vae: Learning disentangled view-common and view-peculiar visual representations for multi-view clustering," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9234–9243.

[18] M. Lee and V. Pavlovic, "Private-shared disentangled multimodal vae for learning of hybrid latent representations," 2020. [Online]. Available: https://arxiv.org/abs/2012.13024

[19] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. R. Glass, "Contrastive audio-visual masked autoencoder," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=QPtMRyk5rb

[20] K. Tanaka, K. Yoshii, S. Dixon, and S. Morishima, "Unsupervised pitch-timbre-variation disentanglement of monophonic music signals based on random perturbation and re-entry training," *APSIPA Transactions on Signal and Information Processing*, vol. 14, no. 1, pp. –, 2025. [Online]. Available: http://dx.doi.org/10.1561/116.20240072

[21] Y.-J. Luo, K. W. Cheuk, T. Nakano, M. Goto, and D. Herremans, "Unsupervised disentanglement of pitch and timbre for isolated musical instrument sounds." in *ISMIR*, 2020, pp. 700–707.

[22] J. Wilkins, S. Ding, M. Fuentes, and J. P. Bello, "Self-supervised multi-view learning for disentangled music audio representations," 2024. [Online]. Available: https://arxiv.org/abs/2411.02711

[23] 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.

[24] 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.

[25] 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.

[26] J. Guinot, E. Quinton, and G. Fazekas, "Leave-one-equivariant: Alleviating invariance-related information loss in contrastive music representations," 2024. [Online]. Available: https://arxiv.org/abs/2412.18955

[27] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019. [Online]. Available: https://arxiv.org/abs/1807.03748

[28] H. F. Garcia, P. Seetharaman, R. Kumar, and B. Pardo, "Vampnet: Music generation via masked acoustic token modeling," 2023. [Online]. Available: https://arxiv.org/abs/2307.04686

[29] M. Cartwright, J. Cramer, A. E. M. Mendez, Y. Wang, H.-H. Wu, V. Lostanlen, M. Fuentes, G. Dove, C. Mydlarz, J. Salamon, O. Nov, and J. P. Bello, "Sonyc-ust-v2: An urban sound tagging dataset with spatiotemporal context," 2020. [Online]. Available: https://arxiv.org/abs/2009.05188