INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Robust Prototype Learning for Anomalous Sound Detection

Xiao-Min Zeng[1], *Yan Song*[1], *Ian McLoughlin*[1,2], *Lin Liu*[3], *Li-Rong Dai*[1]

[1]National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China.
[2]ICT Cluster, Singapore Institute of Technology, Singapore.
[3]iFLYTEK Research, iFLYTEK CO. LTD., Hefei, China.

zxmin115@mail.ustc.edu.cn, {songy, ivm, lrdai}@ustc.edu.cn, liulin@iflytek.com

Abstract

In this paper, we present a robust prototype learning framework for anomalous sound detection (ASD), where prototypical loss is exploited to measure the similarity between samples and prototypes. We show that existing generative and discriminative based ASD methods can be unified into this framework from the perspective of prototypical learning. For ASD in recent DCASE challenges, extensions related to imbalanced learning are proposed to improve the robustness of prototypes learned from source and target domains. Specifically, balanced sampling and multiple-prototype expansion (MPE) strategies are proposed to address imbalances across attributes of source and target domains. Furthermore, a novel negative-prototype expansion (NPE) method is used to construct pseudo-anomalies to learn a more compact and effective embedding space for normal sounds. Evaluation on the DCASE2022 Task2 development dataset demonstrates the validity of the proposed prototype learning framework.

Index Terms: anomalous sound detection, prototype learning, imbalanced learning

1. Introduction

In recent DCASE challenges[1], anomalous sound detection (ASD) is defined as a task that determines whether a machine is behaving abnormally or not through an analysis of its sound. The application is automatic monitoring of machine conditions. This task is challenging since anomalies rarely occur and are highly diverse, especially in the case of domain shifted conditions where factors such as environmental noise or operational conditions often differ between training and test phases [1, 2, 3].

Conventional ASD systems exploit generative methods including AutoEncoder (AE) [1, 4], Gaussian Mixture Model (GMM) [5] and WavNet [6] to model the distribution of normal data in an unsupervised way. Recent discriminative-based methods, such as MobileNetV2 [2, 3], ResNet [7, 8], and STgram-MFN [9], were proposed to learn effective embeddings with deep neural networks (DNNs). With the help of label information (*e.g.* machine-IDs, sections or attributes), they can learn more compact and effective sound clip embeddings than the generative ones, achieving promising ASD performance [1, 2, 3].

It was shown in [10] that AEs act similarly to k-means and Principal Component Analysis (PCA), with an encoder mapping an input to a latent embedding space, and a decoder reconstructing the input. The encoder can be implicitly considered as prototypes (*i.e.* centres in k-means), a description of the distribution of the normal input data. For discriminative-based methods, a classification loss function (*e.g.* softmax cross-entropy loss) is employed to optimize the similarity between samples and weighted vectors, *i.e.* prototypes of each class. From this perspective, it is crucial to learn effective prototypes in both generative and discriminative based ASD systems.

For more challenging domain generalization (DG) ASD tasks in DCASE 2022, several domain-classification-based approaches were reported in the literature [11, 12]. In the former, a multi-task learning framework was proposed to disentangle domain-shared and domain-specific features for domain generalization in ASD. In [8], a pre-trained source-domain network was further fine-tuned with a few samples from the target domain. Additionally, several domain-mixing-based approaches [13, 14, 15] that exploit data augmentation methods like Mixup [16] or SpecAugment [17] were explored. These aim to synthesize more target domain samples to improve the generalization capability of the learned model. However, since source and target domain attributes are non-overlapping and highly imbalanced, it is still difficult to address domain generalization and class imbalance issues simultaneously [11, 12]. Domain-mixing-based methods [13, 14, 15] may suffer from inaccurate synthesized data, where the synthesis may not necessarily follow the distribution of normal data [3].

In this paper, we first present a unified deep neural network framework to learn effective and high quality prototypes, as illustrated in Fig. 1. Prototypical loss is exploited for network optimization by measuring the similarity between samples and the prototype representation of each class. We then extend the framework to address the DCASE 2022 ASD task by casting domain generalization as imbalanced learning from both source and target domains, to improve the robustness of learned prototypes. It is worth noting that this framework can learn a unified embedding space, enabling the same threshold regardless of the domain. Specifically, multiple-prototype expansion (MPE) and balanced sampling strategies are proposed to address the imbalance across non-overlapping attributes of source and target domains. Furthermore, from the perspective of outlier exposure [18], a novel negative-prototype expansion (NPE) strategy is proposed. We apply Mixup [16] between anchor and negative prototypes to construct pseudo-anomalies to learn a more compact and effective embedding space from normal data. This differs from domain-mixing-based methods [13, 14, 15] that treat the augmented samples as normal.

To evaluate the effectiveness of the proposed prototype learning framework, extensive experiments on the DCASE2022 Task2 development dataset reveal its excellent performance for general ASD tasks and demonstrate the flexibility to adapt well to specific tasks.

Yan Song is the corresponding author.

[1]DCASE: Detection and Classification of Acoustic Scenes and Events, https://dcase.community

10.21437/Interspeech.2023-1173

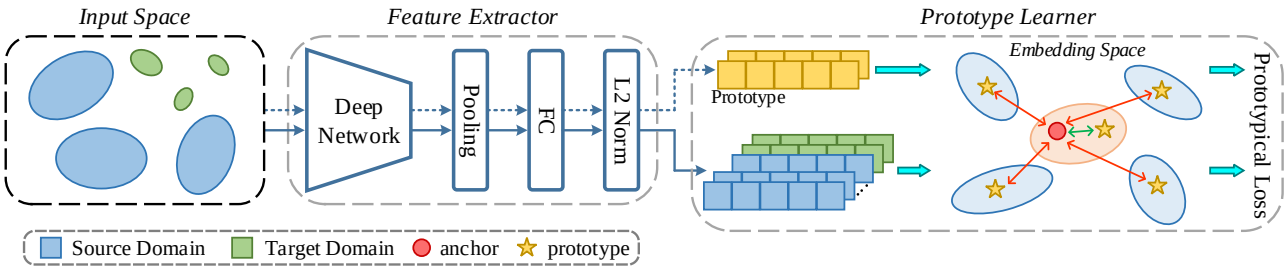


Figure 1: *An overview of the proposed unified prototype learning framework. All training data is sampled randomly to train the feature extractor with the constraint of prototypical loss. The dotted line is implemented before each epoch to obtain prototypes. In embedding space, the anchor is pulled closer to the corresponding prototype and pushed further away from other prototypes, aiming to yield more effective and compact feature representations.*

2. Unified prototype learning framework

As aforementioned, existing ASD methods, both generative and discriminative, can be formulated as tasks that learn prototypical representations from normal samples. From this perspective, we propose a unified prototype learning framework, as shown in Fig. 1. It consists of a feature extractor f_θ parameterized with θ, and a prototype learner using prototypical loss.

Let $\mathcal{X} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ denote the training set and $\mathcal{Y} = \{1, \cdots, C\}$ is the corresponding attribute label space. $\mathcal{S}_c$ denotes a set of all training samples labeled with attribute c. First, we use all available normal data to train feature extractor f_θ by randomly sampling from $\mathcal{X}$. The model f_θ maps each input sample $\mathbf{x}$ into a fixed dimension embedding vector $\mathbf{z}$ with unit length. In the prototype learner, the prototypical loss is exploited to optimize feature extractor f_θ by calculating the distances between prototypes and embeddings, that is,

$$\mathcal{L}_{PL} = -\log \frac{\exp(\mathbf{z}_i \cdot \mathbf{p}_i/\tau)}{\sum_{c=1}^{C} \exp(\mathbf{z}_i \cdot \mathbf{p}_c/\tau)} \tag{1}$$

where τ is a temperature hyper-parameter and $\mathbf{p}_i$ is the attribute prototype of $\mathbf{z}_i$, which is calculated as follows:

$$\mathbf{p}_c = \frac{1}{|\mathcal{S}_c|} \sum_{\mathbf{x}_i \in \mathcal{S}_c} f_\theta(\mathbf{x}_i) \tag{2}$$

Note that $\mathbf{p}_c$ also requires to be normalized.

From the perspective of outlier exposure [18], the framework treats sounds from different attributes as pseudo anomalies. In contrast to calculating centres specifically for inference [8], our framework directly utilizes prototypes to optimize the model, which alleviates the inconsistency between training and testing. Besides, our proposed framework is convenient to manage and modify for specific ASD tasks.

For domain generalization in ASD, it is crucial to learn effective prototypes for the source and target domains. However, the learned prototypes are often not robust enough due to the imbalance between the domains. To address this issue, we propose imbalanced learning strategies and anomaly simulation methods to extend our prototype learning framework.

3. Robust prototype learning

In this section, the extensions related to domain generalization are described in detail, including imbalanced learning strategies (*i.e.* balanced sampling and MPE) and an anomaly simulation method called NPE.

3.1. Imbalanced learning

3.1.1. Multiple-prototype expansion (MPE)

The multiple-prototype expansion (MPE) is presented to generate sub-prototypes for learning more effective feature representation, as shown in Fig. 2(a). For a given set $\mathcal{S}_c$, we randomly select $m\%$ samples to form a new subset $\tilde{\mathcal{S}}_c$ and calculate sub-prototypes $\tilde{\mathbf{p}}_c$ in terms of Eqn. (2). After collecting prototypes $\mathbf{p}$ and sub-prototypes $\tilde{\mathbf{p}}$, for an arbitrary anchor $\mathbf{z}_i$, multiple prototypes are available. Applying the MPE in prototype learner, we update Eqn. (1) to satisfy the multi-prototype situation according to supervised contrastive learning [19],

$$\mathcal{L}_{PL\text{-}MPE} = -\frac{1}{|\hat{\mathcal{P}}_c|} \sum_{\mathbf{p}_c \in \hat{\mathcal{P}}_c} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{p}_c/\tau)}{\sum_{\mathbf{p}_a \in \hat{\mathcal{P}}} \exp(\mathbf{z}_i \cdot \mathbf{p}_a/\tau)} \tag{3}$$

where $\hat{\mathcal{P}}$ consists of all $\mathbf{p}$ and $\tilde{\mathbf{p}}$, and $\hat{\mathcal{P}}_c$ is the set of multi-prototypes corresponding to $\mathbf{z}_i$.

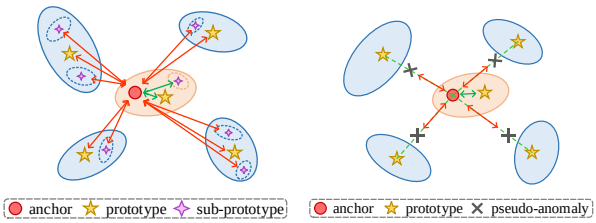
Based on the concept of sub-sampling, the MPE can generate diverse indicative information to guide representation learning effectively. More diverse prototypes are beneficial to reduce intra-class variations, especially for the source domain with a large number of training samples. Furthermore, analyzing the construction of subsets $\tilde{\mathcal{S}}_c$, training data from the target domain is sampled repeatedly due to its size limitation. This operation essentially over-samples the target domain to help alleviate the effect of an imbalance between the two domains.

3.1.2. Balanced sampling

Random sampling adopted in the unified prototype learning framework treats all data in $\mathcal{X}$ equally, leading to insufficient optimization for the target domain, and overfitting for the source domain. Therefore, we extend the idea of sub-sampling to the input space and propose a balanced sampling strategy to rebalance attribute classes in both domains.

During training, a subset $\tilde{\mathcal{Y}}$, composed of N attributes, is sampled from attribute space $\mathcal{Y}$, and then the mini-batch is collected according to $\tilde{\mathcal{Y}}$. For the source domain, a subset $\tilde{\mathcal{S}}_c$ composed of K training data can be directly constructed by random sampling from $\mathcal{S}_c$. For the target domain, K samples are obtained by RandomCrop due to limited sample numbers.

After balanced sampling, the mini-batch is exploited to optimize the feature extractor. Unlike the unified prototype learning framework, prototype $\mathbf{p}_c$ is calculated by replacing $\mathcal{S}_c$ with subset $\tilde{\mathcal{S}}_c$ in Eqn. (2), which allows the prototypes to be updated



(a) MPE (b) NPE

Figure 2: *Illustration of our proposed two expansion modules. (a) Multiple-prototype expansion (MPE) takes the average vector of a subset from the given attribute to obtain sub-prototypes, aiming to promote the diversity of prototypes and reduce intra-class variations. (b) Negative-prototype expansion (NPE) performs linear interpolation on the connection between anchor embedding and other attribute prototypes to generate pseudo-anomalies, resulting in improved inter-class separability.*

before each iteration. The total loss for balanced sampling is,

$$\mathcal{L} = \frac{1}{N \times K} \sum_{N} \sum_{K} \mathcal{L}_{PL} \tag{4}$$

As in [20], the balanced label space is first sampled so that the source and target domains can be trained in a re-balanced manner. Essentially, this addresses the imbalance issue by under-sampling source domain data. The balanced sampling strategy allows limited samples from the target domain to sufficiently optimize the model, resulting in a more efficient representation of the target domain. In addition, when balanced sampling is adopted in our experiments, $\mathcal{L}_{PL}$ is modified due to the introduction of the NPE described in Sec. 3.2.

3.2. Anomaly simulation

In the later stages of training, discrepancies tend to form among every attribute, lacking hard enough negative examples to optimize the model further. Inspired by [21], we propose negative-prototype expansion (NPE) to synthesize pseudo-anomalies to improve the representation learning.

As shown in Fig. 2(b), given an anchor and all prototypes obtained from a mini-batch, more negative pseudo-anomalies are generated by linear interpolation between the anchor and other prototypes. As in [16], the synthesis of pseudo-anomalies is formalized as:

$$\hat{\mathbf{z}}_i = \lambda \mathbf{z}_a + (1 - \lambda)\mathbf{p}_i \tag{5}$$

where, $\mathbf{z}_a$ is the embedding of selected anchor with attribute label c and $\mathbf{p}_i$ is the prototype of the other class, *i.e.* $i \neq c$. λ is the mixing coefficient, which follows a uniform distribution. Once more pseudo-anomalies are generated, the prototypical loss in Eqn. (1) is updated to,

$$\mathcal{L}_{PL\text{-}NPE} = -\log \frac{\exp(\mathbf{z}_a \cdot \mathbf{p}_c/\tau)}{\sum_i \exp(\mathbf{z}_a \cdot \mathbf{p}_i/\tau) + \sum_{j,j\neq c} \exp(\mathbf{z}_a \cdot \hat{\mathbf{z}}_j/\tau)} \tag{6}$$

where all generated vectors $\hat{\mathbf{z}}$ are considered as negatives, *i.e.* simulated anomalies. Analyzing Eqn. (6), the synthesis of pseudo-anomalies is equivalent to adding a margin between the different classes, causing more discriminative representations to be learned.

4. Experiments and results

4.1. Dataset

We conducted extensive experiments on the DCASE2022 Challenge Task2 development dataset that is composed of a subset of ToyADMOS2 [22] and MIMII DG [23] dataset. Each sample is a single-channel, 10-second audio clip. Seven machine types are recorded, with data from each machine type divided into three sections. For each section in training set, there are 990 clips of normal samples in the source domain, but only around 10 normal sounds from the target domain are available. The test set provides 50 clips of normal and anomalous samples in the source and target domains, respectively. Each sound is labeled with attribute information according to various conditions (*e.g.* voltage, velocity or factory noise, etc.) and the specific number of attributes are presented as *machine type(source domain, target domain)* as follows: ToyCar(9,12), ToyTrain(9,12), Bearing(17,13), Fan(6,5), Gearbox(13,13), Slider(12,9), Valve(9,8). Note that there are no shared attributes between the source and target domains, so they have 75 and 72 attributes, respectively.

4.2. Implementation details

We extract the inverted log-Mel spectrogram [24], which emphasizes the high-frequency region, as input features. In detail, 128 inverted Mel filters are used with a window size of 1024 and hop size of 512 for all 16kHz sample rate input clips.

In experiments, we apply the unified prototype framework with MPE to pre-train the feature extractor. Then, balanced sampling and NPE are utilized to fine-tune the model. The feature extractor f_θ follows our previous work [8], which adopts ResNet18 with Time-Frequency Attention Pooling. In both stages, 64 frames are randomly cropped from the input feature, and the model is optimized by SGD with momentum of 0.9 and weight decay of 5e-5. During pre-training, the model is trained for 120 epochs with a step-based decaying learning rate (*i.e.* 0.1×50 epochs, 0.01×40 epochs, 0.001×30 epochs) and mini-batch of 32. During fine-tuning, we construct a mini-batch with $N = 10$ and $K = 10$ to optimize the model over 50 epochs. An initial learning rate of 0.01 is set during the first 30 epochs, declining to 0.001 for the remaining 20 epochs. All temperature factors τ in the loss functions are set to 0.07.

During testing, attribute prototypes from both domains are utilized together to calculate the anomaly score. For each test sound clip, the Mahalanobis distance is found between the test embedding and the prototypes of all attributes in the corresponding section. The minimum distance is the anomaly score. Following [12], the covariance matrix is calculated from all section embeddings without distinguishing domains.

Three standard metrics evaluate ASD performance: the source AUC and target AUC that are calculated by comparing the normal test samples from a given domain against anomalies from both domains, along with pAUC computed over test samples from all domains, where pAUC is calculated as the AUC over a low false-positive-rate (FPR) range [0, 0.1].

4.3. Results

We compare the proposed methods with previous systems and present the performance metrics in Table 1. The baseline system adopts a self-supervised framework [8] to recognize all attributes. Besides this, we also present the results of each training stage to verify the effectiveness of our unified framework, where 'Fine-tune (w/o pre-train)' denotes that fine-tuning is performed from scratch instead of the pre-trained model.

Table 1: *The harmonic mean of three evaluation metrics for each machine type obtained from different methods. The source AUC (%), target AUC (%), and pAUC (%) are harmonic mean across all machine types for each of the three metrics. The overall refers to the harmonic mean of all metrics.*

Methods	ToyCar	ToyTrain	Bearing	Fan	Gearbox	Slider	Valve	Source AUC	Target AUC	pAUC	Overall
AutoEncoder [3]	51.30	39.77	60.64	58.50	63.07	58.00	50.60	70.55	42.27	54.51	53.40
MobileNetV2 [3]	54.40	51.55	60.64	57.53	60.17	51.68	62.13	64.61	50.98	55.77	56.58
Baseline	74.67	58.68	67.84	57.20	74.30	80.12	90.45	78.78	69.45	64.04	70.25
STgram-MFN [9]	68.38	62.25	59.54	63.36	62.75	73.84	60.44	71.21	64.32	57.97	64.05
DG-mix [14]	**79.80**	53.70	70.70	**76.60**	80.28	75.11	79.13	77.55	74.08	66.34	72.35
Disent_Wt [11]	76.95	59.74	72.07	63.91	**81.38**	85.14	**94.50**	**86.09**	71.65	**68.21**	74.57
Pre-train	74.81	62.03	71.32	58.16	74.25	82.69	92.23	84.75	69.70	64.64	72.09
Fine-tune (w/o pre-train)	77.40	59.71	67.19	57.75	70.80	83.24	76.02	74.52	71.74	62.52	69.20
Fine-tune (w/ pre-train)	79.08	**62.95**	**78.53**	61.34	76.50	**85.34**	93.71	85.41	**76.03**	66.63	**75.25**

Table 2: *The harmonic mean source AUC (%), target AUC (%), and pAUC (%) for different selected ratio $m\%$ in ablation experiments on MPE in pre-training.*

ratio $m\%$	Source AUC	Target AUC	pAUC	Overall
w/o MPE	83.53	68.60	63.53	70.94
50%	84.54	**70.02**	63.31	71.59
25%	**84.75**	69.70	**64.64**	**72.09**
12.5%	83.98	68.49	64.58	71.44

In terms of overall results, our proposed prototype learning methods significantly outperform most ASD systems. We believe the primary reason is that the unified prototype learning framework is so flexible that our proposed domain-related methods can be conveniently applied. Compared to the baseline, the results of pre-training show that the unified prototype framework achieves an improvement for the source domain. After fine-tuning, our framework obtains superior results in the target domain while further improving the performance in the source domain, achieving a trade-off between the two domains. Moreover, it is worth noting that the performance of 'Fine-tune (w/o pre-train)' shows a little gap between the source and target domains, which indicates the effectiveness of balanced sampling for addressing the imbalanced learning issue. However, skipping the pre-training also weakens the representational ability of the source domain. This is because balanced sampling under-samples source domain data, causing some loss of information from that domain.

In Table 2, we present the results of ablation experiments on the MPE module, where the selected ratio $m\%$ indicates the percentage of samples over which the sub-prototype is calculated. It is not hard to observe that our proposed MPE method improves the overall performance. The best overall performance is achieved when 25% samples are utilized to obtain sub-prototypes. That is because the sub-prototypes provide more information on positives and negatives, promoting the descriptiveness and discrimination of the learned representations. Since more diverse sub-prototypes are offered for representation learning, performance on the source domain achieves a consistent improvement. For the target domain, the oversampling in MPE also benefits its performance, but $m\%$ needs to be chosen carefully due to the limited number of samples. When the selected ratio is set to 12.5%, the gain of MPE is limited. This may be because the obtained sub-prototypes introduce bias in the description of original attributes, which affects the representation learning for ASD.

Table 3: *The harmonic mean source AUC (%), target AUC (%), and pAUC (%) of ablation experiments on NPE with different ranges of λ.*

Range of λ	Source AUC	Target AUC	pAUC	Overall
Pre-train	84.75	69.70	64.64	72.09
w/o NPE	84.35	73.37	65.40	73.57
[0, 0.25]	85.26	73.16	66.55	74.21
[0.25, 0.5]	**85.41**	**76.03**	**66.63**	**75.25**
[0.5, 0.75]	79.34	74.61	65.11	72.52
[0.75, 1.0]	75.00	72.65	61.44	69.17

We also study the influence of NPE module mixing coefficient λ on performance, as shown in Table 3. Compared with pre-training, even without NPE, fine-tuning with balanced sampling not only significantly improves target domain performance but also maintains performance in the source domain. After applying NPE, more pseudo-anomalies are synthesized by mixing, which allows both domains to be further optimized. However, once the value of λ is too large, simulated anomalies are too close to the anchor, resulting in inaccurate representation learning for the anchor. The negative effect can be noticed in Table 3, where the source domain performance declines significantly if $\lambda > 0.5$. Therefore, it is crucial to select a reasonable range of λ. When λ follows a uniform distribution between $[0.25, 0.5]$, the synthetic negatives are hard enough and far enough away from the anchor, achieving the best performance on all three metrics.

5. Conclusion

In this paper, we argue that the quality of a prototype plays an important role in both generative and discriminative ASD methods. Based on this unified perspective, we propose a prototype learning framework, which can be extended to various ASD tasks. Specifically, we apply the framework to address domain generalization in ASD, where a high degree of imbalance exists between source and target domains. The proposed balanced sampling and MPE strategies within the framework can mitigate this imbalance. Meanwhile, an NPE strategy is employed to simulate anomalies to improve inter-class representation and thus enhance the discriminative ability. Extensive experiments on the DCASE2022 Task2 development dataset demonstrate the flexibility of the prototype learning framework, and the effectiveness of its extension for domain generalization in ASD.

6. References

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, Tokyo, Japan, November 2020, pp. 81–85.

[2] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 186–190.

[3] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, and Y. Kawaguchi, "Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[4] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 271–275.

[5] F. Xiao, Y. Liu, Y. Wei, J. Guan, Q. Zhu, T. Zheng, and J. Han, "The DCASE2022 challenge task 2 system: Anomalous sound detection with self-supervised attribute classification and GMM-based clustering," DCASE2022 Challenge, Tech. Rep., July 2022.

[6] E. Rushe and B. M. Namee, "Anomaly detection in raw audio using deep autoregressive networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3597–3601.

[7] H. Hojjati and N. Armanfard, "Self-supervised acoustic anomaly detection via contrastive learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3253–3257.

[8] H. Chen, Y. Song, L.-R. Dai, I. McLoughlin, and L. Liu, "Self-supervised representation learning for unsupervised anomalous sound detection under domain shift," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 471–475.

[9] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 816–820.

[10] G. E. Hinton and R. Zemel, "Autoencoders, minimum description length and helmholtz free energy," in *Advances in Neural Information Processing Systems*, J. Cowan, G. Tesauro, and J. Alspector, Eds., vol. 6. Morgan-Kaufmann, 1993. [Online]. Available: https://proceedings.neurips.cc/paper/1993/file/9e3cfc48eccf81a0d57663e129aef3cb-Paper.pdf

[11] S. Venkatesh, G. Wichern, A. Subramanian, and J. Le Roux, "Improved domain generalization via disentangled multi-task learning in unsupervised anomalous sound detection," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[12] Y. Zeng, H. Liu, L. Xu, Y. Zhou, and L. Gan, "Robust anomaly sound detection framework for machine condition monitoring," DCASE2022 Challenge, Tech. Rep., July 2022.

[13] I. Kuroyanagi, T. Hayashi, K. Takeda, and T. Toda, "Two-stage anomalous sound detection systems using domain generalization and specialization techniques," DCASE2022 Challenge, Tech. Rep., July 2022.

[14] I. Nejjar, J. Meunier-Pion, G. Frusque, and O. Fink, "Dg-mix: Domain generalization for anomalous sound detection based on self-supervised learning," in *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 2022.

[15] S. Verbitskiy, M. Shkhanukova, and V. Vyshegorodtsev, "Unsupervised anomalous sound detection using multiple time-frequency representations," DCASE2022 Challenge, Tech. Rep., July 2022.

[16] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *6th International Conference on Learning Representations*, 2018.

[17] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2613–2617.

[18] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *International Conference on Learning Representations*, 2018.

[19] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

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

[21] B. Ko and G. Gu, "Embedding expansion: Augmentation in embedding space for deep metric learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7255–7264.

[22] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," in *Proceedings of the 6th Detection and Classification of Acoustic Scenes and Events 2021 Workshop (DCASE2021)*, Barcelona, Spain, November 2021, pp. 1–5.

[23] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," *In arXiv e-prints: 2205.13879*, 2022.

[24] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison of features for synthetic speech detection," in *Proc. Interspeech 2015*, 2015, pp. 2087–2091.