# DG-MIX: DOMAIN GENERALIZATION FOR ANOMALOUS SOUND DETECTION BASED ON SELF-SUPERVISED LEARNING

*Ismail Nejjar*[1,2]*, Jean Meunier-Pion*[1,3]*, Gaetan Frusque*[1]*, Olga Fink*[1]*,*

[1] EPFL, IMOS, Lausanne, Switerzland, {ismail.nejjar, gaetan.frusque, olga.fink}@epfl.ch
[2] ETH Zürich, Chair of Intelligent Maintenance Systems, Zürich, Switerzland, {inejjar}@ethz.ch
[3] CentraleSupélec, Gif-sur-Yvette, France, {jean.meunier-pion}@student-cs.fr

## ABSTRACT

Detecting anomalies in sound data has recently received significant attention due to the increasing number of implementations of sound condition monitoring solutions for critical assets. In this context, changing operating conditions impose significant domain shifts resulting in performance drops if a model trained on a set of operating conditions is applied to a new operating condition. An essential challenge is distinguishing between anomalies due to faults and new operating conditions. Therefore, the high variability of operating conditions or even the emergence of new operating conditions requires algorithms that can be applied under all conditions. Therefore, domain generalization approaches need to be developed to tackle this challenge. In this paper, we propose a novel framework that leads to a representation that separates the health state from changes in operating conditions in the latent space. This research introduces DG-Mix (Domain Generalization Mixup), an algorithm inspired by the recent Variance-Invariance-Covariance Regularization (VICReg) framework. Extending the original VICReg algorithm, we propose to use Mixup between two samples of the same machine type as a transformation and apply a geometric constraint instead of an invariance loss. This approach allows us to learn a representation that distinguishes between the operating conditions in an unsupervised way. The proposed DG-Mix enables the generalization between different machine types and diverse operating conditions without an additional adaptation of the hyperparameters or an ensemble method. DG-Mix provides superior performance and outperforms the baselines on the development dataset of DCASE 2022 challenge task 2. We also demonstrate that training using DG-Mix and then fine-tuning the model to a specific task significantly improves the model's performance.

***Index Terms—*** Unsupervised Learning, Mixup, Domain Generalization, Self-Supervised Learning, Anomalous Sound Detection

## 1. INTRODUCTION

Anomalous sound detection (ASD) is the task of identifying whether the sound emitted by a machine is normal or abnormal [1, 2, 3]. One of the main challenges of this task is to distinguish between novel operating conditions or novel background noise and real anomalies caused by a machine fault or malfunction. Moreover, the sound emitted by machines of the same type but operated differently or installed under different conditions may differ significantly. Deep learning models have recently demonstrated excellent performance in detecting abnormal sounds under different scenarios. The directions pursued to tackle these challenges range from Unsupervised Anomalous sound detection [4, 5, 6], where only normal sound samples are used for training, to domain adaptation techniques for bridging domain shifts [7]. Domain shifts are discrepancies in the acoustic signals between a source and a target domain mainly caused by differences in machine operating conditions or ambient noise. The shifts result in performance drops if a model trained on a set of operating conditions is applied to a new operating condition.

In real-world applications, the background noise of the machine can be affected by various sound sources surrounding the machine. Therefore, it is difficult to identify the distinct causes of the changes and attribute them to the domain shift. Consequently, it is necessary to develop a method that can be generalized to different changes in operating conditions without relying on the detection of domain shifts.

In this paper, we propose a novel algorithm DG-Mix (Domain Generalization Mixup), which aims to learn representations that distinguish between the different operating and recording conditions in an unsupervised manner. Three objectives are thereby pursued: (1) reveal the impact of attributes on the data by enforcing embeddings in the same batch to be different, (2) obtain uncorrelated embedding features containing specific information, (3) respect defined geometrical constraints between the different domains. We also investigate how self-supervised learning pre-training helps our model to learn more robust and more general representations that generalize across various operating conditions for different machine types, and are robust to different noise levels and noise types. To this end, our proposed algorithm is compared on the one hand to another popular self-supervised approach, VICReg, and on the other hand to our proposed model trained from scratch. We evaluated the proposed approach and submitted it to task 2, "Unsupervised anomalous sound detection (ASD) for machine condition monitoring applying domain generalization techniques" of the DCASE challenge 2022 [8]. In experimental evaluations, it is shown that the proposed technique significantly outperforms both baseline approaches but also the originally proposed VICReg on the source and target domains of the development set.

## 2. DATASET

The dataset of this task was generated from the MIMII DG [9] (Malfunctioning Industrial Machine Investigation and Inspection for Domain Generalization) and ToyADMOS2 [10] (Anomaly Detection in Machine Operating Sounds) datasets consisting of normal and anomalous operating sounds of seven types of toy/real machines. ToyCar and ToyTrain machine types are extracted from ToyADMOS2 dataset while fan, gearbox, bearing, slide rail, and valve are extracted from MIMII DG dataset. Each recording is a singlechannel 10-second audio signal sampled at 16 kHz. The signals originate from a mixture between machine sounds and environmental noise samples of several real-world factories. Each machine type contains six sections. The dataset is divided into a development set of three sections and an evaluation set consisting of the other three sections. For each section, there are 1000 training samples, including 990 source samples and ten target samples. In this research, we use all the training data in the development dataset and the additional training dataset for training the models.

## 3. METHOD

We propose a new framework inspired by the Variance-Invariance-Covariance Regularization (VICReg), a self-supervised algorithm proposed in [11]. We give an overview of our proposed approach in Figure 1. The framework is composed of (1) a self-supervised learning algorithm, (2) a subsequent fine-tuning step and (3) an anomaly detection phase based on k-Nearest Neighbors (k-NN) [12].

The objective of the self-supervised task is to learn an encoder that provides meaningful representations of audio samples. This pre-trained encoder is then used and fine-tuned to perform supervised classification of the section ID. Finally, we use the embeddings from the encoder to compute an anomaly score with k-NN.

### 3.1. Audio Prepocessing

To simplify the task, all audio samples are provided with only one channel. Each audio sample is then transformed into a log-mel spectrogram. The input given to our model is a two-dimensional image-like feature $X \in \mathcal{R}^{P \times F}$. The frame size of the Short-Time Fourier Transform (STFT) is 64 ms, and the hop size is 32 ms. We also set the number of Mel bins $F$ to 128. The number of frames of the context window $P$ is fixed to 64. The context window is shifted by $L$ frames resulting in $B$ extracted images, with $B = \lceil \frac{T-P}{L} \rceil$ with $L = 8$. Given the previous parameters, the total spectrogram size $T$ is equal to 313.

### 3.2. CNN Architectures

We used the MobileNetV2 [13] backbone trained from scratch in this work. The off-the-shelf Pytorch [14] implementation of MobileNetV2 is used. The width multiplier parameter is set to 0.5, and the last layers are adapted to obtain a 320-dimensional vector per input image. Table 1 provides a detailed summary of the applied CNN architecture, with a total parameter number of 1.45M.

### 3.3. DG-Mix : Self-Supervised Pre-training

The first step of our approach is based on a self-supervised learning algorithm inspired by VICReg [11]. This framework provides a good feature representation for image classification problems.

**Background:** VICReg is an algorithm based on Siamese networks. VICReg aims to prevent a collapse by regularising the variance and covariance of the network outputs. The objective function of VICReg contains three main terms: a variance term, an invariance term, and a covariance term.

1. **Variance**: Regularization term that prevents mode collapse
2. **Covariance**: Regularization term that prevents dimensional collapse.
3. **Invariance**: Similarity metric to be minimized between two augmented views of the same source image.

| Input | Operator | t | c | n | s |
|---|---|---|---|---|---|
| $128 \times 64 \times 3$ | conv2d 3×3 | - | 16 | 1 | 2 |
| $64 \times 32 \times 16$ | bottleneck | 1 | 8 | 1 | 1 |
| $64 \times 32 \times 8$ | bottleneck | 6 | 16 | 2 | 2 |
| $32 \times 16 \times 16$ | bottleneck | 6 | 16 | 3 | 2 |
| $16 \times 8 \times 16$ | bottleneck | 6 | 32 | 4 | 2 |
| $8 \times 4 \times 32$ | bottleneck | 6 | 48 | 3 | 1 |
| $8 \times 4 \times 48$ | bottleneck | 6 | 80 | 3 | 2 |
| $4 \times 2 \times 80$ | bottleneck | 6 | 160 | 1 | 1 |
| $4 \times 2 \times 160$ | conv2d 1×1 | - | 320 | 1 | 1 |
| $4 \times 2 \times 320$ | conv2d 4×2 | - | 320 | 1 | 1 |
| $1 \times 1 \times 320$ | conv2d 1×1 | - | 320 | 1 | |

Table 1: Modified MobileNetV2 architecture used for all experiments. Each row represents the sequence of layers, repeated n times, with c channels, and stride s

**Proposed Approach:** Figure 1 provides an overview of DG-MIX. The proposed loss comprises three parts. The last two terms correspond to the variance and covariance losses presented in the VICReg [11] implementation, while we propose to substitute the third part (representing the invariance criterion in VICReg) with a term that enforces the embedding features of the source and target domains to be distinguishable. Thereby, we are able to learn a specific representation for both domains. For each machine type, we train a Siamese architecture where the three branches are similar and share the same weights. Each branch is composed of an encoder $f_\theta$ which corresponds to the modified MobileNetV2 presented in Table 1, followed by an expander $h_\phi$. The expander is composed of three fully-connected layers of size 1280. Each of the layers is followed by a batch normalization layer [15] and a ReLU [16] activation function.

In order to mitigate the gap between the source and target domains, we propose to extend the VICReg framework [11] by using Mixup [17] to augment the target domain. Furthermore, a novel loss is proposed to take into account the added Mixup branch, acting as a regularization term and improving domain generalization.

While in the original VICReg approach, a data augmentation approach is applied, we propose to impose similarity between the sample representations. Given all the $S$ log Mel-spectrograms from both the source and target domains of all sections for each machine type, two different samples, $X$ and $X'$ are selected. For each such pair of samples, a linear combination with respect to $\lambda$ is obtained. This combination gives rise to a new sample denoted as $X_\lambda$. Formally, $\lambda$ is a realization of a beta distribution $Beta(\alpha, \beta)$ and represents the mixup rate. In our case we set $\alpha = \beta = 0.5$.

First $X, X'$ and $X_\lambda$ are encoded by $f_\theta$ resulting in $Y, Y'$ and $Y_\lambda$, and then mapped by the expander on the embeddings, $Z, Z'$ and $Z_\lambda$. The loss is composed of three terms and computed at the embedding level on $Z$, $Z'$ and $Z_\lambda$. For a batch of size $N$, we denote $Z = [z_1, ..., z_N]$, with $z \in \mathcal{R}^D$ and $D$ the expander dimension.

The proposed consistency term seeks to generate new virtual domains. A linear interpolations of feature vectors should lead to linear interpolations of their corresponding domain. Therefore the proposed loss aims at minimzing the distance between the embedding vector of $X_\lambda$ and the linear combination of embeddings of $X$ and $X'$:

$$s(Z_\lambda, Z, Z', \lambda) = \frac{1}{N} \sum_{i=1}^{N} \| z_{\lambda,i} - (\lambda z_i + (1-\lambda) z'_i) \|_2^2 \qquad (1)$$

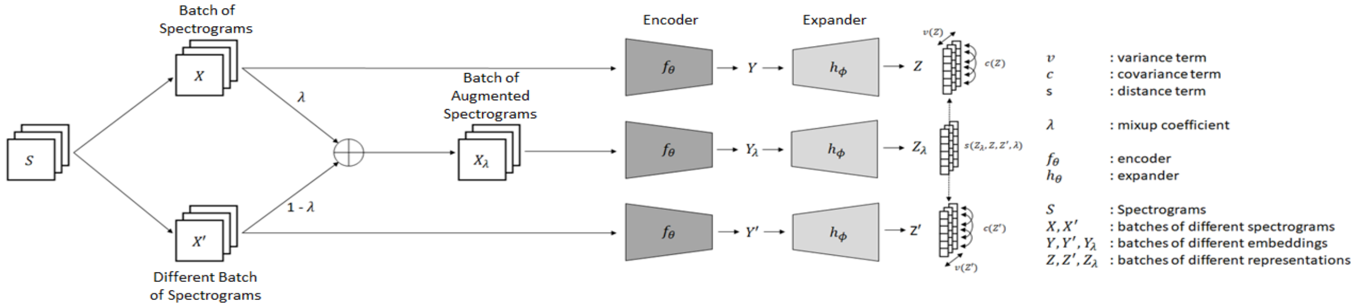


Figure 1: DG-Mix : Self-Supervised Framework for Domain Generalization

The second loss term forces the variance inside each batch to be equal to 1, preventing a mode collapse:

$$v(Z) = \frac{1}{D}\sum_{i=1}^{D} max(0, 1 - S(z^j)) \tag{2}$$

where $z^j$ is the j$^{th}$ row of the matrix $Z$ in the batch, and $S(z) = \sqrt{Var(z)}$ is the standard deviation.

Finally, the last loss term aims to learn uncorrelated features for each embedding, by forcing the off-diagonal elements to be zero, resulting in a rich embedding. The covariance matrix is defined as:

$$C(Z) = \frac{1}{N-1}\sum_{i=1}^{N}(z_i - \bar{z})(z_i - \bar{z})^T \tag{3}$$

with $\bar{z} = \frac{1}{N}\sum_{i=1}^{N} z_i$ representing the mean embedding over a mini-batch.

$$c(z) = \frac{1}{d}\sum_{i \neq j}[C(Z)]^2_{i,j} \tag{4}$$

The final loss used to train the model is:

$$\gamma s(Z_\lambda, Z, Z', \lambda) + \mu(v(Z) + v(Z')) + \nu(c(Z) + c(Z')) \tag{5}$$

$\gamma, \mu, \nu$ are hyper-parameters that we set in this report to $25, 25, 1$ respectively. The networks are trained using the Large Batch Training of Convolutional Networks (LARS) [18] optimizer, with a learning rate of $0.8$, weight decay of $10^{-4}$ and a batch size of 1024 for 100 epochs. In addition, ten warmup epochs were used, and the learning rate followed a cosine decay schedule starting from 0 and finishing at $0.002$. The expander size and the hyperparameters were not finetuned on the task but rather taken as reported in the VICReg paper. After the pre-training, only the encoder model was used for the downstream classification task presented in the next section.

### 3.4. Fine-tuning on Section ID Classification

Similar to the baseline method, our proposed model is fine-tuned to identify the section ID of an audio sample.

The pre-trained encoder is used, and a classifier composed of two fully connected layers (320-128-6) is added. To improve the robustness of the model, a mixup strategy on the source and target data for each section is used to generate augmented data of intra-domain and inter-domain samples.

The KL-divergence loss between the classifier output and the mixed section ID is used for this task along with the geometrical constraint presented in equation 1 as a regularization term. However, this time it is directly applied to the encoder. Finally, the networks are fine-tuned using AdamW [19] optimizer, with a learning rate of $10^{-4}$, weight decay of $10^{-4}$, and a batch size of 64.

### 3.5. Anomaly Detection

After the fine-tuning step, we apply a k-NN algorithm [12] to compute the anomaly score. We use the mean embedding vector from the 10-s audio recording as input feature to the k-NN algorithm.

We used the Euclidean metric as the anomaly score, and the number of nearest neighbors was set to 1. In other words, the larger the distance from the training embeddings, the more abnormal the sample is.

## 4. RESULTS

We tested our model on the development set of the DCASE 2022 Task 2 described in section 2. We first present how our proposed method without pre-training already outperforms the baseline. The main differences between the baseline and our proposed method are (1) change of the CNN, (2) use of Mixup for data augmentation, (3) addition of a regularization term from the Mixup transformation, and (4) use of k-NN for anomaly score computation. Table 2 displays the harmonic means of the AUC Source, AUC Target, and pAUC computed over all three sections for each machine type using the baselines and the proposed method without pre-training. The harmonic means of the AUC Source, AUC Target, and pAUC are also reported for each method.An absolute improvement of more than $10\%$ of the baseline approach is obtained.

The importance of pre-training becomes apparent from the results reported in Table 3 where the performance of two self-supervised approaches, VICReg using SpecAugment [20] and the proposed DG-Mix method is compared. Moreover, we use the SpecAugment procedure for VICReg because Mixup is not adapted for this method. VICReg provides an improvement of almost $1\%$ in terms of the overall harmonic mean and DG-Mix has a performance gain of $4\%$, suggesting that they both successfully helped to get better and more robust representations.

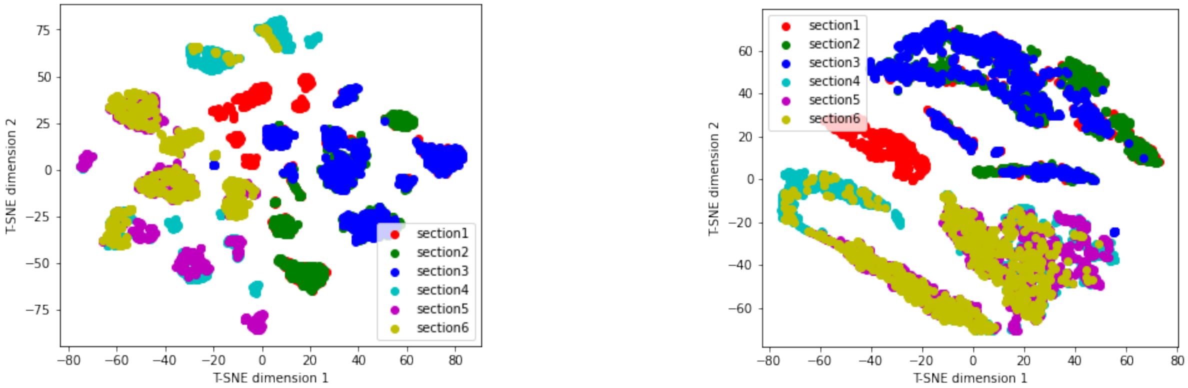
Pre-training with DG-Mix outperforms by $3\%$ the pre-training obtained using VICReg. This is due to Mixup and its regularization term that provide a richer representation of the dataset. This is illustrated in Figures 2a and 2b presenting the t-SNE [21] plots of the embeddings obtained from a ToyCar training sample with DG-Mix and VICReg. There are more clusters with DG-Mix, showing more granularity and a better separation of the dataset into different modalities.

| | Baseline MobileNet-V2 | | | | Proposed anomaly detection method without pre-training | | | |
|---|---|---|---|---|---|---|---|---|
| Machine | AUC Source | AUC Target | pAUC | Harmonic Mean | AUC Source | AUC Target | pAUC | Harmonic Mean |
| ToyCar | 58,97 | 52,26 | 52,39 | 54,37 | **79,82** | **73,62** | **60,11** | **70,18** |
| ToyTrain | **58,59** | 46,07 | **51,56** | 51,57 | 46,45 | **61,25** | 51,43 | **52,36** |
| Bearing | **62,88** | 61,81 | **57,35** | **60,58** | 59,24 | **69,06** | 50,13 | 58,47 |
| Fan | 71,35 | 48,53 | 57,10 | 57,54 | **88,85** | **70,27** | **69,45** | **75,22** |
| Gearbox | 69,98 | 56,60 | 56,18 | 60,29 | **79,83** | **70,32** | **60,78** | **69,44** |
| Slider | 66,03 | 40,72 | 54,77 | 51,76 | **90,70** | **66,02** | **64,89** | **72,15** |
| Valve | 67,75 | 58,01 | **62,70** | 62,57 | **71,74** | **65,99** | 59,05 | **65,18** |
| **Overall** | 64.73 | 51.06 | 55.80 | 56.65 | **72.71** | **70.93** | **62.27** | **68.32** |

Table 2: Results for the MobileNet-V2 baseline and our proposed anomaly detection method without pre-training (in %).

| | VICReg with SpecAugment | | | | DG-Mix | | | |
|---|---|---|---|---|---|---|---|---|
| Machine | AUC Source | AUC Target | pAUC | Harmonic Mean | AUC Source | AUC Target | pAUC | Harmonic Mean |
| ToyCar | 85.09 | 78,68 | 59,05 | 72.47 | **93.28** | **81.73** | **68.32** | **79.80** |
| ToyTrain | 48,34 | **56,20** | **53,69** | 52,54 | **52.50** | 55.48 | 53.20 | **53.70** |
| Bearing | **77,99** | 70,49 | **73,22** | **73,77** | 66.99 | **82.90** | 64.75 | 70.70 |
| Fan | **83,83** | **77,17** | 65,93 | 74,9 | 83.37 | 76.73 | **70.74** | **76.60** |
| Gearbox | **94,99** | 68,77 | 68,30 | 75,55 | 88.96 | **82.74** | **71.22** | **80.28** |
| Slider | **94,99** | **68,74** | 68,25 | **75,51** | 94.52 | 67.35 | **68.91** | 75.11 |
| Valve | 75,02 | 66,48 | 60,60 | 66,86 | **85.09** | **81.93** | **71.65** | **79.13** |
| **Overall** | 76.44 | 68.78 | 63.53 | 69.18 | **77.55** | **74.08** | **66.33** | **72.34** |

Table 3: Results obtained when pre-training with VICReg compared to DG-Mix (in %).



(a) t-SNE plot of the embedding result when Pre-training with DG-Mix

(b) t-SNE plot of the embedding result when Pre-training with VICReg

However, SpecAugment is less suited than Mixup for anomaly detection because it uses time warping, frequency masking, and time masking transformations that can interfere with the anomaly patterns. In contrast, Mixup does not assume anything about possible anomalies but only mixes samples. Using Mixup between and across multiple domains [22] allows us to sample the augmented training data from the heterogeneous Mixup distribution and get a more robust feature extractor at the end, which improves the results.

## 5. CONCLUSION

In this paper, we proposed a novel sound anomaly detection framework for domain generalization composed of a self-supervised algorithm followed by a supervised task using Mixup on the input log-Mel spectrograms. In this work, our goal was to develop a unified framework that is robust and performs well across all machine types. We used the same hyperparameters for each machine type to achieve this goal. Experimental evaluation shows that the proposed approach significantly outperforms all baseline approaches. In addition, we demonstrated that pre-training an encoder improved the generalization ability of this encoder. Extending the framework to other tasks and datasets is left for future research.

## 6. ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation under Grant PP00P2_176878

## 7. REFERENCES

[1] Y. Koizumi, S. Saito, H. Uematsu, Y. Kawachi, and N. Harada, "Unsupervised detection of anomalous sound based on deep learning and the neyman–pearson lemma," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 212–224, 2018.

[2] Y. Kawaguchi, R. Tanabe, T. Endo, K. Ichige, and K. Hamada, "Anomaly detection based on an ensemble of dereverberation and anomalous sound extraction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2019, pp. 865–869.

[3] 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).* IEEE, 2020, pp. 271–275.

[4] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, *et al.*, "Description and discussion on dcase2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2006.05822*, 2020.

[5] R. Giri, F. Cheng, K. Helwani, S. V. Tenneti, U. Isik, and A. Krishnaswamy, "Group masked autoencoder based density estimator for audio anomaly detection," in *Detection and Classification of Acoustic Scenes and Events Workshop 2020*, 2020.

[6] K. Wilkinghoff, "Sub-cluster adacos: Learning representations for anomalous sound detection," in *2021 International Joint Conference on Neural Networks (IJCNN).* IEEE, 2021, pp. 1–8.

[7] 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 sound detection for machine condition monitoring under domain shifted conditions," *arXiv preprint arXiv:2106.04492*, 2021.

[8] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, 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 arXiv e-prints: 2206.05876*, 2022.

[9] 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.

[10] 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.

[11] A. Bardes, J. Ponce, and Y. LeCun, "Vicreg: Variance-invariance-covariance regularization for self-supervised learning," 2022. [Online]. Available: https://hal.inria.fr/hal-03541297/file/vicreg_iclr_2022.pdf

[12] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," *International Statistical Review/Revue Internationale de Statistique*, vol. 57, no. 3, pp. 238–247, 1989.

[13] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2018. [Online]. Available: https://arxiv.org/abs/1801.04381

[14] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[15] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[16] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," *arXiv preprint arXiv:1803.08375*, 2018.

[17] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond Empirical Risk Minimization," *CoRR*, vol. abs/1710.09412, 2017. [Online]. Available: http://arxiv.org/abs/1710.09412

[18] Y. You, I. Gitman, and B. Ginsburg, "Scaling SGD Batch Size to 32k for ImageNet Training," *CoRR*, vol. abs/1708.03888, 2017. [Online]. Available: http://arxiv.org/abs/1708.03888

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017. [Online]. Available: https://arxiv.org/abs/1711.05101

[20] 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," *arXiv preprint arXiv:1904.08779*, 2019.

[21] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[22] Y. Wang, H. Li, and A. C. Kot, "Heterogeneous domain generalization via domain mixup," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2020, pp. 3622–3626.