



Article

Regularized Contrastive Masked Autoencoder Model for Machinery Anomaly Detection Using Diffusion-Based Data Augmentation

Esmaeil Zahedi ¹, Mohamad Saraee ^{2,*}, Fatemeh Sadat Masoumi ³ and Mohsen Yazdinejad ¹

¹ Faculty of Computer Engineering, University of Isfahan, Isfahan 8174673441, Iran; zahedi@eng.ui.ac.ir (E.Z.); mohsen.yazdinejad@eng.ui.ac.ir (M.Y.)

² School of Science, Engineering and Environment University of Salford, Manchester M5 4WT, UK

³ Faculty of Mathematics, Statistics and Computer Science, Allameh Tabataba'i University, Tehran 1485643449, Iran; fatemeh_masoumi@atu.ac.ir

* Correspondence: m.saraee@salford.ac.uk; Tel.: +44-78-0370-7440

Abstract: Unsupervised anomalous sound detection, especially self-supervised methods, plays a crucial role in differentiating unknown abnormal sounds of machines from normal sounds. Self-supervised learning can be divided into two main categories: Generative and Contrastive methods. While Generative methods mainly focus on reconstructing data, Contrastive learning methods refine data representations by leveraging the contrast between each sample and its augmented version. However, existing Contrastive learning methods for anomalous sound detection often have two main problems. The first one is that they mostly rely on simple augmentation techniques, such as time or frequency masking, which may introduce biases due to the limited diversity of real-world sounds and noises encountered in practical scenarios (e.g., factory noises combined with machine sounds). The second issue is dimension collapsing, which leads to learning a feature space with limited representation. To address the first shortcoming, we suggest a diffusion-based data augmentation method that employs ChatGPT and AudioLDM. Also, to address the second concern, we put forward a two-stage self-supervised model. In the first stage, we introduce a novel approach that combines Contrastive learning and masked autoencoders to pre-train on the MIMII and ToyADMOS2 datasets. This combination allows our model to capture both global and local features, leading to a more comprehensive representation of the data. In the second stage, we refine the audio representations for each machine ID by employing supervised Contrastive learning to fine-tune the pre-trained model. This process enhances the relationship between audio features originating from the same machine ID. Experiments show that our method outperforms most of the state-of-the-art self-supervised learning methods. Our suggested model achieves an average AUC and pAUC of 94.39% and 87.93% on the DCASE 2020 Challenge Task2 dataset, respectively.



Citation: Zahedi, E.; Saraee, M.; Masoumi, F.S.; Yazdinejad, M. Regularized Contrastive Masked Autoencoder Model for Machinery Anomaly Detection Using Diffusion-Based Data Augmentation. *Algorithms* **2023**, *16*, 431. <https://doi.org/10.3390/a16090431>

Academic Editor: Frank Werner

Received: 29 July 2023

Revised: 19 August 2023

Accepted: 31 August 2023

Published: 8 September 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Maintaining and supervising the performance of industrial machines is an important task in every industry. By effectively analyzing machine behavior, sudden faults can be anticipated and prevented. These analyses can be performed either by experts or by automatic processes. Automatic anomaly sound detection has received significant attention in the research community and numerous methods have been proposed, including one-class classifiers [1], autoencoders [2,3], Contrastive learning and self-supervised learning (SSL) [4–6]. These methods are typically trained on datasets that primarily consist of normal sounds [2–12]. Thus, a major challenge in designing an anomaly detector lies in finding a more informative data representation that can effectively differentiate between normal

and anomalous sounds. Furthermore, while real-world environments are characterized by domain shifting, diverse noise sources, and various anomalies, there are not any robust data sets for real-world scenarios.

In response to the optimal data representation problem, deep learning-based approaches leverage audio augmentation techniques to enhance the model's representation ability [13,14]. However, most existing audio augmentation methods, such as SpecAugment [15], Time Stretching and Pitch Shifting [16], Noise Injection [17], Reverberation [18], and Time and Frequency Masking [19], have shown limited effectiveness, particularly in modifying high-level semantic attributes in augmented data, such as introducing new environmental sounds or domain-specific characteristic, because these methods often lack diversity and generality along with vital semantic dimensions. In this regard, one popular alternative approach is the use of Generative models, particularly Generative Adversarial Networks (GANs) [20], which have been widely employed for generating synthesized data in computer vision tasks [21,22]. Synthetic data generated from GANs has demonstrated benefits in representation learning and training classifiers in various applications [20,23–27]. Another option in the case of having limited real data is utilizing diffusion models [28]. Although diffusion-based data augmentations have primarily focused on image modality, there is an opportunity to extend their application to audio data. Motivated by this research gap, in this study, we propose a novel diffusion-based data augmentation method for enhancing the generality and diversity of audio datasets in the context of anomaly sound detection. This method leverages textual prompts from ChatGPT [29], which is a Large Language Model (LLM) developed by OpenAI, as input to AudioLDM [30], which is a text-to-audio diffusion model and produces high-quality audio samples.

The major challenge in anomaly sound detection is choosing a proper method to improve the performance of this task; earlier works in this field focused on the supervised detection of anomalies [1]. However, a significant problem with supervised anomaly detection is the assumption that all possible anomalous sounds are accessible during the training phase, which may not be feasible in many practical applications. In response to this, unsupervised approaches have been widely explored and in the DCASE 2020 task 2 challenge [2–31] they have showcased promising results [9,32]. Researchers used the Hidden Markov Model (HMM) [33], Gaussian Mixture Model (GMM) [34], Non-negative Matrix Factorization (NMF) [35], and deep learning-based approaches, such as autoencoders and Generative models [36], to model and analyze normal sounds. By training these models to compress and reconstruct normal sounds, the underlying properties of normal sounds were learned effectively in the latent space, and when an abnormal sample was introduced into these models, the resulting large reconstruction errors indicated its deviation from the learned normal sound distribution. Consequently, unsupervised deep learning-based methods have demonstrated their effectiveness in improving the performance of sound anomaly detection by obtaining better latent space features of normal sounds. Among these methods, self-supervised techniques have emerged as one of the most efficient approaches [37,38].

In recent years, self-supervised learning has gained significant attention for its effectiveness in representation learning. This learning method can be divided into Contrastive and Generative learning. Contrastive learning's objective is to unveil distinctive audio features by comparing positive and negative samples in a high-dimensional space, and Generative self-supervised learning focuses on learning audio representation through the reconstruction of audio signals. Several kinds of research have been proposed to leverage the SSL method of learning for audio representation learning and anomalous sound detection. Examples include AADCL [2], which uses Contrastive learning to learn audio representation; STgram-MFN [39], a spectral-temporal fusion-based self-supervised method; Glow Aff [8] a flow-based self-supervised method; GMADE [3], which uses an ensemble of Group Masked Autoencoders for Density Estimation; CAV-MAE [40], a Contrastive Audio-Visual Masked Autoencoder; and CLP-SCF [37], a Contrastive learning method with self-supervised classification-based fine-tuning.

Contrastive learning trains the encoder to be invariant to semantics-preserving data variations, while masked autoencoders (MAEs) [41] focus on learning spatial statistical correlations. Furthermore, MAE methods treat each sample independently in the loss function, whereas Contrastive methods explicitly consider the relationships between all samples in a batch by adjusting embedding distances. Given these differences, we posit that these two approaches are complementary [42], which leads to extracting different discriminative features from a given input. On the other hand, the generalization capability of Contrastive learning is influenced by four crucial factors [13]. The first one is positive sample alignment, which indicates to how well similar audio samples are grouped together in the learned representation space. The second factor is class center divergence, which measures the separation between different classes of audio samples to facilitate better discrimination. The third factor is related to feature collapse prevention, which is essential for learning informative vectors without redundancies. These first three factors mainly pertain to the properties of the learned audio representations. The fourth factor that impacts the generalization capability is the quality of augmented data [13] used during training. Data augmentation techniques play a vital role in creating diverse and representative augmented data, enhancing the model's ability to generalize to unseen samples. To meet the requirement for factor one, we suggest a combined method, aiming to leverage the strengths of both Contrastive learning and masked autoencoders to enhance the model's representational power. To address factor number three, we introduce Variance—Covariance Regularization [43], which is a regularization term added to the Contrastive and MAE loss functions. Additionally, in order to improve factor number four, we incorporate a diffusion-based data augmentation technique.

Consequently, our contributions to this research are as follows:

- (1) In response to the issue of not having a robust dataset for various real-world noises, we propose a novel diffusion-based data augmentation method. Our approach incorporates textual prompts from ChatGPT as input to AudioLDM. The use of diffusion-based data augmentation allows us to add noise or artifacts with limited knowledge of any target domain solely through text prompts.
- (2) To improve the generalization of Contrastive learning, specifically to prevent feature collapse and learn informative representation, we propose a novel learning framework that combines Contrastive learning and masked autoencoders (MAEs) with Variance—Covariance Regularization. Contrastive learning and MAEs complement each other, extracting different discriminative features from the input data. By simultaneously leveraging these two approaches, we enhance the model's representational power and better discriminate between normal and anomalous sounds.

The rest of the content of this paper is structured as follows: In Section 2, we introduce our proposed learning framework and its components. Further, the validation of our method through experimental results on Dcase 2020 challenge datasets [31,44] is presented in Section 3, and Finally, the conclusion and the scope for further research are outlined in Section 4.

2. Materials and Methods

In this section, the proposed Regularized Contrastive Masked AutoEncoder (RC_MAE) model (as shown in Figure 1) for machinery anomaly detection using diffusion-based data augmentation is presented in detail. The required material is discussed, and the proposed method is fully explained.

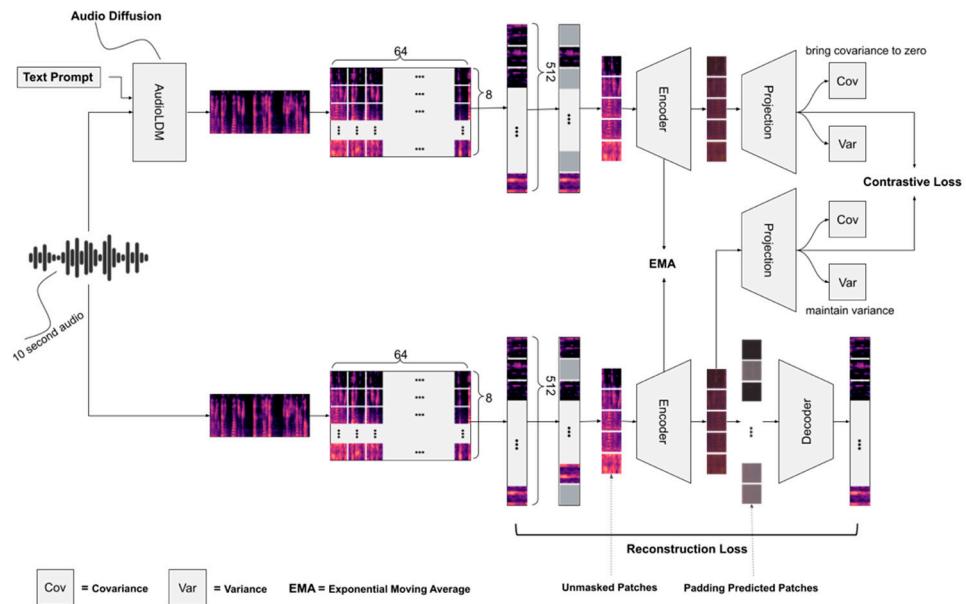


Figure 1. Proposed framework.

2.1. Data Augmentation Method

Diffusion models have proven to be highly effective in producing realistic output based on text queries (prompts) [45]. In our study, we employed AudioLDM, which is built upon a latent diffusion architecture to generate synthetic audio. We leveraged ChatGPT using prior knowledge about the target domain for generating prompts. Each prompt can consist of multiple words; for instance, “add sharp and metallic screeching sound created by metal-cutting saws”. During the model pretraining phase, both the original samples and their augmented versions, using a randomly selected prompt from ChatGPT prompts, were fed into the model for pre-training purposes. The workflow of the suggested model is as follows:

To generate audio, instead of creating it from scratch, we integrated real audio into the generation process of the diffusion model, which is a reverse process with S steps. Firstly, we insert a real audio x_0^{ref} with noise $\epsilon \sim N(0, 1)$ at timestep $\lfloor S_{t_0} \rfloor$, where $t_0 \in [0, 1]$ is a hyperparameter controlling the insertion position of the audio. Then, the process of reverse diffusion is applied by utilizing the spliced audio at timestep $\lfloor S_{t_0} \rfloor$, before we proceed iteratively to apply Equation (1) until a sample is generated at timestep 0.

$$x_{\lfloor S_{t_0} \rfloor} = \sqrt{\tilde{\alpha}_{\lfloor S_{t_0} \rfloor}} x_0^{ref} + \sqrt{1 - \tilde{\alpha}_{\lfloor S_{t_0} \rfloor}} \epsilon \quad (1)$$

This conditioned generation process is guided by a prompt that includes the embedding \vec{w}_i for the prompt. We demonstrate this type of augmentation surpasses traditional augmentation methods without diffusion and helps the improvement in few-shot audio classification tasks. In this context, four key challenges associated with Generative augmentation were identified [46,47], which are as follows: prompt ambiguity—certain prompts possess multiple meanings, leading to the generation of incorrect samples; diversity—ensuring the generation of audio samples with sufficient variation poses another challenge; domain shifting—augmentation methods must demonstrate flexibility in adapting to new distributions; and finally, fidelity—Generative methods must produce high-quality samples that closely resemble real audio. In this study, we address these challenges by employing clear and diverse prompts, generating high-quality samples.

2.2. Masked Autoencoder

The masked autoencoder (MAE) [41] is a straightforward yet powerful self-supervised technique that can be applied to audio data [40]. In this approach, input data and their

augmentation are represented as $a = [a_1, a_2, \dots, a_{512}]$ and $aug = [aug_1, aug_2, \dots, aug_{512}]$ tokens. A portion of the input tokens are then masked, and only the unmasked tokens are fed into a Transformer-based encoder–decoder model. The primary objective of the MAE is to reconstruct the masked tokens while minimizing the Mean Square Error (MSE) as defined as Loss of MAE in Equation (2),

$$L_{MAE} = \frac{1}{n} \sum_{i=1}^n \left(\frac{(a_{mask}^i - f_\theta(a_{mask}^i))^2}{\# \text{ of masked audio patches}} \right) \quad (2)$$

where n represents the mini-batch size, and $a_{mask}^i, f_\theta(a_{mask}^i)$ refers to the original and predicted masked patches, respectively. The loss calculation is performed only on the masked portion of the input. Through minimizing this loss, the model learns a meaningful representation of the input audio data. The MAE approach offers several advantages [48]. Firstly, it simplifies the training pipeline by using the original input as the prediction target. Secondly, by only inputting the unmasked tokens to the encoder and employing a high masking ratio, MAEs significantly reduce computational overhead. Lastly, MAEs have demonstrated strong performance in audio representation learning [48].

2.3. Contrastive Learning

Contrastive learning helps the model extract relevant features by analyzing the similarities and differences between data points. It uses the principle that similar instances should be closer in the learned space, and dissimilar ones should be farther apart. By contrasting positive and negative examples, the model learns to differentiate between classes or categories to discover patterns and structure in the data, improving the model's performance. In this paper, we used the output of the audio and augmented audio, c_i^a and c_i^{aug} , for Contrastive learning; the contrastive loss L_c can be seen in Equation (3),

$$L_c = -\frac{1}{N} \sum_{i=1}^N \log \left[\frac{\exp(\frac{sim_{i,a}}{\tau})}{\sum_{k \neq i} \exp(\frac{sim_{i,k}}{\tau}) + \exp(\frac{sim_{i,a}}{\tau})} \right] \quad (3)$$

where $sim_{i,a} = \|c_i^a\|^T \|c_i^{aug}\|$ and τ is the temperature.

2.4. Variance—Covariance Regularization

Figure 1 depicts the whole proposed framework in this paper. As can be seen in this figure, Variance—Covariance Regularization is applied to both branches of the architecture separately, thereby preserving the information content of each embedding at a certain level and preventing informational collapse [43] independently for the two branches. The Variance preservation term ensures that the embedding vectors do not collapse toward zero by explicitly preventing shrinkage. It employs a hinge loss to maintain the standard deviation of each embedding dimension above a specified threshold within a batch. Equation (4) enforces differentiation between the embedding vectors of samples within the batch,

$$v(Z) = \frac{1}{d} \sum_{j=1}^d \max \left(0, 1 - \sqrt{Var(z^j)} + \epsilon \right) \quad (4)$$

where ϵ is a small scalar to prevent numerical instabilities. This criterion ensures that the Variance within the current batch equals one along each dimension, thus avoiding a collapse where all inputs are mapped onto the same vector.

The Covariance criterion avoids informational collapse by tackling redundancy among the embedding variables [49]. It includes a term that minimizes the Covariances between every pair of embedding variables within a batch toward zero. Equation (5) promotes decorrelation among the variables within each embedding and avoids situations where the variables vary together or are highly correlated, thus preventing informational collapse. This regularization does not impose the constraint of weight sharing between the two branches.

Additionally, integrating Variance preservation into other self-supervised joint-embedding methods enhances training stability and improves downstream task performance [43].

$$C(Z) = \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})(z_i - \bar{z})^T, \text{ where } \bar{z} = \frac{1}{n} \sum_{i=1}^n z_i \quad (5)$$

$$c(Z) = \frac{1}{d} \sum_{i \neq j} [C(Z)]_{i,j}^2 \quad (6)$$

The Covariance regularization term in Equation (6) can be defined as the summation of squared off-diagonal coefficients in the Covariance matrix $C(Z)$ and scaled by a factor of $1/d$ where d is the dimension of vector. This regularization term helps to mitigate excessive correlation among dimensions. The purpose of this term is to enforce the off-diagonal coefficients of the Covariance matrix $C(Z)$ to approach zero. Doing so aims to reduce the correlation between different dimensions of the embeddings and prevent them from encoding redundant information. Equation (7) is the average of the Variance and Covariance terms as an overall VCR loss function.

$$L_{VCR} = \frac{1}{2}(v(Z) + c(Z)) \quad (7)$$

2.5. Proposed Model

In the suggested Regularized Contrastive Masked AutoEncoder (RC-MAE) model, the loss function is computed by combining the contrastive loss (L_{con}) multiplied by the weight (λ_{con}), the Variance—Covariance loss (L_{VCR}) multiplied by the weight (λ_{VCR}), and the reconstruction loss (L_{MAE}) multiplied by the weight (λ_{MAE}). Hence, the overall loss, denoted as L_{RC-MAE} , can be seen in Equation (8),

$$L_{RC-MAE} = \lambda_{Con} L_{Con} + \lambda_{VCR} L_{VCR} + \lambda_{MAE} L_{MAE} \quad (8)$$

where λ_{Con} , λ_{VCR} , and λ_{MAE} are the hyperparameters of contrastive loss, Variance—Covariance loss, and reconstruction loss. In the proposed method, due to the scale of the gradient of (L_{con}) being larger than (L_{MAE}), we empirically find that performance is robust when we set the mentioned hyperparameters to 0.3, 0.3, and 0.4, respectively.

2.5.1. Implementation Detail

Firstly, in the context of audio augmentation, AudioLDM [30] was employed, which is a diffusion-based text-to-audio model. This model takes real audio as the initial input and a text prompt as a condition to generate augmented audio. The process is illustrated in Figure 2, where different text prompts are employed to generate different variations of augmented audio. To make these textual prompts, we utilized ChatGPT [29], as outlined in Table 1. The combination of AudioLDM and ChatGPT enables the generation of diverse and customized augmented audio samples for enhanced audio data augmentation. In this research, we empirically introduce a set of hyperparameters for text-to-audio diffusion models that control the diversity of the generated audio. The specific values for these hyperparameters can be found in Table 2.

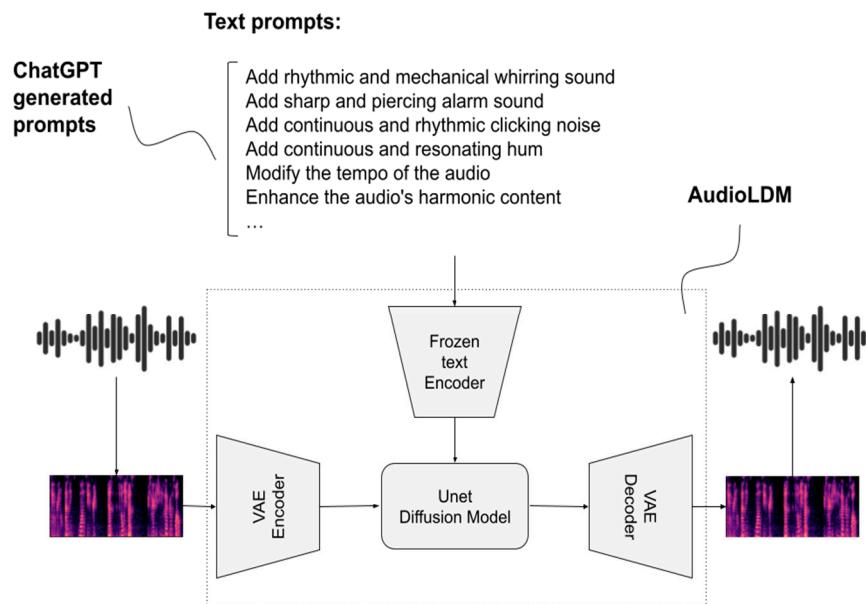


Figure 2. Scheme of the suggested method for audio data augmentation.

Table 1. Samples of retrieved prompts from ChatGPT.

ChatGPT Prompts: Different Factory Noises
“add sharp and metallic screeching sound created by metal-cutting saws in a fabrication workshop as they cut through various metal materials”
“add rhythmic and pulsating buzzing sound produced by printing presses in a publishing house as they rapidly print and collate pages of books and magazines”
“add continuous and powerful rumbling noise accompanied by the grinding and crushing of rocks in a mining quarry as large crushers pulverize rocks into smaller fragments”
“add continuous and rhythmic clicking noise produced by injection molding machines in a plastic manufacturing facility as they shape molten plastic into various products”
“add rhythmic and repetitive whirring noise accompanied by the sound of cutting tools in a woodworking shop as they shape and carve wood materials”
“add continuous and powerful whirring noise accompanied by the sound of pneumatic tools in an automobile assembly line as vehicle parts are fastened”

Table 2. AudioLDM hyperparameters.

Hyperparameter Name	Value
The sampling step for DDIM (DDIM_STEPS)	Random (10, 50)
The duration of the samples (DURATION)	10 s
Random seed (SEED)	Random for each sample
Text prompt for audio generation (TEXT)	add (noises) exist in (factory)
Guidance scale (GUIDANCE_SCALE)	2.5

With regard to the proposed model in this paper, when working with a pair of original audio and diffusion-based augmented audio, they were initially pre-processed to a log-scaled Mel spectrogram with a sampling frequency of 16,000 Hz, window size of 1024, hop size of 512, and Mel-spaced frequency bins 128 in the range 50–8000 Hz. Then, they were tokenized as $a = [a_1, a_2, \dots, a_{512}]$ and $aug = [aug_1, aug_2, \dots, aug_{512}]$. The next step involved masking a specific portion (50–75%) of the original and augmented audio data, denoted as a and aug , respectively. In the subsequent step, we exclusively inputted

the unmasked tokens, represented as a_{unmask} , into an audio encoder, resulting in the output a'_{unmask} . The encoder and decoders are based on ViT [14], a model for image classification that employs a Transformer-like architecture over patches of the image, Ref [50] with a depth of 12 layers and 6 heads. The proposed method comprises two branches. In the first and second branches, the outputs of the encoder f_θ are mapped onto the embeddings R^{512} -dimensional space, through linear projection layers $z_a = proj(f_\theta(a))$ and $z_{aug} = proj(f_\theta(aug))$, then a Variance—Covariance regularization calculated separately on each z_a and z_{aug} and NCE contrastive loss is also computed based on both z_a and z_{aug} . In the next step, the predicted a'_{unmask} unmasked tokens are padded with masked tokens, while preserving their original positions. This creates a transformed representation a' , which is then fed to the decoder to generate the reconstructed audio (denoted as \hat{a}); finally MAE reconstruction loss is calculated for a and \hat{a} . Therefore, the training objective is minimizing a combinatorial loss (5) that comprises two different losses: (1) and (2), which are regularized by (3). The optimization process helps the model to accurately reconstruct the input audio by simultaneously making similar samples closer to each other and making dissimilar ones farther from the similar ones. This approach ensures a diverse audio representation while reducing correlations in the embedding space. In the pre-training phase, we empirically used the Adam optimizer [18] with a learning rate of 0.0001 for model optimization. The batch size and number of epochs were set to 128 and 400, respectively. The temperature score τ in Equation (2) was empirically chosen as 0.05 based on [12]. Each model was trained on 2 NVIDIA 3090Ti GPUs using distributed data-parallel training.

2.5.2. Fine-tuning Stage

During the pre-training phase, the proposed self-supervised model was trained on audio data from various machines to learn audio representations; we did not utilize the metadata associated with the audio files, such as label or machine type and ID, which can provide valuable information about the machines' states or properties.

To improve the learning of audio features from machines with different IDs, we employed supervised Contrastive learning methods that leverage machine IDs as labels in order to capture the relationship between audio signals and their corresponding machine IDs. This approach improves the capability to identify sounds from different IDs and strengthens the connection between sounds originating from the same ID. In this stage, we randomly selected machine sound, a_i , from a set of N input machine sounds to serve as the anchor and create contrasts with the remaining $(N - 1)$ machine sounds, and the label for each sample was determined by the associated machine ID. In the next step we derived individual sound embeddings, z_j , using a pre-trained model. To effectively capture the relationship among audio embeddings from the same ID and to differentiate audio embeddings from different IDs, we sought to maximize the cosine similarity score for embeddings from the same ID while minimizing the cosine similarity score for embeddings from distinct IDs. The Supervised Contrastive Loss (SCL) [51], defined in Equation (9), encapsulates this process,

$$L_{SCL} = -\frac{1}{N} \sum_{i=1}^N \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{j \neq i}^N \exp(z_i \cdot z_j / \tau)} \quad (9)$$

where the \cdot symbol denotes the inner dot product, τ is the temperature parameter, P_i denotes the samples that have the same machine ID, and $|P(i)|$ is the cardinality of P_i . To further enhance the distinguishing ability of the learnt audio representation, we used CrossEntropy loss (CE), so the final combinatorial loss function is as Equation (10),

$$L_{Stage2} = \lambda_{SCL} L_{SCL} + \lambda_{CE} L_{CE} \quad (10)$$

where λ_{SCL} and λ_{CE} are the hyperparameters, which show the contribution of each loss of supervised contrast loss (SCL) and CrossEntropy loss (CE). The primary aim of our proposed method is to detect anomalous sounds in the audio domain by accurately predicting the type and ID of machine sounds. To accomplish this, we calculated an

anomaly score, which is determined by taking the negative logarithm of the probability of the current machine sound and its corresponding ID. This methodology relies on the assumption that typical sounds are unlikely to be associated with an inappropriate ID. During the inference stage, if the predicted ID deviates significantly from the actual ID, the sound is classified as an anomaly.

3. Results

3.1. Dataset

During the initial pre-training phase, we used all audios in the DCASE 2020 Challenge Task2 MIMII [31] and ToyADMOS [44] datasets which contain audio samples for detecting malfunctioning industrial machinery; similar papers utilize these samples for assessing performance [2,7,8,39,52,53]. To evaluate the effectiveness of our approach, we conducted fine-tuning experiments using only the development training dataset. This dataset consists of machine-generated sounds from various sources, including four machine types (Fan, Pump, Slider, and Valve) derived from the MIMII dataset and two machine types (ToyCar and ToyConveyor) obtained from the ToyADMOS dataset. Each machine type has seven distinct variations, except for ToyConveyor, which has six variations. In total, we obtained audio recordings from 41 unique machines, with an average duration of approximately 10 s. For the fine-tuning phase, we combined the training data from the development dataset to create a comprehensive training set, with the goal of effectively identifying and classifying all machine IDs using our model. To assess the performance of our model, we utilized the test data, which included both normal and anomaly sounds, from the DCASE 2020 Challenge Task2 development dataset.

3.2. Performance Evaluation

Regarding the evaluation metrics, we utilized those suggested in the DCASE challenge (DCASE, 2020). These metrics include Area Under the ROC Curve (AUC) and partial AUC (pAUC). The formulas for calculating AUC and pAUC are provided as Equations (11) and (12), respectively,

$$AUC = \frac{1}{N_- N_+} \sum_{i=1}^{N_-} \sum_{j=1}^{N_+} H(A_\theta(x_j^+) - A_\theta(x_i^-)) \quad (11)$$

$$pAUC = \frac{1}{\lfloor pN_- \rfloor N_+} \sum_{i=1}^{\lfloor pN_- \rfloor} \sum_{j=1}^{N_+} H(A_\theta(x_j^+) - A_\theta(x_i^-)) \quad (12)$$

where $H()$ returns 1 when $x > 0$ and 0 otherwise and A_θ denotes the model output given the parameters θ . x^- and x^+ are normal and anomalous samples. N_- and N_+ are the number of normal and anomalous test samples, respectively. Based on this formula, the anomaly score of the audio samples serves as the threshold for determining if a sample is normal or anomalous, as opposed to relying on the decision results. The purpose of utilizing pAUC, as explained in [2], is to enhance the True Positive Rate (TPR) while maintaining a low False Positive Rate (FPR).

3.3. Performance Comparison

In this experimental study, we evaluate our proposed model using two distinct data augmentation strategies. The first strategy, known as the “Baseline”, implements a standard data augmentation approach involving random Time and Frequency Masking with parameters that depend on the dataset. The second strategy utilizes the audio diffusion method, where the original audio serves as the real guidance, and different prompts are employed to generate synthetic audio. As illustrated in Table 3, our findings show that the proposed data augmentation strategy outperforms the first one in terms of the average pAUC metric. This highlights its superior effectiveness, achieved through the utilization of diverse prompts that generate high-quality samples across various domains.

Table 3. Comparing RC-MAE with and without Diffusion-Based Data Augmentation (DDA).

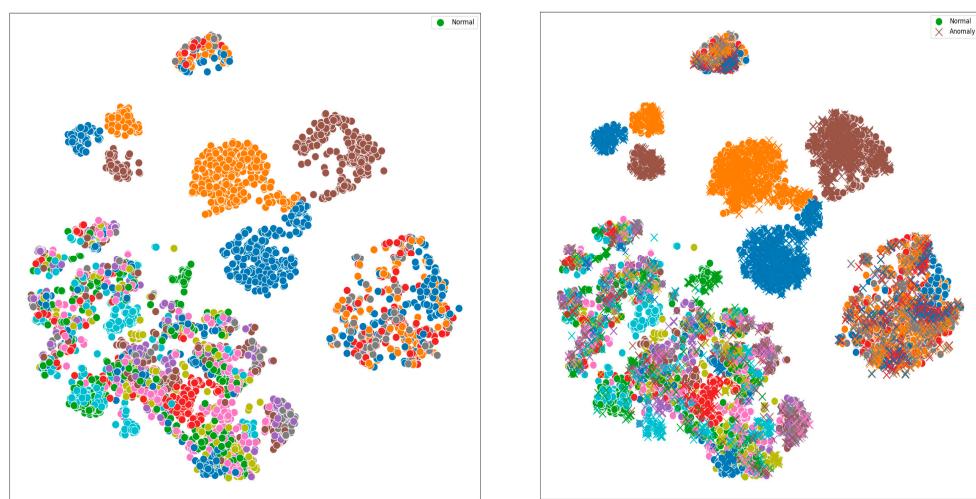
Methods	RC-MAE without DDA (pAUC)	RC-MAE with DDA (pAUC)
Valve	98.01	98.37
Fan	87.44	88.10
Pump	83.89	83.25
Slider	97.16	97.84
ToyCar	90.10	91.35
ToyConveyor	65.38	68.70
Average	86.99	87.93

A comparative analysis of our proposed RC_MAE method is presented in Table 4, where it is evaluated against other state-of-the-art techniques, namely GMADE [7], MobileNetV2 [53], AADCL [2], STgram-MFN [39], Glow-Aff [8], and autoencoders. The results demonstrate that our proposed RC_MAE method improves the average performance of anomaly sound detection. Specifically, compared to the best performance achieved by other methods in the literature, our method demonstrates a 2.03% improvement in average AUC and 1.59% improvement in average pAUC. These findings highlight the effectiveness of our method in enhancing abnormal sound detection performance. These results confirm that, in general, the combination of Contrastive and Generative self-supervised learning proves to be a powerful tool for anomaly detection. Another contributing factor to achieving this improvement is the consideration of various generated real-world noises in audio, allowing the model to better distinguish between noises and anomalies.

Table 4. Comparing RC-MAE with SOTA methods.

Methods	Valve		Fan		Pump		Slider		ToyCar		ToyConveyor		Average	
	AUC	pAUC												
AE Baseline [52]	66.28	50.98	65.83	52.45	72.89	59.99	84.76	66.53	78.77	67.58	72.53	67.58	73.51	65.85
GMADE [7]	99.07	96.20	83.06	79.55	87.87	82.38	97.62	89.70	95.57	91.54	81.46	66.62	90.77	84.33
MobileNetV2 [53]	88.65	87.98	80.19	74.40	82.53	76.50	95.27	85.22	87.66	85.92	69.71	56.43	84.34	77.74
Glow_AFF [8]	91.40	75.00	74.90	65.30	83.40	73.80	94.60	82.80	92.20	84.10	71.50	59.00	85.20	73.90
STgram-MFN [39]	99.64	98.44	94.04	88.97	91.94	81.75	99.55	97.61	94.44	87.68	74.57	63.30	92.36	86.34
AADCL [2]	68.62	55.03	85.27	68.93	86.75	70.85	77.74	61.62	88.79	75.95	71.26	57.40	79.74	64.96
Our Method	99.52	98.37	95.12	88.10	92.82	83.25	99.10	97.84	95.02	91.35	84.80	68.70	94.39	87.93

To illustrate the effect of the learned audio feature representation, Figures 3 and 4 display the t-distributed stochastic neighbour embedding (t-SNE) visualization of the latent features from two distinct phases: pre-training and fine-tuning. Figure 3 illustrates how the pre-training stage can proficiently distinguish various machine sounds in the absence of supervision, solely aided by diffusion augmentation. The Figure 4 clearly demonstrates that incorporating additional metadata during the fine-tuning process enhances the ability to distinguish between different audio features of different machines. Furthermore, our proposed method effectively minimizes the overlap between the latent features associated with normal and anomalous machine IDs, as depicted in Figure 4. This figure confirmed that anomalies and normal sounds were mapped to distinct clusters and were distinguishable in the latent space. This result further demonstrates the effectiveness of the proposed method.



(a) t-SNE visualization of pretrained embedding of just normal sounds

(b) t-SNE visualization of pretrained embedding of normal and anomaly sounds

Figure 3. The t-SNE visualizations of pre-trained model output features on the Train and Test datasets for the different machine IDs. Different colors represent different machine IDs. The shape “●” and shape “×” denote normal and anomalous classes, respectively.

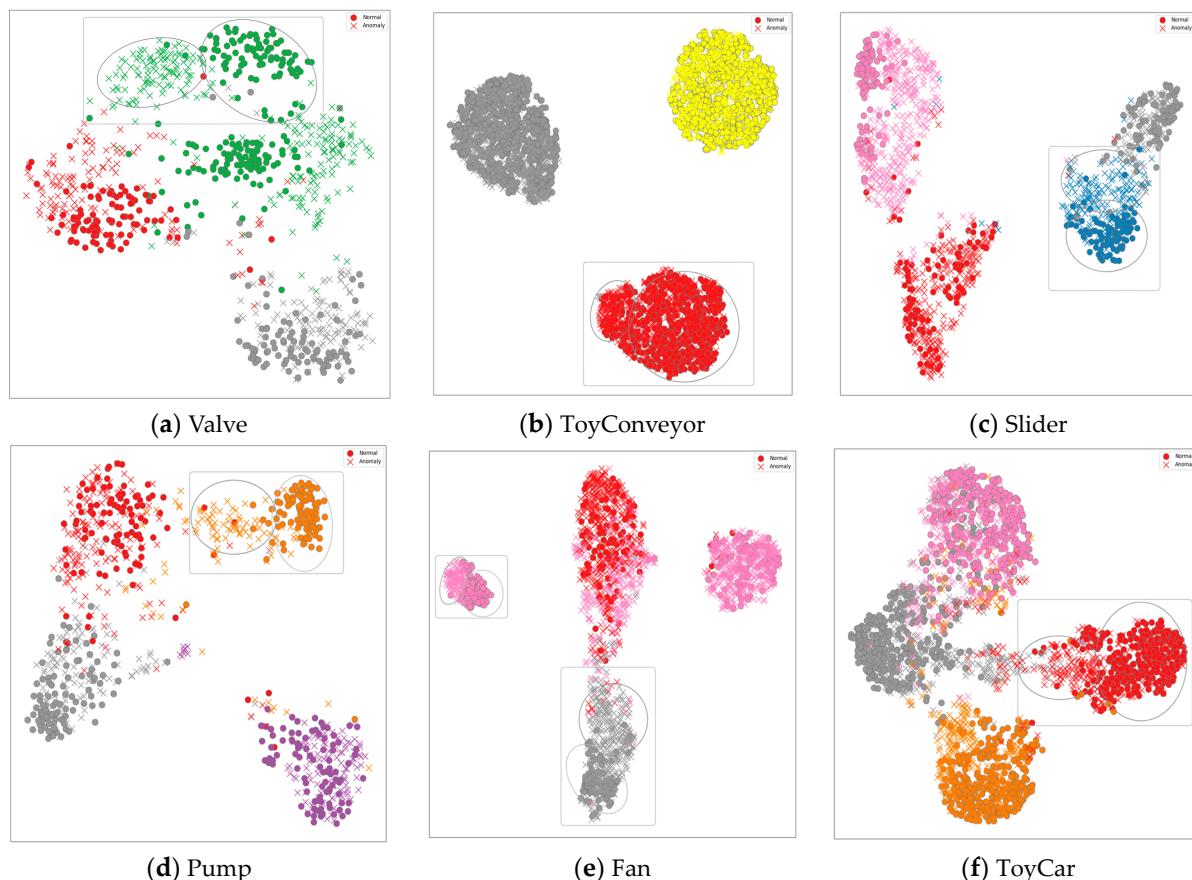


Figure 4. The t-SNE visualizations of fine-tuned model output features on the Train and Test datasets for the different machine IDs (a–f); different colors represent different machine IDs. The shape “●” and shape “×” denote normal and anomalous classes, respectively.

4. Conclusions and Future Works

Predictive detection of machine anomalies is an important concern in many industries, which can be performed by the analysis of machine sounds. In this paper, we introduce a novel two-stage self-supervised anomalous sound detection method named the Regularized Contrastive Masked Autoencoder (RC-MAE) model. In the first stage, we employ a diffusion-based audio augmentation method called AudioLDM for pre-training. By applying our proposed method to both the augmented and real sounds, we aim to achieve a higher-quality audio representation from unlabeled data. To further enhance the model's performance, we then fine-tune the pre-trained model using supervised Contrastive learning. Our approach is pre-trained and fine-tuned using the MIMII and ToyADMOS2 datasets. Specifically, we assessed the model's performance on the MIMII test dataset. The obtained results demonstrate that our model exhibits impressive average Area Under Curve (AUC) and partial Area Under Curve (pAUC) values of 94.39% and 87.93%, respectively. We perform qualitative analyses on different visualizations of the learned representations generated by the RC-MAE encoder. Our observations revealed that fine-tuning the model with metadata, such as machine type and machine ID, enhances the ability to distinguish between normal and anomalous sounds. The t-SNE visualization of the learned representations encodes meaningful information associated with the underlying structures in the pre-trained stage. In this study, the quantitative findings have demonstrated the efficacy of RC-MAE, while the qualitative observations have indicated valuable insights within the acquired representations. To the best of our knowledge, no studies have been conducted with a mixture of Contrastive and Generative self-supervised learning methods; also, there has been no prior research undertaken on diffusion-based data augmentation in the area of anomalous sound detection.

Our research on this topic is ongoing. Our future work will aim to improve the model's ability to generalize different machine types and explore this ability with other datasets and metadata. Furthermore, other types of self-supervised models and their incorporation, and other data augmentation methods can be studied.

Author Contributions: Conceptualization, E.Z., M.S., F.S.M. and M.Y.; methodology, E.Z., M.S. and F.S.M.; software, F.S.M. and M.Y.; validation, E.Z., M.S. and F.S.M.; formal analysis, E.Z., F.S.M. and M.Y.; investigation, E.Z., M.S. and F.S.M.; resources, E.Z., M.S., F.S.M. and M.Y.; data curation, F.S.M. and M.Y.; writing—original draft preparation, E.Z., M.S. and F.S.M.; writing—review and editing, E.Z., M.S., F.S.M. and M.Y.; visualization, F.S.M. and M.Y.; supervision, M.S.; project administration, E.Z. and M.S.; funding acquisition, M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: <https://dcase.community/challenge2020/task-unsupervised-detection-of-anomalous-sounds#audio-dataset> (accessed on 2 July 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

- Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* **2009**, *41*, 1–58. [[CrossRef](#)]
- Daniluk, P.; Gozdziewski, M.; Kapka, S.; Kosmider, M. Ensemble of Auto-Encoder Based Systems for Anomaly Detection. DCASE2020 Challenge. 2020. Available online: https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Daniluk_140_t2.pdf (accessed on 2 June 2023).
- Suefusa, K.; Nishida, T.; Purohit, H.; Tanabe, R.; Endo, T.; Kawaguchi, Y. Anomalous sound detection based on interpolation deep neural network. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 271–275.
- Hojjati, H.; Armanfard, N. Self-supervised acoustic anomaly detection via contrastive learning. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 3253–3257.
- Xiao, F.; Liu, Y.; Wei, Y.; Guan, J.; Zhu, Q.; Zheng, T.; Han, J. The DCASE2022 Challenge Task 2 System: Anomalous Sound Detection with Self-Supervised Attribute Classification and GMM-Based Clustering. Challenge, Technical Report. Available online: https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Guan_24_t2.pdf (accessed on 2 June 2023).

6. Ruff, L.; Kauffmann, J.R.; Vandermeulen, R.A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T.G.; Müller, K.-R. A unifying review of deep and shallow anomaly detection. *Proc. IEEE* **2021**, *109*, 756–795. [[CrossRef](#)]
7. Giri, R.; Cheng, F.; Helwani, K.; Tenneti, S.V.; Isik, U.; Krishnaswamy, A. Group masked autoencoder based density estimator for audio anomaly detection. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2020, Tokyo, Japan, 2–3 November 2020.
8. Dohi, K.; Endo, T.; Purohit, H.; Tanabe, R.; Kawaguchi, Y. Flow-based self-supervised density estimation for anomalous sound detection. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 336–340.
9. Dohi, K.; Imoto, K.; Harada, N.; Niizumi, D.; Koizumi, Y.; Nishida, T.; Purohit, H.; Endo, T.; Yamamoto, M.; Kawaguchi, Y. Description and discussion on DCASE 2022 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques. *arXiv* **2022**, arXiv:2206.05876.
10. Wei, Y.; Guan, J.; Lan, H.; Wang, W. Anomalous Sound Detection System with Self-Challenge and Metric Evaluation for DCASE2022 Challenge Task 2. DCASE2022 Challenge, Technical Report. Available online: https://dcase.community/documents/challenge2022/technical_reports/DCASE2022_Wei_22_t2.pdf (accessed on 2 June 2023).
11. Chen, B.; Bondi, L.; Das, S. Learning to adapt to domain shifts with few-shot samples in anomalous sound detection. In Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 133–139.
12. Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M.R.; Venkatesh, S.; van den Hengel, A. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1705–1714.
13. Huang, W.; Yi, M.; Zhao, X. Towards the generalization of contrastive self-supervised learning. *arXiv* **2021**, arXiv:2111.00743.
14. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
15. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv* **2019**, arXiv:1904.08779.
16. Van den Oord, A.; Dieleman, S.; Schrauwen, B. Deep content-based music recommendation. In *Advances in Neural Information Processing Systems 26*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2013.
17. Ko, T.; Peddinti, V.; Povey, D.; Khudanpur, S. Audio augmentation for speech recognition. In Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association, Dresden, Germany, 6–10 September 2015.
18. Luo, Y.; Chen, Z.; Hershey, J.R.; Le Roux, J.; Mesgarani, N. Deep clustering and conventional networks for music separation: Stronger together. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 61–65.
19. Luo, Y.; Mesgarani, N. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 696–700.
20. Jahanian, A.; Puig, X.; Tian, Y.; Isola, P. Generative models as a data source for multiview representation learning. *arXiv* **2021**, arXiv:2106.05258.
21. Antoniou, A.; Storkey, A.; Edwards, H. Data augmentation generative adversarial networks. *arXiv* **2017**, arXiv:1711.04340.
22. Tran, T.; Pham, T.; Carneiro, G.; Palmer, L.; Reid, I. A bayesian data augmentation approach for learning deep models. In *Advances in Neural Information Processing Systems 30*; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2017.
23. Dos Santos Tanaka, F.H.; Aranha, C. Data augmentation using GANs. *arXiv* **2019**, arXiv:1904.09135.
24. Dat, P.T.; Dutt, A.; Pellerin, D.; Quénot, G. Classifier training from a generative model. In Proceedings of the 2019 International Conference on Content-Based Multimedia Indexing (CBMI), Dublin, Ireland, 4–6 September 2019; pp. 1–6.
25. Yamaguchi, S.; Kanai, S.; Eda, T. Effective data augmentation with multi-domain learning gans. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 6566–6574.
26. Besnier, V.; Jain, H.; Bursuc, A.; Cord, M.; Pérez, P. This dataset does not exist: Training models from generated images. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 1–5.
27. Haque, A. EC-GAN: Low-sample classification using semi-supervised algorithms and GANs (Student Abstract). In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 15797–15798.
28. He, R.; Sun, S.; Yu, X.; Xue, C.; Zhang, W.; Torr, P.; Bai, S.; Qi, X. Is synthetic data from generative models ready for image recognition? *arXiv* **2022**, arXiv:2210.07574.
29. Available online: <https://chat.openai.com/chat> (accessed on 2 June 2023).
30. Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; Plumley, M.D. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv* **2023**, arXiv:2301.12503.
31. Purohit, H.; Tanabe, R.; Ichige, K.; Endo, T.; Nikaido, Y.; Suefusa, K.; Kawaguchi, Y. MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection. *arXiv* **2019**, arXiv:1909.09347.

32. Primus, P. Reframing Unsupervised Machine Condition Monitoring as a Supervised Classification Task with Outlier-Exposed Classifiers. Techniacl Report, DCASE2020 Challenge. 2020. Available online: https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Primus_36_t2.pdf (accessed on 2 June 2023).
33. Dorj, E.; Altangerel, E. Anomaly detection approach using hidden Markov model. In Proceedings of the International Forum on Strategic Technology, IFOST, Ulaanbaatar, Mongolia, 28 June–1 July 2013; Volume 2, pp. 141–144.
34. Ntalampiras, S.; Potamitis, I.; Fakotakis, N. Probabilistic novelty detection for acoustic surveillance under real-world conditions. *IEEE Trans. Multimed.* **2011**, *13*, 713–719. [CrossRef]
35. Sasou, A.; Odontsengel, N. Acoustic novelty detection based on AHLAC and NMF. In Proceedings of the 2012 International Symposium on Intelligent Signal Processing and Communications Systems, Tamsui, Taiwan, 4–7 November 2012; pp. 872–875.
36. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning representations by back-propagating errors. *Nature* **1986**, *323*, 533–536. [CrossRef]
37. Guan, J.; Xiao, F.; Liu, Y.; Zhu, Q.; Wang, W. Anomalous Sound Detection Using Audio Representation with Machine ID Based Contrastive Learning Pretraining. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
38. Koizumi, Y.; Kawaguchi, Y.; Imoto, K.; Nakamura, T.; Nikaido, Y.; Tanabe, R.; Purohit, H.; Suefusa, K.; Endo, T.; Yasuda, M.; et al. Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring. *arXiv* **2020**, arXiv:2006.05822.
39. Liu, Y.; Guan, J.; Zhu, Q.; Wang, W. Anomalous sound detection using spectral-temporal information fusion. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 816–820.
40. Gong, Y.; Rouditchenko, A.; Liu, A.H.; Harwath, D.; Karlinsky, L.; Kuehne, H.; Glass, J.R. Contrastive audio-visual masked autoencoder. In Proceedings of the Eleventh International Conference on Learning Representations, Virtual, 25–29 April 2022.
41. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 16000–16009.
42. Mishra, S.; Robinson, J.; Chang, H.; Jacobs, D.; Sarna, A.; Maschinot, A.; Krishnan, D. A simple, efficient and scalable contrastive masked autoencoder for learning visual representations. *arXiv* **2022**, arXiv:2210.16870.
43. Bardes, A.; Ponce, J.; LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv* **2021**, arXiv:2105.04906.
44. Koizumi, Y.; Saito, S.; Uematsu, H.; Harada, N.; Imoto, K. ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection. In Proceedings of the 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019; pp. 313–317.
45. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 10684–10695.
46. Burg, M.F.; Wenzel, F.; Zietlow, D.; Horn, M.; Makansi, O.; Locatello, F.; Russell, C. A data augmentation perspective on diffusion models and retrieval. *arXiv* **2023**, arXiv:2304.10253.
47. Trabucco, B.; Doherty, K.; Gurinas, M.; Salakhutdinov, R. Effective data augmentation with diffusion models. *arXiv* **2023**, arXiv:2302.07944.
48. Mao, J.; Yin, X.; Chang, Y.; Zhou, H. Improvements to Self-Supervised Representation Learning for Masked Image Modeling. *arXiv* **2022**, arXiv:2205.10546.
49. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 7–23 July 2022; pp. 12310–12320.
50. Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; Hu, H. Simmim: A simple framework for masked image modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 21–24 June 2022; pp. 9653–9663.
51. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. In *Advances in Neural Information Processing Systems* 33; Neural Information Processing Systems Foundation, Inc.: La Jolla, CA, USA, 2020; pp. 18661–18673.
52. Chen, Y.; Song, Y.; Cheng, T. Anomalous Sounds Detection Using a New Type of Autoencoder Based on Residual Connection. DCASE2020 Challenge. 2020. Available online: https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Chen_25_t2.pdf (accessed on 2 June 2023).
53. Giri, R.; Tenneti, S.V.; Cheng, F.; Helwani, K.; Isik, U.; Krishnaswamy, A. Self-supervised classification for detecting anomalous sounds. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop 2020, Tokyo, Japan, 2–4 November 2020.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.