ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447941

ANOMALOUS SOUND DETECTION BY FEATURE-LEVEL ANOMALY SIMULATION

Vitjan Zavrtanik *Matija Marolt* *Matej Kristan* *Danijel Skočaj*

Faculty of Computer and Information Science, University of Ljubljana

ABSTRACT

Recently a growing number of works focus on machine defect detection from anomalous audio patterns. The datasets for the machine audio domain are scarce and recent methods that perform well on benchmarks such as DCASE2020 Task 2, rely on auxiliary information such as annotated data from other training classes in the domain to extract information that can be used in deep-learning classification-based anomaly detection approaches. However, in practical scenarios, annotated data from the same domain may not be readily available so annotation-free methods that can learn appropriate audio representations from unannotated data are needed. We propose AudDSR, a simulation-based anomaly detection method that learns to detect anomalies without additional annotated data and instead focuses on a discrete feature space sampling method for an anomaly simulation process. AudDSR outperforms competing methods that do not rely on annotated data on the DCASE2020 anomalous sound detection benchmark and even matches the performance of some methods that utilize additional annotation information.

Index Terms— anomalous sound detection, anomaly simulation, vector quantization

1. INTRODUCTION

The audio anomaly detection problem typically focuses on detecting defective machine activity from audio recordings of that machine. During training no anomalous recordings are provided and the typical audio anomaly detection models learn to model an anomaly-free distribution of sound by training only on anomaly-free recordings. The standard audio anomaly detection benchmark, DCASE2020 [1], provides recordings of several different machine types and several instances of the same machine type. For each recording the machine type and machine instance (machine id) are annotated. These annotations provide significant information about possible deviations in the sound representation for each machine. Most top performing methods use these available annotations of all classes to train a classification network, and use the assumption that the trained network will fail to classify anomalous recordings, making them detectable. Such recordings can also be used in an outlier exposure [2] process as anomalous samples.

Having a diverse set of annotated data from the same domain can not often be expected in practice [3], therefore methods that can detect anomalies without relying on additional information from such annotated datasets are necessary and are a promising direction for further research effort. Previous generative approaches such as autoencoder models [1] and masked autoencoder models [4, 5] do not require annotated data, but rely merely on their reconstruction capability to detect anomalies. These reconstruction-based methods are trained on anomaly-free recordings of an individual machine instance and are assumed to poorly reconstruct anomalous recordings thus making anomalies detectable purely through a reconstruction error metric. A known problem of reconstruction-based methods is that they tend to generalize very well, therefore some near-in-distribution anomalous frames may be accurately reconstructed despite not occurring in the training set.

In computer vision, anomaly detection has become a very active research field in recent years. Although some methods, such as flow-based models [6, 7], are used both in audio and vision-based anomaly detection, methods focusing on anomaly simulation [8, 9, 10] that are the state-of-the-art on visual anomaly detection problems have not been previously used in the audio setting. One of the issues in adapting such methods for audio anomaly detection is the lack of a well defined anomaly generation process. Additionally, vision-based methods often rely on large pretrained networks. Large pretrained audio backbone networks do not necessarily capture the nuances necessary for audio machine defect detection, due to being trained on data from a different domain.

As our main contribution we thus propose AudDSR, a novel audio anomaly detection model that does not rely on annotated data but instead first learns a general discrete representation of machine audio and then simulates realistic anomalies by sampling from the learned discrete feature space. On the standard DCASE2020 Task 2 dataset [1], the proposed method outpeforms competing generative audio anomaly detection models, that do not require additional annotations, by approximately 4 percentage points on the standard metrics. Additionally, AudDSR even reaches the performance of some strong audio anomaly detection baselines that rely on using additional annotations. The framework of the proposed method is also extendable which enables it to incorporate additional information provided by annotations in which case it outperforms several recent annotation-reliant models.

2. RELATION TO PRIOR WORK

The DCASE2020 is a commonly used audio anomaly detection benchmark [1]. The benchmark contains 6 machine types with each machine type containing recordings of three or four machines. These machine instances are labeled with machine IDs, leading to a total of 23 classes. The top performing methods rely on the annotations of the 23 classes to learn specific features that are useful for the anomaly detection task and are difficult to obtain without having a relatively large annotated dataset from the task domain. Since having such an annotated dataset available for all practical scenarios is not guaranteed this is a major drawback of such methods. In flow-based methods [6] a normalizing flow model is trained on the data of a specific class, while other classes are used as outliers to adapt the negative log likelihood in out-of-distribution examples. In the MobileNetV2 [11] approach the network is trained as a classifier for individual machine IDs and the softmax score of the correct class is used as the anomaly score. In STGram [12] additional features extracted from the waveform are used to improve the result. In the top performing method GeCo [4] both the reconstructive method PAE [5] and a discriminative network are used. The anomaly score is then the combination of the reconstruction error by PAE and the discriminative output. Generative methods that do not require additional annotations are mostly focused on reconstruction [1, 5, 13] or on density estimation [14]. Reconstruction-based approaches can either generalize and also accurately reconstruct anomalous regions or fail to reconstruct events that are rare but are not considered anomalies. Our approach, AudDSR, focuses on generating simulated anomalies by sampling from a discrete feature space, thus not relying on machine ID annotations but also avoiding the issues of reconstruction-based methods.

3. OUR APPROACH: AUDDSR

We propose a novel audio anomaly detection model, AudDSR, based on the Audio Dual Subspace Reconstruction [9]. AudDSR is trained in two stages. First, it utilizes a vector-quantized autoencoder, namely VQ-VAE-2 [15], to learn a general discrete feature space for the domain of machine audio recordings. Then, the learned discrete space is used to sample realistic anomalies from the learned distribution. A discriminative network is then trained to detect simulated anomalies and generalizes well to real-world anomalies. Note that annotations of machine instances are not used to train AudDSR, thus not requiring such a dataset during training. In Section 3.1 the first stage of training is described where the general discrete feature space is learned. In Section3.2, the architecture of AudDSR is described which is based on the visual surface anomaly detection method DSR [9] and in Section 3.3, the anomaly simulation process is defined.

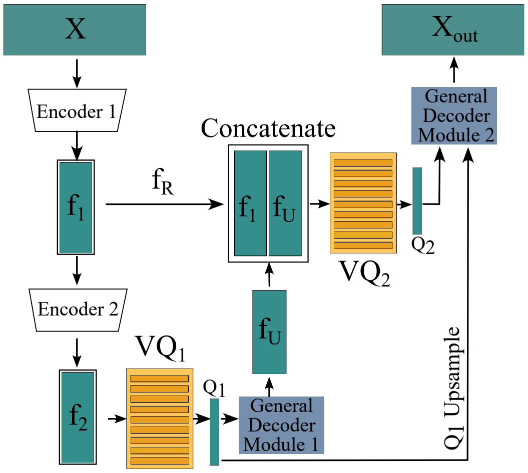

Fig. 1. The discrete autoencoder architecture of AudDSR.

3.1. Discrete latent space learning

The vector-quantized autoencoder VQ-VAE-2 [15] is used to learn a general discrete latent space. A two stage architecture is used to ensure an accurate reconstruction of the input data x. The complete architecture of the discrete autoencoder used is shown in Figure 1. The input spectrogram is first encoded to f_1 and f_2 by Encoder 1 and Encoder 2, respectively. f_1 is then quantized to the nearest vector in codebook VQ_1, resulting in the quantized feature map Q_1. Q_1 is then decoded and upsampled by the General Decoder Module 1, resulting in f_U, which is then concatenated with f_1. The concatenation is then quantized to the nearest vectors in codebook VQ_2 resulting in Q_2. Q_1 is upsampled by bilinear interpolation to the shape of Q_2, concatenated and input into the General Decoder Module 2, which produces a reconstruction X_{out} of the input spectrogram X. The standard VQ-VAE-2 loss is used to train the discrete autoencoder:

$$\begin{aligned}\mathcal{L}_{\text{ae}} = \lambda_x L_2(\mathbf{X}, \mathbf{X}_{\text{out}}) \\ + L_2(sg[\mathbf{f_2}], \mathbf{Q_1}) + \lambda_K L_2(\mathbf{f_2}, sg[\mathbf{Q_1}]) \\ + L_2(sg[\mathbf{f_1}], \mathbf{Q_2}) + \lambda_K L_2(\mathbf{f_1}, sg[\mathbf{Q_2}]), \quad (1)\end{aligned}$$

where $L_2(\cdot)$ is the Euclidean distance and $sg[\cdot]$ is the stop gradient operator. λ_K is fixed to 0.25 in all experiments following [15].

3.2. AudDSR architecture

AudDSR follows the architecture of DSR [9] and is shown in Figure 2. The Mel spectrogram of the input, X, is input into the encoder of the discrete autoencoder, where the quantized feature maps Q_1 and Q_2 are extracted. Then, Q_1 and Q_2 are

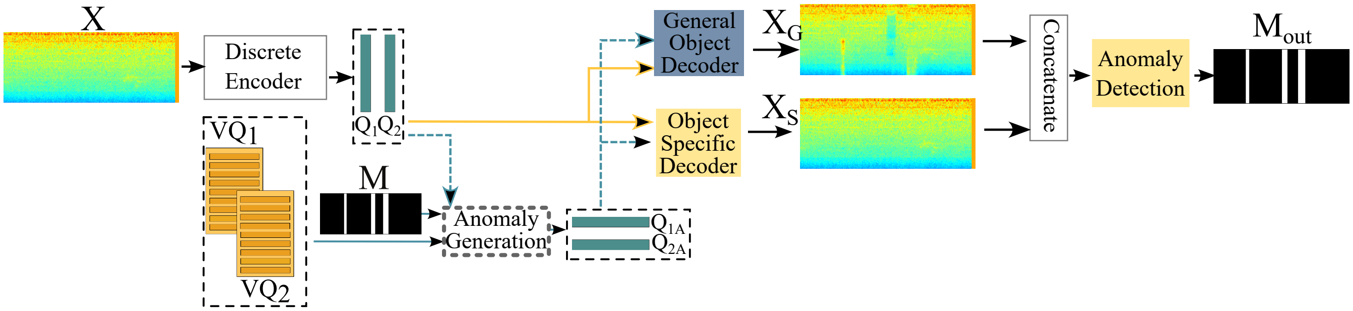

Fig. 2. The architecture of AudDSR. The modules marked in yellow are trainable during the second stage of training. Steps marked with blue dotted arrows are only done during training.

input into two decoder networks. The general object decoder maintains the information in Q_1 and Q_2 and directly reconstructs the input spectrogram X which may be anomalous at inference so any potential anomalies are reconstructed in X_G. The general object decoder consists of the General decoder modules of the discrete autoencoder in Figure 1 and is not trained in the second stage. The second decoder, the object specific decoder, aims to restore the anomaly-free spectrogram so any potential anomalies are instead restored to their anomaly-free values in the corrected spectrogram X_S. The resulting spectrograms X_G and X_S are then concatenated and input into the Anomaly detection module which outputs a segmentation map of the anomalous sections of X.

Only the Object specific decoder and the Anomaly detection module are trained in the second stage. The object specific decoder is trained to reconstruct anomaly-free spectrograms from synthetically corrupted quantized feature maps Q_{1A} and Q_{2A} using an L2 loss. The anomaly detection module is trained using the Focal loss following [8, 9].

3.3. Anomaly sampling

During training, anomalies are simulated by augmenting the quantized feature maps Q_1 and Q_2. First an anomaly map M is generated by one of two processes. In one process, a diverse set of anomalous parts of the spectrogram are generated by thresholding and binarizing a Perlin noise map. Perlin noise maps are commonly used in discriminative visual anomaly detection methods [9, 8], but do not accurately model the anomalies that are occur in spectrograms, where the anomalies often span across the entire duration of the recording or only across a specific frequency band. A new anomaly shape simulation process is therefore defined. First a frequency band where anomalies will be generated is chosen. Then, several time segments are chosen in which consecutive frames will be augmented. The chosen time segments and the frequency band are then marked in the anomaly map M. M is then resized to fit to the spatial dimensions of Q_1 and Q_2. Feature vectors of Q_1 and Q_2 in regions corresponding to positive values of M are replaced with feature vectors sampled from codebooks VQ_1 and VQ_2 respectively. The resulting augmented feature maps containing simulated anomalies are marked with Q_{1A} and Q_{2A} in Figure 2.

4. EXPERIMENTS

AudDSR is evaluated on the DCASE2020 Task 2 benchmark [1] that contains 6 machine types (ToyCar, ToyConveyor, Fan, Pump, Slider, and Valve). Each machine type consists of recordings of four machines with the exception of ToyConveyor that contains 3 machines. The recordings of different machine instances are labeled with machine IDs, leading to 23 total classes. The AUROC and pAUC metrics are used for evaluation which are standard for this benchmark. pAUC is the AUROC with a maximum false-positive-rate of 0.1. The average metrics over all machine IDs for a specific machine type is reported in all experiments.

4.1. Implementation details

AudDSR takes log-Mel spectrograms as the input with 128 Mel filters, a window size of 1024 and a hop size of 512. The sample rate of the DCASE2020 benchmark samples is 16kHz.

In the first training stage, the discrete autoencoder was trained for 20000 iterations with a batch size of 256. The Adam optimizer was used and the learning rate was set to 0.0002. The goal of the first stage is to learn a general discrete audio feature space. In the second stage of training, AudDSR was trained for 10000 iterations with a batch size of 16. The Adam optimizer was used and the learning rate was set to 0.0002. During training, half of the samples in each batch were set to contain simulated anomalies, while the rest were anomaly-free. The anomaly-score for each input spectrogram was defined as the mean value of the output anomaly map.

4.2. Results

AudDSR is compared to other top-performing recent methods on the DCASE2020 benchmark [1]. The results are listed in Table 1, where they are split between methods that

Table 1. Performance on the DCASE2020 Task 2 benchmark. Performance in terms of AUC and pAUC are reported. Metrics where AudDSR outperforms competing annotation-free methods are written in bold.

Methods	ToyCar		ToyConveyor		Fan		Pump		Slider		Valve		Average	
	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC	AUC	pAUC
Annotation-free														
AE [1]	80.90	69.90	73.40	61.10	66.20	53.20	72.90	60.30	85.50	67.80	66.30	51.20	74.20	60.58
IDNN [13]	80.19	71.87	75.74	61.26	69.15	53.53	74.06	61.26	88.32	69.07	88.31	65.67	79.30	63.78
ANP [16]	72.50	67.30	67.00	54.50	69.20	54.40	72.80	61.80	90.70	74.20	86.90	70.70	76.52	63.82
PAE [5]	75.35	69.70	77.58	61.37	72.94	54.37	74.27	62.01	91.92	74.39	95.41	81.24	81.25	67.18
AudDSR	**91.89**	**82.90**	**78.02**	**64.60**	**73.82**	**64.98**	**85.91**	**74.32**	90.16	71.54	90.05	70.20	**84.97**	**71.45**
Annotation-reliant														
MobileNetV2 [11]	87.66	85.92	69.71	56.43	80.19	74.40	82.53	76.50	95.27	85.22	88.65	87.98	84.00	77.74
GlowAff [6]	92.20	84.10	71.50	59.00	74.90	65.30	83.40	73.80	94.60	82.80	91.40	75.00	85.20	73.90
STgram [12]	88.80	87.38	72.93	63.62	91.30	86.73	91.25	81.69	99.36	96.84	94.44	91.58	89.68	84.64
AudDSR$_{annot}$	93.60	90.65	81.57	71.23	77.46	75.39	88.52	79.16	98.56	93.00	98.90	94.76	90.12	84.59
GeCo [4]	96.62	89.33	74.69	65.82	92.73	85.19	93.09	86.89	98.61	95.26	99.06	95.52	92.47	86.34

rely on machine ID annotations and the methods that do not. AudDSR outperforms all the annotation-free methods, even outperforming PAE [5] by 3.7 percentage points in terms of mean AUC and 4.3 percentage points in terms of mean pAUC. AudDSR also outperforms MobileNetV2 [11] in terms of AUC, but falls behind in terms of pAUC. Nonetheless AudDSR narrows the gap between Annotation-free and Annotation-required methods.

Additionally, AudDSR is extendable and can also utilize additional information if available. In experiment AudDSR$_{annot}$, annotated examples from other machine instances of the same type are used as outliers for training the Anomaly detection module in addition to the simulated anomalies. AudDSR$_{annot}$ outperforms MobileNetV2 [11] and GlowAff [6], achieves comparable performance to STGram [12], but does not quite reach the performance of the best performing method GeCo [4]. There is however, room for improvement. The training anomaly map for outlier spectrograms was simply set to 1 in all regions which may not be ideal. Additionally, AudDSR$_{annot}$ does not utilize individual machine ID annotations and only learns to differentiate between the in-distribution class and the rest, which does not utilize the entire information available. This experiment demonstrates the extendable nature of AudDSR and shows the potential of AudDSR to work with additional information when available.

4.3. Discrete autoencoder training data

The impact of the dataset used to learn the general discrete representation of the discrete autoencoder is evaluated. In most experiments the discrete autoencoder is trained on the entire DCASE2020 Task 2 dataset without taking machine ID annotations into account. Table 2 lists the mean AUC and pAUC performance using the discrete autoencoder trained on the unannotated DCASE2020 data in experiment AudDSR$_{dcase}$ and on a subset of the AudioSet [17] data in experiment AudDSR$_{AudioSet}$. Both achieve comparable performance which demonstrates that training a discrete latent space for anomaly simulation is robust to the dataset choice.

Experiment	AUC	pAUC
AudDSR$_{dcase}$	84.97	71.45
AudDSR$_{AudioSet}$	84.86	71.61

Table 2. The impact of the discrete autoencoder training dataset on the anomaly detection performance.

5. CONCLUSION

We present AudDSR, an audio anomaly detection method that does not rely on an annotated classification dataset of the same domain for accurate anomaly detection. It focuses on generating simulated anomalies in a discrete feature space to learn an anomaly detection module that is capable of generalizing to real-world anomalies. On the standard DCASE2020 Task 2 anomaly detection benchmark, AudDSR achieves state-of-the-art results outperforming all methods that are not annotation-reliant on both AUC and pAUC metrics by approximately 3 and 4 percentage points, respectively. The AudDSR framework is flexible and with a slight modification to the training process, AudDSR can also incorporate the additional annotation information. In this case it achieves excellent results and outperforms most annotation-reliant methods. In future work, the concept of AudDSR could be extended to waveform-based pretrained quantized features spaces such as EnCodec [18] which may improve results or offer interesting insights.

Acknowledgement *This work was supported by Slovenian research agency programs L2-3169, J2-2506, P2-0214. Vitjan Zavrtanik was supported by the Young researcher program of the ARRS.*

6. REFERENCES

[1] Yuma Koizumi, Yohei Kawaguchi, Keisuke Imoto, Toshiki Nakamura, Yuki Nikaido, Ryo Tanabe, Harsh Purohit, Kaori Suefusa, Takashi Endo, Masahiro Yasuda, et al., "Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2006.05822*, 2020.

[2] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich, "Deep anomaly detection with outlier exposure," in *International Conference on Learning Representations*, 2019.

[3] Kota Dohi, Keisuke Imoto, Noboru Harada, Daisuke Niizumi, Yuma Koizumi, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei Kawaguchi, "Description and discussion on DCASE 2023 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2305.07828*, 2023.

[4] Xiao-Min Zeng, Yan Song, Zhu Zhuo, Yu Zhou, Yu-Hong Li, Hui Xue, Li-Rong Dai, and Ian McLoughlin, "Joint generative-contrastive representation learning for anomalous sound detection," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[5] Xiao-Min Zeng, Yan Song, Li-Rong Dai, and Lin Liu, "Predictive autoencoders are context-aware unsupervised anomalous sound detectors," in *National Conference on Man-Machine Speech Communication*. Springer, 2023, pp. 101–113.

[6] Kota Dohi, Takashi Endo, Harsh Purohit, Ryo Tanabe, and Yohei Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 336–340.

[7] Marco Rudolph, Tom Wehrbein, Bodo Rosenhahn, and Bastian Wandt, "Asymmetric student-teacher networks for industrial anomaly detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2592–2602.

[8] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj, "Draem-a discriminatively trained reconstruction embedding for surface anomaly detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8330–8339.

[9] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj, "Dsr–a dual subspace re-projection network for surface anomaly detection," in *European conference on computer vision*. Springer, 2022, pp. 539–554.

[10] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang, "Simplenet: A simple network for image anomaly detection and localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20402–20411.

[11] Ritwik Giri, Srikanth V Tenneti, Fangzhou Cheng, Karim Helwani, Umut Isik, and Arvindh Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proceedings of the Detec- tion and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020.

[12] Youde Liu, Jian Guan, Qiaoxi Zhu, and Wenwu 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.

[13] Kaori Suefusa, Tomoya Nishida, Harsh Purohit, Ryo Tanabe, Takashi Endo, and Yohei 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.

[14] Ritwik Giri, Fangzhou Cheng, Karim Helwani, Srikanth V Tenneti, Umut Isik, and Arvindh Krishnaswamy, "Group masked autoencoder based density estimator for audio anomaly detection," in *Proceedings of the Detec- tion and Classification of Acoustic Scenes and Events 2020 Workshop (DCASE2020)*, 2020.

[15] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals, "Generating diverse high-fidelity images with vq-vae-2," *Advances in neural information processing systems*, vol. 32, 2019.

[16] Gordon Wichern, Ankush Chakrabarty, Zhong-Qiu Wang, and Jonathan Le Roux, "Anomalous sound detection using attentive neural processes," in *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2021, pp. 186–190.

[17] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[18] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *Transactions on Machine Learning Research*, 2023.