

# ID-CONDITIONED AUTO-ENCODER FOR UNSUPERVISED ANOMALY DETECTION

*Stawomir Kapka*

Samsung R&D Institute Poland  
Artificial Intelligence  
Warsaw, Poland  
s.kapka@samsung.com

## ABSTRACT

In this paper, we introduce ID-Conditioned Auto-Encoder for unsupervised anomaly detection. Our method is an adaptation of the Class-Conditioned Auto-Encoder (C2AE) designed for the open-set recognition. Assuming that non-anomalous samples constitute of distinct IDs, we apply Conditioned Auto-Encoder with labels provided by these IDs. Opposed to C2AE, our approach omits the classification subtask and reduces the learning process to the single run. We simplify the learning process further by fixing a constant vector as the target for non-matching labels. We apply our method in the context of sounds for machine condition monitoring. We evaluate our method on the ToyADMOS and MIMII datasets from the DCASE 2020 Challenge Task 2. We conduct an ablation study to indicate which steps of our method influences results the most.

**Index Terms**— DCASE 2020 Challenge Task 2, Unsupervised anomaly detection, Machine Condition Monitoring, Conditioned Auto-Encoder

## 1. INTRODUCTION

Unsupervised anomaly detection is a problem of detecting anomalous samples under the condition that only non-anomalous (normal) samples have been provided during training phase. In this paper, we focus on unsupervised anomaly detection in the context of sounds for machine condition monitoring – i.e., detecting mechanical failure by listening.

Many techniques have been studied for detecting anomalous sounds. Among others there are solutions based on SVMs [1, 2], sparse coding [3, 4], GMMs [5, 6], Neyman-Pearson lemma [7, 8], signal processing [9, 10, 11], Interpolation DNN [12], and Auto-Encoders [8, 13, 14, 15]. Beyond sounds, much more techniques for anomaly detection based on deep learning can be found in the survey [16].

Unsupervised anomaly detection can be viewed as a special case of the open-set recognition [17]. In fact, since during training we are provided only with normal samples and we have to predict whether new samples are normal or anomalous, we can look at it as a binary classification problem with only one given class during training phase.

Recently, Class-Conditioned Auto-Encoder (C2AE) [18] has been introduced for the open-set recognition problem. According to the survey on the open-set recognition [19], it is currently the state-of-the-art in the open-set recognition problem. In section 2, we introduce an ID-Conditioned Auto-Encoder (IDCAE), which is the adapted version of C2AE applicable to the unsupervised anomaly detection.

Task 2 in this year IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE 2020) [20, 21] focuses precisely on the unsupervised detection of anomalous sounds for machine condition monitoring. The data for this task is ToyADMOS [22] and MIMII Dataset [23] which consists of sounds of six types of operating machines. In section 3, we develop the model for this challenge, and we evaluate its performance via an ablation study in section 4.

## 2. PROPOSED METHOD

Let us consider an arbitrary but fixed machine type, henceforth called the machine unless otherwise specified. We assume that we have various IDs of the machine. It is precisely the case in task 2 in the DCASE 2020 Challenge. In the nomenclature from [18] we treat machines with different IDs as distinct classes.

Our system constitutes of three main parts:

- encoder  $E : \mathcal{X} \rightarrow \mathcal{Z}$  which maps *feature vector*  $X$  from input space  $\mathcal{X}$  to the *code*  $E(X)$  in the latent space  $\mathcal{Z}$ ,
- decoder  $D : \mathcal{Z} \rightarrow \mathcal{X}$  which takes the code  $Z$  from  $\mathcal{Z}$  and outputs the vector  $D(Z)$  of the same shape as feature vectors from  $\mathcal{X}$ ,
- conditioning made of two functions  $H_\gamma, H_\beta : \mathcal{Y} \rightarrow \mathcal{Z}$  which take the one-hot label  $l$  from  $\mathcal{Y}$  and map it to the vectors  $H_\gamma(l), H_\beta(l)$  of the same size as codes from  $\mathcal{Z}$ .

During feed-forward, the code  $Z$  is combined with  $H_\gamma(l), H_\beta(l)$  to form  $H(Z, l) = H_\gamma(l) \cdot Z + H_\beta(l)$ , which is an affine transformation of the latent space conditioned by  $l$  [24]. Thus, our whole system takes two inputs  $X, l$  from  $\mathcal{X}$  and  $\mathcal{Y}$  respectively and outputs  $D(H(E(X), l))$ .

Given an input  $X$  with some ID, we call label corresponding to this ID by the *match* and all other labels by *non-matches*. We wish that our system reconstructs  $X$  faithfully if and only if it is a normal sample conditioned by the matching label.

Given an input  $X$ , we set the label  $l$  to the match with probability  $\alpha$  or to a randomly selected non-match with probability  $1 - \alpha$ , where  $\alpha$  is predefined. Thus, for a batch  $X_1, X_2, \dots, X_n$  approximately  $\alpha$  fraction of samples will be conditioned by matches and  $1 - \alpha$  by non-matches. If  $l$  is the match, then the loss equals difference between the system’s output and  $X$ , that is  $\|D(H(E(X), l)) - X\|$ . If  $l$  is a non-match, then the loss equals difference between the system’s output and some predefined constant vector  $C$  with the same shape as  $X$ , that is  $\|D(H(E(X), l)) - C\|$ . In our setting  $\|\cdot\|$  is either  $L_1$  or the square of  $L_2$  norm.

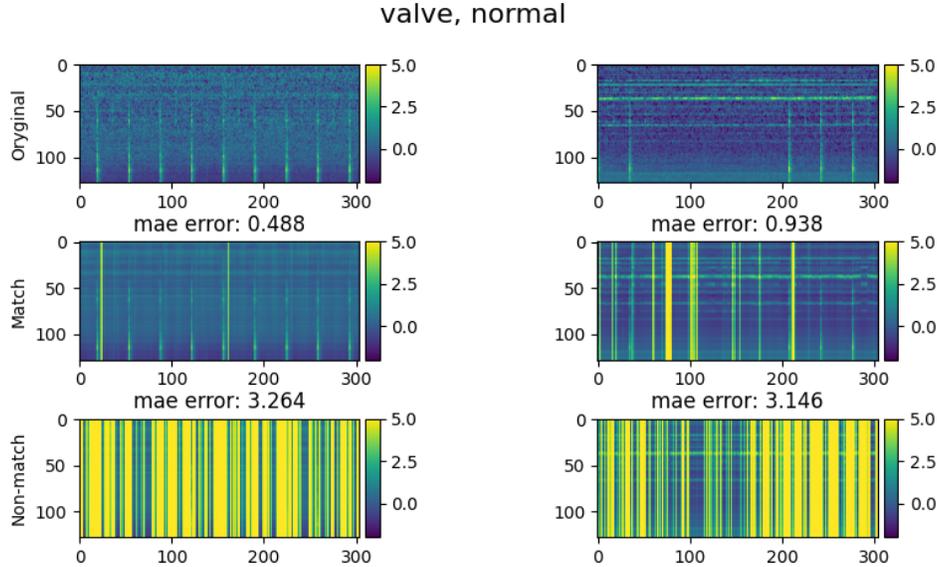


Figure 1: Two exemplary spectrograms of some valve normal samples with their corresponding reconstructions. In the top row we have original spectrograms of the whole recordings. In the middle row we have reconstructions with matching labels. We see that reconstructions are rather faithful. In the bottom row we have reconstructions with non-matching labels, and as expected the reconstructed spectrograms are quite off.

During the inference we always feed the network with matching labels. If a sample is non-anomalous, we expect the reconstruction to be faithful resulting in low reconstruction error (see Figure 1). If the sample is anomalous, there may be two cases. If the sample is nothing like any sample during training, then generally auto-encoder wouldn't be able to reconstruct it resulting in high reconstruction error. However, if the sample resemble normal samples with other IDs, then the auto-encoder will try to reconstruct the vector  $C$  resulting again in high error (see Figure 2).

In general, the point of auto-encoder-based solutions for anomaly detection is to separate distributions of reconstruction errors for normal and anomalous samples. In our case, the distribution of reconstruction errors for anomalies is shifted further to the right due to the samples that resemble samples with other IDs. Thus, it indicates that our method should work at least as good as regular auto-encoder. Overall, distribution of reconstruction errors for anomalies with matching labels should be between distributions of reconstruction errors for normal samples with matching and non-matching labels (see Figure 3).

The additional advantage of our approach is that feeding the network with more IDs may result in better performance. In fact, in the section 4 we show that it holds in the case of machine condition monitoring.

### 3. MODEL

#### 3.1. Features

In our model, we feed the network with fragments of normalised log-mel power spectrograms. Feature vector space  $\mathcal{X}$  consists of vectors of the shape  $(F, M)$ , where  $F$  is the frame size and  $M$  is the number of mels. Given an audio signal we first compute its Short Time Fourier Transform with 1024 window and 512 hop size, we transform it to the power mel-scaled spectrogram with  $M$  mels,

Table 1: The architecture of IDCAE

Encoder ( $E$ )	Decoder ( $D$ )	Conditioning ( $H_\gamma, H_\beta$ )
Input ( $F, M$ )	Input 16	Input #IDs
Flatten	DenseBlock 128	Dense 16
DenseBlock 128	DenseBlock 128	sigmoid
DenseBlock 64	DenseBlock 128	Dense 16
DenseBlock 32	DenseBlock 128	
DenseBlock 16	Dense $F \cdot M$	
	Reshape ( $F, M$ )	

DenseBlock  $n$ : Dense  $n$ , Batch-norm, relu

and we take its logarithm with base 10 and multiply it by 10. Finally, we standardize all spectrograms frequency-wise to zero mean and unit variance, and sample frames of size  $F$  as an input to our system.

#### 3.2. Architecture

As described in subsection 2, our model constitutes of the encoder  $E$ , the decoder  $D$  and the conditioning  $H_\gamma, H_\beta$ . In our case, all these components are fully connected neural networks. Thus, we have to flatten feature vectors and reshape the output to  $(F, M)$  for the sake of the dimension compatibility. The dense layers in  $E$  and  $D$  are followed by batch-norm and relu activation function, while the dense layers in  $H_\gamma, H_\beta$  are followed just by sigmoid activation functions.  $E$  has three hidden dense layers with 128, 64 and 32 units followed by the latent dense layer with 16 units.  $D$  is made of four hidden dense layers each with 128 units.  $H_\gamma$  and  $H_\beta$  have both a single hidden dense layer with 16 units. We summarise the architecture in Table 1.

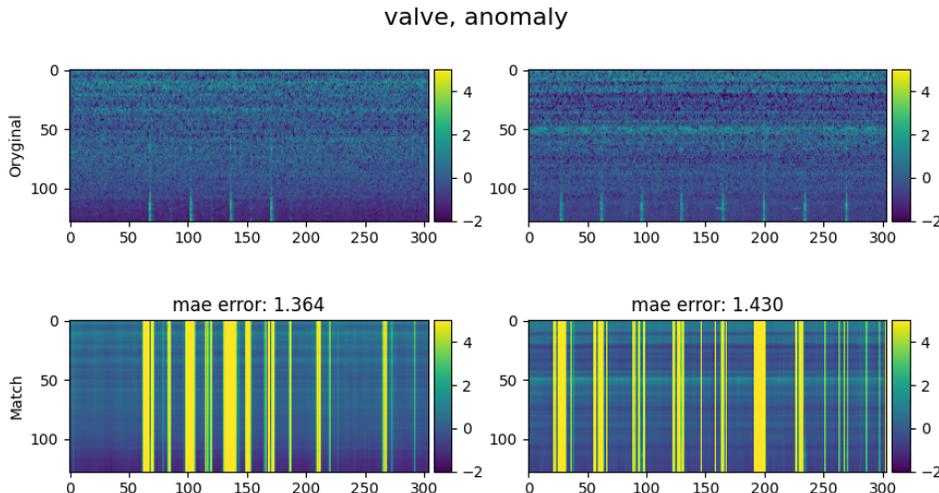


Figure 2: Two exemplary spectrograms of some valve anomalous samples with their corresponding reconstructions. In the top row, we have original spectrograms of the whole recordings. In the bottom row, we show reconstructed spectrograms with matching labels. The reconstruction error is mostly boosted in the windows containing anomalous valve sounds.

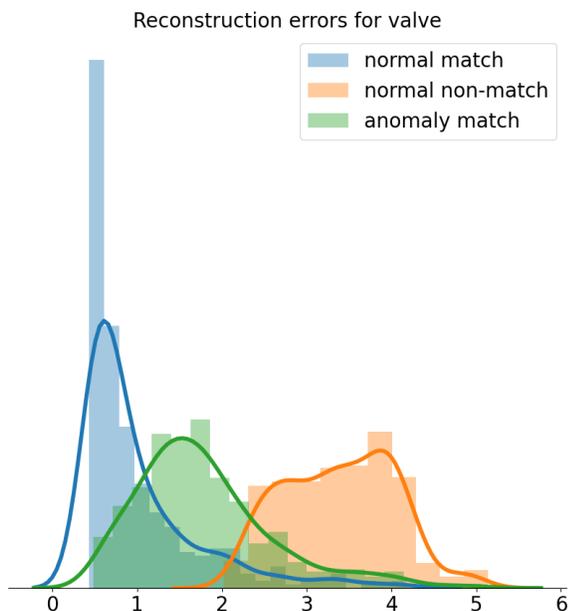


Figure 3: Distributions of reconstruction errors for all valve samples from the test split from the development dataset.

### 3.3. Parameters

We used  $\alpha = 0.75$ ,  $C = \mathbf{5}$  (the bold-face denotes that it is a constant vector with value 5 everywhere) with a frame size  $F = 10$  and number of mels  $M = 128$ . As for complexity for this setup, the encoder, decoder and conditioning has 175,792, 218,880 and 800 parameters respectively, making total of 395,472 parameters.

### 3.4. Training

We train our models using Adam optimizer with default parameters [25] setting mean absolute error as a loss function. For each machine, we train our network for 100 epochs with exponential learning rate decay by multiplying the learning rate by 0.95 every 5 epochs. For every epoch we randomly sample 300 frames from each spectrogram.

### 3.5. Ensemble

$\alpha$ ,  $C$  and number of mels  $M$  are hyperparameters for our model. Unfortunately, there is no single couple of these parameters that works best for all the machines. We set  $F = 10$  and did a grid search with  $\alpha \in \{0.9, 0.75, 0.5\}$ ,  $C \in \{\mathbf{0}, \mathbf{2.5}, \mathbf{5}, \mathbf{10}\}$ ,  $M \in \{128, 256\}$  trying mean square and mean absolute errors. We selected 3 models for each machine that maximize mAUC (the average of AUC and pAUC with  $p = 0.1$ ) on the test split from the development dataset, and for each machine we did an ensemble by selecting 3 weights such that the weighted anomaly score maximize mAUC. The ensemble is illustrated in Figure 4.

## 4. RESULTS

We develop and evaluate our system on the dataset from task 2 from DCASE 2020 Challenge [20], which consists of recording on 6 different machine types (Toy Car, Toy Conveyor, Fan, Pump, Slider and Valve). The available datasets consists of development and additional datasets. The development dataset consists of 3 or 4 IDs per machine while the additional dataset consist of 3 IDs per machine. All results in this section are evaluated on the test split from the development dataset using the area under the receiver operating characteristic (ROC) curve (AUC) and the partial-AUC (pAUC) with  $p = 0.1$ .

We conduct the ablation study starting from the DCASE baseline system and ending on the ensemble. We list the following steps, where each step is built upon the previous one:

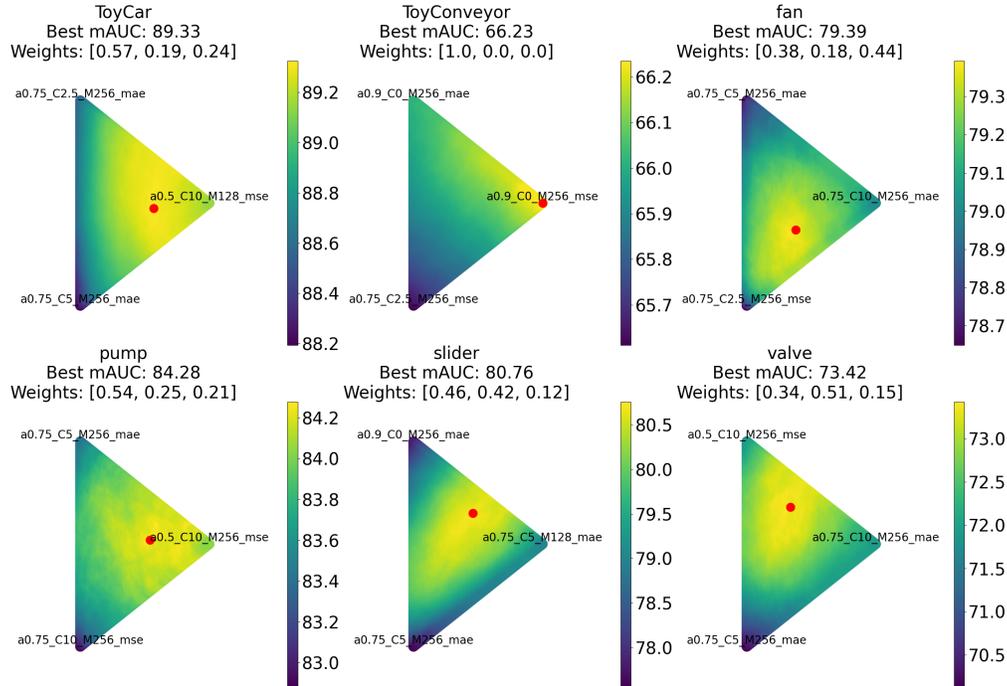


Figure 4: Visualisation of the ensemble of the best three models for each machine type. The vertices indicate mAUC of individual models and the interior points of triangles indicate mAUC of the convex combination of anomaly scores. That is, for given weights, new anomaly score is obtained as the linear combination of these weights with corresponding anomaly scores, and the mAUC is computed of that scores. The red dots indicate the weights that maximize mAUC.

Table 2: Ablation Study; the **boldface** denotes the best scores, and the underline denotes the greatest improvements.

	Toy Car		Toy Conveyor		Fan		Pump		Slider		Valve		Average	
	AUC	pAUC												
Baseline	78.77	67.58	<b>72.53</b>	<b>60.43</b>	65.83	52.45	72.89	59.99	84.76	66.53	66.28	50.98	73.51	59.66
Architect	81.18	68.78	70.37	57.21	70.63	56.44	74.15	<u>66.89</u>	85.68	67.92	66.47	50.79	74.75	61.34
Scaler	79.09	65.26	71.23	57.38	70.51	56.32	74.23	66.08	86.28	68.33	69.46	51.65	75.13	60.84
Condition	78.07	74.22	70.29	59.46	<u>77.45</u>	<u>70.32</u>	77.29	70.33	80.04	68.25	<u>78.26</u>	55.80	76.90	<u>66.40</u>
AddDataset	88.66	85.54	68.26	57.37	79.56	74.21	83.28	76.75	82.50	68.76	81.97	57.05	80.71	69.95
Ensemble	<b>91.28</b>	<b>87.36</b>	<u>72.23</u>	<u>60.23</u>	<b>81.80</b>	<b>76.98</b>	<b>88.27</b>	<b>80.28</b>	<b>86.80</b>	<u>74.71</u>	<b>84.56</b>	<b>62.28</b>	<b>84.16</b>	<b>73.64</b>

- Baseline - DCASE Baseline system.
- Architect - Encoder changed to the smaller one with 128, 64 and 32 units, latent layer set to 16 units, frame size set to 10, and training conducted as in subsection 3.4.
- Scaler - Spectrograms standardised frequency-wise to zero mean and unit variance.
- Condition - Conditioning added with  $\alpha = 0.75$  and  $C = 5$ .
- AddDataset - Training on the train splits from both development and additional datasets combining labels from these datasets.
- Ensemble - Ensemble as described in the subsection 3.5.

We summarise the results of the ablation study in Table 2. Even though results vary over machines, we will focus on the average scores which are placed in the last two columns of the Table. Our architecture changes as in Architect and normalization Scaler have minor influence on AUC and pAUC scores. Adding condi-

tioning layer in Condition significantly improves pAUC score, which proves that IDCAE is advantageous over standard AE. Moreover as expected, adding more IDs like in AddDataset improves both AUC and pAUC scores significantly. Finally, the ensemble from Ensemble due to its nature again significantly boosts both AUC and pAUC.

## 5. CONCLUSION

In this paper we introduced a method on how to enhance the auto-encoder designed for unsupervised anomaly detection. Assuming distinct IDs, we condition the latent space in order to enforce auto-encoder to reconstruct samples faithfully if and only if they are non-anomalous and conditioned with matching labels. Our method outperforms significantly the DCASE baseline system. In the ablation study, we proved that the conditioning we introduced is crucial for the performance improvement, and that feeding the system with more IDs gets even better results.

## 6. REFERENCES

- [1] S. Lecomte, R. Lengellé, C. Richard, F. Capman, and B. Ravera, “Abnormal events detection using unsupervised one-class svm - application to audio surveillance and evaluation -,” in *2011 8th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2011, pp. 124–129.
- [2] F. Aurino, M. Folla, F. Gargiulo, V. Moscato, A. Picariello, and C. Sansone, “One-class svm based approach for detecting anomalous audio events,” in *2014 International Conference on Intelligent Networking and Collaborative Systems*, 2014, pp. 145–151.
- [3] Y. Kawaguchi, T. Endo, K. Ichige, and K. Hamada, “Non-negative novelty extraction: A new non-negativity constraint for nmf,” in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018, pp. 256–260.
- [4] R. Giri, A. Krishnaswamy, and K. Helwani, “Robust non-negative block sparse coding for acoustic novelty detection,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, New York University, NY, USA, October 2019, pp. 74–78.
- [5] A. Ito, A. Aiba, M. Ito, and S. Makino, “Detection of abnormal sound using multi-stage gmm for surveillance microphone,” in *2009 Fifth International Conference on Information Assurance and Security*, vol. 1, 2009, pp. 733–736.
- [6] S. Ntalampiras, I. Potamitis, and N. Fakotakis, “Probabilistic novelty detection for acoustic surveillance under real-world conditions,” *Multimedia, IEEE Transactions on*, vol. 13, pp. 713 – 719, 09 2011.
- [7] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, “Optimizing acoustic feature extractor for anomalous sound detection based on neyman-pearson lemma,” in *2017 25th European Signal Processing Conference (EUSIPCO)*, 2017, pp. 698–702.
- [8] 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, 2019.
- [9] 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)*, 2019, pp. 865–869.
- [10] D. Conte, P. Foggia, G. Percannella, A. Saggese, and M. Vento, “An ensemble of rejecting classifiers for anomaly detection of audio events,” in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*, 2012, pp. 76–81.
- [11] R. Bardeli and D. Stein, “Uninformed abnormal event detection on audio,” in *Speech Communication; 10. ITG Symposium*, 2012, pp. 1–4.
- [12] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, “Anomalous sound detection based on interpolation deep neural network,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 271–275.
- [13] Y. Kawaguchi and T. Endo, “How can we detect anomalies from subsampled audio signals?” in *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017, pp. 1–6.
- [14] D. Oh and I. Yun, “Residual error based anomaly detection using auto-encoder in smd machine sound,” *Sensors*, vol. 18, no. 5, p. 1308, Apr 2018. [Online]. Available: <http://dx.doi.org/10.3390/s18051308>
- [15] Y. Koizumi, S. Saito, M. Yamaguchi, S. Murata, and N. Harada, “Batch uniformization for minimizing maximum anomaly score of dnn-based anomaly detection in sounds,” in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 6–10.
- [16] R. Chalapathy and S. Chawla, “Deep learning for anomaly detection: A survey,” 2019.
- [17] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boulton, “Toward open set recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, 2013.
- [18] P. Oza and V. M. Patel, “C2AE: Class conditioned auto-encoder for open-set recognition,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [19] C. Geng, S.-J. Huang, and S. Chen, “Recent advances in open set recognition: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, 2020.
- [20] <http://dcase.community/challenge2020/task-unsupervised-detection-of-anomalous-sounds>.
- [21] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda, and N. Harada, “Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring,” in *arXiv e-prints: 2006.05822*, June 2020, pp. 1–4. [Online]. Available: <https://arxiv.org/abs/2006.05822>
- [22] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, “ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection,” in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, November 2019, pp. 308–312. [Online]. Available: <https://ieeexplore.ieee.org/document/8937164>
- [23] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, “MIMII Dataset: Sound dataset for malfunctioning industrial machine investigation and inspection,” in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE2019)*, November 2019, pp. 209–213. [Online]. Available: [http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop\\\_Purohit\\\_21.pdf](http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop\_Purohit\_21.pdf)
- [24] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville, “Film: Visual reasoning with a general conditioning layer,” in *AAAI*, 2018.
- [25] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014. [Online]. Available: <https://arxiv.org/abs/1412.6980>