SELF-SUPERVISED CLASSIFICATION FOR DETECTING ANOMALOUS SOUNDS

Ritwik Giri[*]*, Srikanth V. Tenneti*[*]*, Fangzhou Cheng, Karim Helwani,*
Umut Isik, Arvindh Krishnaswamy

Amazon Web Services, Palo Alto, CA, USA

ABSTRACT

Representation learning, using self-supervised classification has recently been shown to give state-of-the-art accuracies for anomaly detection on computer vision datasets. Geometric transformations on images such as rotations, translations and flipping have been used in these recent works to create auxiliary classification tasks for feature learning. This paper introduces a new self-supervised classification framework for anomaly detection in audio signals. Classification tasks are set up based on differences in the metadata associated with the audio files. Synthetic augmentations such as linearly combining and warping audio-spectrograms are also used to increase the complexity of the classification task, to learn finer features. The proposed approach is validated using the publicly available DCASE 2020 challenge task 2: *Unsupervised Detection of Anomalous Sounds for Machine Condition Monitoring dataset*. We demonstrate the effectiveness of our approach by comparing against the baseline autoencoder model, showing an improvement of over 12.8% in the average AUC metrics using a MobileNetV2 based model. Ensembles of these models with our concurrently published Group-Masked Auto-Encoder won the top 3 positions in the DCASE 2020 challenge task 2.

***Index Terms*—** Self-supervised anomaly detection, machine audio, unsupervised anomaly detection, ArcFace

1. INTRODUCTION

Anomaly detection is a popular problem in machine learning, manifesting in several different flavors across diverse applications. In this work, our focus is on detecting anomalies in audio recordings. Given a training set with audio recordings labeled as "normal sounds", our goal is to flag sounds that are significantly dissimilar to these normal sounds. Typical application include acoustic scene monitoring systems, where unexpected/concerning events such as glass breaking, gun shots, babies crying are needed to be detected [1, 2, 3].

Early works in this field are based on supervised detection of anomalies [4, 5]. However, one limitation of supervised anomaly detection is that, in many practical applications, one may not have access to all possible anomalous sounds at training time. A more realistic assumption is that, we may have access to a collection of what constitutes as "normal" sounds. It may then be possible to flag events that are not seen in training as "anomalous". The recently released 2020 DCASE challenge introduces such a dataset, by combining two recent audio datasets recorded from machines, namely ToyADMOS [6] and MIMII [7]. The objective is to flag anomalous machines in the test set, when the training set itself consists of audio recordings from machines operating normally.

[*] Equal contribution.

Operation	t	c	n	s
Conv2D	-	16	1	2
Bottleneck	1	8	1	1
Bottleneck	6	16	2	2
Bottleneck	6	16	3	2
Bottleneck	6	32	4	2
Bottleneck	6	48	3	1
Bottleneck	6	80	3	2
Bottleneck	6	160	1	1
Conv2D	-	1280	1	1
Avg Pool	-	1280	1	-
Dense	-	num classes	1	-

Table 1: MobileNetV2 architecture used in this work. Each line describes a sequence of 1 or more identical (modulo stride) layers, repeated n times. All layers in the same sequence have the same number c of output channels. The first layer of each sequence has a stride s and all others use stride 1. All spatial convolutions use 3 3 kernels. The expansion factor t is always applied to the input size as described in [8]

The proposed approach in this paper is to use self-supervised classification based on metadata accompanying the audio files, to learn compact representations of the "normal" data. In different applications, this metadata could take various forms, such as locations of the microphones relative to the source, information about the ambient settings, information about the source itself, etc. By learning features, that can discern such "weak/auxiliary" labels that the audio data is accompanied by, we may be able to learn fine enough representations of the normal sounds to discern them from anomalies. We show that using such metadata on the 2020 DCASE dataset, yields significant improvements in the anomaly detection accuracy over the challenge baseline, as shown in Table 2. We also introduce a family of audio-inspired augmentations such as mix-up and spectral warping to create additional data/metadata pairs, to increase the complexity of the auxiliary classification task to learn finer features.

Outline: An overview of related prior works on self-supervised representation learning is presented in Sec. 2. Following this, our proposed approach is discussed in Sec. 3. Results are presented in Sec. 4. Finally, Sec. 5 concludes the papers, and talks about some future directions of this work.

2. PRIOR WORK

Within the framework of unsupervised representational learning, self-supervision involves withholding certain aspects of the data, and tasking a network to predict it. The features learnt by such a

Algorithm	ToyCar	ToyConveyor	Fan	Pump	Slider	Valve
Baseline	78.77 (67.58)	72.53 (60.43)	65.83 (52.45)	72.89 (59.99)	84.76 (66.53)	66.28 (50.98)
MobileNetV2, no aug.	87.66 (85.92)	69.71 (56.43)	80.19 (74.40)	82.53 (76.50)	95.27 (85.22)	88.65 (87.98)
MobileNetV2, with aug.	88.60 (86.15)	78.36 (64.40)	80.61 (78.11)	83.23 (76.33)	96.26 (84.61)	91.26 (84.82)
MobileNetV2, with ArcFace	88.16 (85.13)	57.01 (53.79)	79.44 (75.47)	80.14 (72.87)	92.15 (85.41)	86.86 (86.10)
ResNet, no aug.	88.69 (86.15)	65.04 (61.71)	78.87 (74.43)	83.50 (80.00)	90.49 (74.51)	86.24 (84.24)
ResNet, with aug.	89.95 (87.88)	70.30 (58.50)	80.78 (74.69)	85.81 (79.80)	90.57 (74.03)	84.45 (80.89)
GroupMADE	80.51 (71.89)	76.03 (60.70)	70.10 (53.62)	75.68 (68.97)	93.29 (83.46)	89.68 (70.95)
ens: MobileNetV2 + ResNetV2 with aug + GroupMADE	**95.57 (91.54)**	**81.46 (66.62)**	**82.39 (78.23)**	**87.64 (82.37)**	**97.28 (88.03)**	**98.46 (94.87)**

Table 2: Results over the development dataset

network are then used for further downstream tasks.

A variety of auxiliary tasks have been used in prior works for self-supervision, in many different applications. For example, [9] uses predicting co-occurance of audio and video streams to learn features, for applications such as source localization and action recognition. Predicting relative positions of image and video patches has been used in works such as [10, 11, 12]. Predicting frame ordering is used in [13, 14], while impainting images with missing patches is used in [15] to learn representations. A survey of many other works on self-supervised learning is presented in [16].

Of particular relevance to this paper are prior self-supervision works, that use classification based auxiliary tasks for learning features for anomaly detection. For example, in [17, 18, 19], the learning tasks involve networks to discriminate between multiple geometric transformations such as rotations, flipping and translations, applied to images. A different approach is presented in [20], where data is transformed onto a finite number of subspaces, before learning a feature mapping that maximizes the difference between inter-class and intra-class separations.

We employ a different strategy here. We leverage accompanying metadata, combined with different types of audio-inspired data augmentations to set up various classification tasks. Specifically for the DCASE dataset, for each machine type, we train networks on the normal data from all the machine IDs to:

1. Identify the machine ID of an audio sample. Apart from the provided samples, we also consider randomized linear combinations of the existing machine IDs to simulate new synthetic machine IDs. The network is then tasked to identify the mixing proportions of the original IDs.
2. Distinguish a sample from a set of synthetically perturbed versions of it. We use spectral warping to create the perturbations.

3. PROPOSED APPROACH

3.1. Auxiliary Tasks

In this subsection, we present the auxiliary classification tasks in detail, that have been used to learn compact representations of the normal data. For each machine type, we train a network to identify the machine ID that a recording belongs to. For example, for machine type ToyCar, the DCASE challenge dataset consists of 7 machine IDs. We train a network to identify the ID that a training sample belongs to. The softmax classification score of a test sample, measured at the output corresponding to its true machine ID, is taken as a measure of a sample's "inlier" score. Its negative is taken as the anomaly score.

We use two additional variations of the above idea:

3.1.1. Linear Combination Augmentations

Data from different machine IDs are combined in pairs using randomized linear combinations, and the network is trained to learn to identify the mixing proportions. For example, for an input sample that is a mixture of $(0.4 * x_1) + (0.6 * x_2)$, where x_1 and x_2 are samples from IDs 1 and 2, the network is trained to output $[0.4, 0.6, 0, 0, \ldots]$. KL divergence is used as the loss for this task. Linear combinations, both before and after taking the log, on mel-spectrograms have been considered.

3.1.2. Spectral Warping Augmentation

We perturb samples from existing machine IDs to create new machine IDs, using image warping. Specifically, for each machine ID, we synthesize two "perturbed machine ID", by warping along the frequency axis. Warping is performed using OpenCV's geometric image transformations, to map a frequency f in the original machine ID's spectrogram to:

$$f \rightarrow f_{max} \times \left(\frac{f}{f_{max}}\right)^{\alpha} \quad (1)$$

where f_{max} is the largest frequency in the spectrogram. The parameter α was chosen as 0.95 and 1.05 for the two perturbed classes respectively.

Augmentations have been widely used in prior audio processing literature, mostly in classification settings. A set of spectral augmentations that involve time warping, frequency masking, and time masking for audio were recently proposed in [21]. These were extended in [22] to include frequency warping, loudness control, and time length control. The frequency warping proposed therein is different from ours, and involves shifting a selected frequency by a fixed amount. Time stretching, pitch shifting, dynamic range compression and background noise were used in [23, 24]. In our experiments, we tried various pitch shifting and time stretching augmentations, but they did not improve the accuracy further than what could be achieved by the simple frequency warping we described above.

An important practical question is, are there criteria that can be used to predict which types of augmentations would work well for anomaly detection? We used the following insight: For anomaly

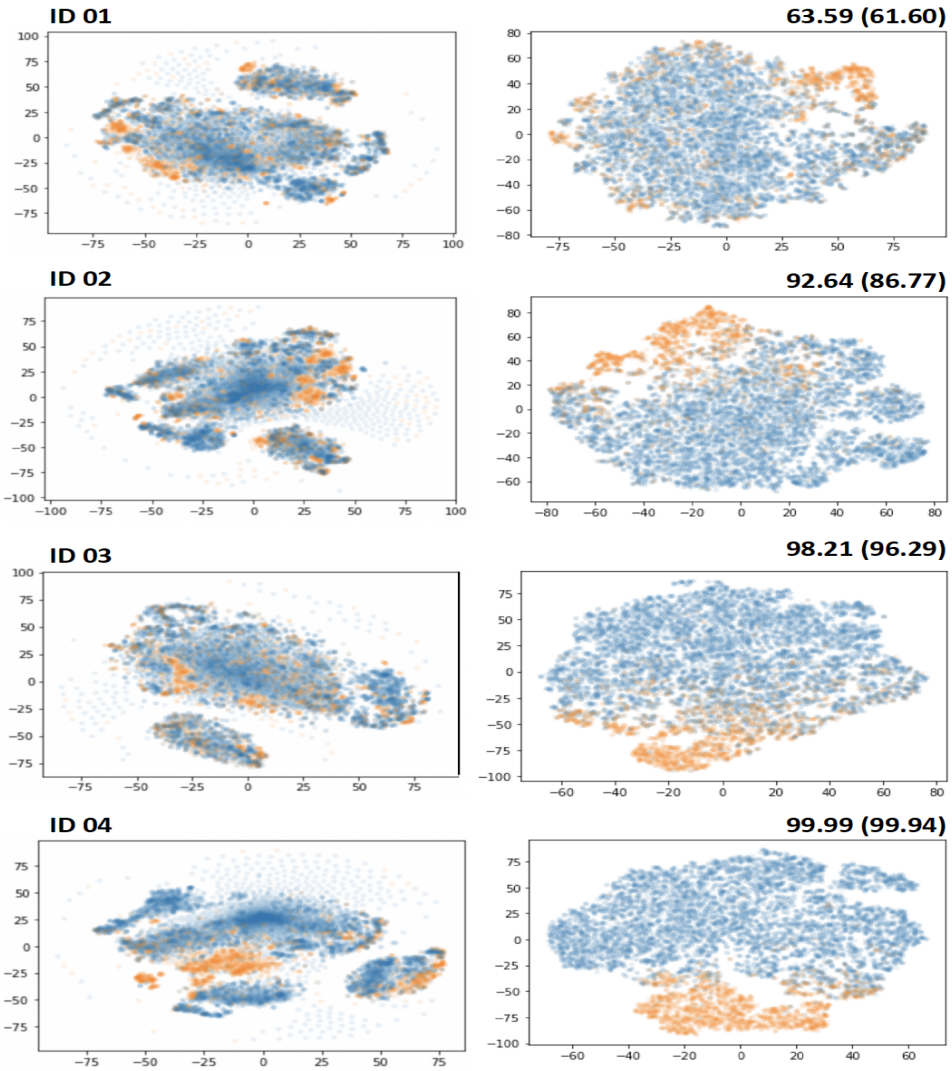

Figure 1: Per-ID ToyCar t-SNE plots of (left) the 64×128 input spectrograms and (right) 1280×1 features leant by the MobileNetV2 model trained without any augmentations. The AUC(pAUC) numbers are also shown for each machine ID. Blue points are normal samples, while the orange ones are the anomalous ones.

detection, the motivation for augmentations is quite different from that in classification settings. The augmentations referenced above were mostly used in classification settings in prior works, where they are meant to increase the robustness of a classifier to perturbations. However, in anomaly detection, especially in the way we used augmentations here, they are meant to increase the sensitivity of the models to fine differences in the inputs. If the perturbations in the augmentations are "too obvious", then the self-supervised classification task is too easy, and the model does not learn any finer discriminative features. For instance, while frequency masking and time masking were found to be very useful in [21], they were not very beneficial in our experiments here. In comparison, the more subtle frequency warping described in Eq. (1) was more useful.

3.2. Classifier Architectures

For the classification task, we employ two different architectures; MobileNetV2 and ResNet-50. MobileNetV2 is introduced in [8] as a computationally efficient convolutional neural network for visual recognition tasks such as object detection, classification and semantic segmentation. We use off-the-shelf Keras implementation of MobileNetV2, with the width multiplier parameter set to 0.5 . We set the "weights" argument as *None* while invoking the MobileNetV2 model to train it from scratch with random initialization. A summary of the architecture is given in Table 1.

The ResNet-50 [25] (Residual Network) model consists of 5 stages each with a convolution and Identity block. Each convolution block has 3 convolution layers and each identity block also has 3 convolution layers. For ResNet-50, we also use an off-the-shelf Keras implementation. Similar to the MobileNetV2 model, the ResNet-50 model is also trained from scratch with random ini-

Machine Type	Augmentation	Loss	Epochs
ToyCar	post-log linear combinations, spectral warping	KL-divergence	20
ToyConveyor	pre-log linear combinations	KL-divergence	10
Fan	pre-log linear combinations	KL-divergence	10
Pump	post-log linear combinations, spectral warping	KL-divergence	20
Slider	pre-log linear combinations	KL-divergence	10
Valve	post-log linear combinations	KL-divergence	10

Table 3: Combinations of augmentations that gave best results on DCASE 2020 challenge dev dataset.

tialization.

Even though, the Keras implementation of both the models allows one to load the pre-trained ImageNet weights, we do not use it because our aim is to learn features that are specific to different machine IDs, whereas ImageNet based initialization will force the model to learn more generic image classification related filters, for example edge detectors, which is not optimal for our task.

3.3. Inputs

The inputs to the classifier models are 64×128 images, which are the log-mel spectrograms, computed using the following parameters:

1. Each input 10s file is split into frames of length 64ms, with hop length of 32ms between frames.
2. 1024-FFT and 128 mel bins are used to featurize each frame.
3. 64 featurized frames are stacked to form a 64×128 image.
4. Successive 64×128 images have an overlap of 56 frames.

3.4. Additive Angular Margin (ArcFace) Loss

We also experiment with, recently proposed Additive Angular Margin Loss (ArcFace) for classifier models instead of traditional softmax loss to increase the discriminative power of the feature embeddings learned by the deep convnet models. Specifically this loss function helps our model to learn features that enhance the intra-class compactness along with the inter-class discrepancy. Details of this loss function can be found in the original paper [26]. ArcFace is used for the task of identifying machine ID of an audio sample.

4. RESULTS

Table 2 shows the receiver operating characteristics curve's area under curve (AUC) and partial area under curve (pAUC) [27] obtained on the DCASE challenge dataset. The development subset of the dataset is used in the results shown. For different machine types, different combinations of the augmentations mentioned above are observed to give the best results. The "with augmentations" results shown in Table 2 for MobileNetV2 are obtained using the augmentations indicated in Table 3. The benefits of augmentations are particularly evident for the ToyConveyor class. For ResNet, the "with augmentations" results were obtained using post-log linear combinations. The networks are trained using the ADAM optimizer, with a learning rate of 0.0001. The auxiliary classification task was observed to quickly converge for certain machine types. The number of epochs is varied across machine types accordingly, as shown in Table 3, to avoid over-fitting. For the ArcFace results shown in Table 2, the over-fitting problem is not as evident, and fixing the number of epochs at 25 for all machine types seem to suffice.

Fig. 1 shows t-SNE [28] plots of features extracted from four machine IDs from the ToyCar set, using the MobileNetV2 architecture and no augmentations. For machine IDs 02, 03 and 04, the separation between the normal and anomalous test samples is evident. For ID 01, self-supervised models do not perform as well. The AUC and pAUC numbers, also shown in Fig 1, also reflect this. To obtain these plots, 5000 normal points are sampled from the union of training and test sets, and 5000 anomalous points are sampled from the test set. For comparison purposes we also present the t-SNE plots of the 64×128 input log mel spectrograms, and the separation between normal and anomalous samples is evident in learned embedding space.

4.1. Ensembling with Group-MADE

In our experiments, we observe that the self-supervised approach described above performs very well when ensembled with other anomaly detection approaches. For instance, [29] proposes a Group-Masked Autoencoder based Density Estimation (Group-MADE) approach for audio anomaly detection. It is observed that ensembling the Group-MADE approach with the self-supervised approaches yields the best results in our experiments on all machine types. The ensmebling is done in the following fashion: we transform the anomaly scores of each model into a standardized scale, before combining them. The standardization transformation for any given model is applied in a per-machine ID fashion, by computing the mean and variance of its anomaly scores over the training data for that machine ID. The anomaly scores are then transformed to have zero mean and unit variance over the training data of that machine ID. Standardized anomaly scores across different models are then combined using mean or max ensembling. Table 2 shows the results of ensembling across multiple MoblieNetV2, ResNet-50 and Group-Made models for each machine type.

5. CONCLUSION

In this paper, a self-supervised classification based approach is presented for detecting anomalous sounds from machines. The classification task learns features that are discriminative enough to identify the multiple individual machine IDs within any given machine type. The proposed models significantly outperform the baseline autoencoder based approach that was provided by the challenge authors. Synthesizing the new data by taking linear combination of data from existing machines, as well as by warping the spectrograms, is employed to further increase the complexity of the self-supervised task.

6. REFERENCES

[1] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj, and T. Virtanen, "DCASE 2017 challenge setup: Tasks, datasets and baseline system," in *Proceedings of DCASE 2017 Workshop*, November 2017, pp. 85–92.

[2] 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*. IEEE, 2012, pp. 76–81.

[3] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1D convolutional recurrent neural networks," DCASE2017 Challenge, Tech. Rep., September 2017.

[4] M. Valera and S. A. Velastin, "Intelligent distributed surveillance systems: a review," *IEE Proceedings-Vision, Image and Signal Processing*, vol. 152, no. 2, pp. 192–204, 2005.

[5] G. Valenzise, L. Gerosa, M. Tagliasacchi, F. Antonacci, and A. Sarti, "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 21–26.

[6] 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.

[7] 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 DCASE 2019 Workshop*, November 2019, pp. 209–213.

[8] M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation," *CoRR*, vol. abs/1801.04381, 2018. [Online]. Available: http://arxiv.org/abs/1801.04381

[9] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648.

[10] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1422–1430.

[11] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European Conference on Computer Vision*. Springer, 2016, pp. 69–84.

[12] D. Kim, D. Cho, and I. S. Kweon, "Self-supervised video representation learning with space-time cubic puzzles," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8545–8552.

[13] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *European Conference on Computer Vision*. Springer, 2016, pp. 527–544.

[14] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang, "Unsupervised representation learning by sorting sequences," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 667–676.

[15] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[16] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[17] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 15 663–15 674.

[18] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 9758–9769.

[19] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *International Conference on Learning Representations*, 2018.

[20] B. Liron and H. Yedid, "Classification-based anomaly detection for general data," *arXiv preprint arXiv:2005.02359*, 2020.

[21] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[22] Y. Hwang, H. Cho, H. Yang, D.-O. Won, I. Oh, and S.-W. Lee, "Mel-spectrogram augmentation for sequence to sequence voice conversion," 2020.

[23] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, 2017.

[24] J. Abeßer, "A review of deep learning based methods for acoustic scene classification," *Applied Sciences*, vol. 10, no. 6, 2020.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[26] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[27] 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

[28] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[29] R. Giri, F. Cheng, K. Helwani, S. V. Tenneti, U. Isik, and A. Krishnaswamy, "Group masked auto-encoder based density estimation for audio anomaly detection," in *DCASE 2020 Workshop*.