USING UMAP TO INSPECT AUDIO DATA FOR UNSUPERVISED ANOMALY DETECTION UNDER DOMAIN-SHIFT CONDITIONS

Andres Fernandez, Mark D. Plumbley

Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, UK
{andres.fernandez, m.plumbley}@surrey.ac.uk

ABSTRACT

The goal of Unsupervised Anomaly Detection (UAD) is to detect anomalous signals under the condition that only non-anomalous (*normal*) data is available beforehand. In UAD under Domain-Shift Conditions (UAD-S), data is further exposed to contextual changes that are usually unknown beforehand. Motivated by the difficulties encountered in the UAD-S task presented at the 2021 edition of the Detection and Classification of Acoustic Scenes and Events (DCASE) challenge[1], we visually inspect Uniform Manifold Approximations and Projections (UMAPs) for log-STFT, log-mel and pretrained Look, Listen and Learn (L3) representations of the DCASE UAD-S dataset. In our exploratory investigation, we look for two qualities, *Separability (SEP)* and *Discriminative Support (DSUP)*, and formulate several hypotheses that could facilitate diagnosis and developement of further representation and detection approaches. Particularly, we hypothesize that input length and pretraining may regulate a relevant tradeoff between SEP and DSUP. Our code as well as the resulting UMAPs and plots are publicly available[2].

Index Terms— DCASE2021, Unsupervised Anomaly Detection, Domain Shift, UMAP, Interpretability

1. INTRODUCTION

The goal of Unsupervised Anomaly Detection (UAD) is to detect anomalous instances under the condition that only non-anomalous (i.e. *normal*) instances are available beforehand. This has relevance in monitoring applications where anomalous data is hard to collect whereas normal data is abundant. Unsupervised Anomaly Detection under Domain-Shift Conditions (UAD-S) presents an extra challenge: both normal and anomalous data can be exposed to *domain shifts*, i.e. changes in the environment that cause an impact in the data and are usually unknown beforehand. This can result in false negatives, if the detector is not sensitive enough, or false positives, if the detector does not tolerate or adapt to domain shifts.
In the audio domain, UAD has attracted attention as a promising Predictive Maintenance (PdM)[1] solution for Industrial Sound Analysis (ISA)[2]: Sound monitoring is non-invasive, is robust to occlusions, can be carried out during production, and anomalous sounds can signal issues long before critical faults occur. UAD-S is a natural extension to UAD, since even in controlled environments, new non-anomalous sources of sound can arise (e.g. due to maintenance work or upgrades). Given that it is difficult or undesirable to completely isolate the analyzed sound source from its environment, PdM-ISA solutions must be able to detect slight deviations (including short-duration events like clicks) while embracing stronger environmental changes.
The 2020 and 2021 editions of Detection and Classification of Acoustic Scenes and Events (DCASE) have incorporated UAD (2020, task 2) and UAD-S (2021, task 2) challenges. In both editions, a broad variety of approaches has been explored, but the results achieved in 2021 were significantly lower than in 2020 in terms of numeric performance. This may indicate a higher complexity of the 2021 task, independently of the choice of model and training scheme. In order to gain further insights, we propose to inspect the data distribution itself. Specifically, our proposed contributions are:

- We showcase a method for of UAD-S data exploration via visual inspection of Uniform Manifold Approximations and Projections (UMAPs) and assessment of 2 beneficial qualities: **Separability (SEP)** and **Discriminative Support (DSUP)**.
- We apply the proposed analysis procedure to the DCASE 2021 dataset, revealing insights on its macro- and microstructure.
- Based on the analysis and literature, we formulate a series of verifiable hypotheses that we believe can facilitate diagnosis and developement of further approaches.

Section 2 reviews UAD in DCASE. Section 3 describes our methodology. Section 4 describes our experiments. Section 5 presents and discusses some results. Section 6 concludes and proposes future work.

2. UAD IN DCASE

The DCASE 2020 dataset was a result of combining two recently curated datasets (ToyAdmos[3] and MIMII[4]), each featuring 10-second audio segments from different well-functioning *devices* (toy car, valve, fan, etc). Each segment was mixed with different background sounds to simulate real environments. The devices were then intentionally damaged/disrupted to provide anomalous data, which was only available for validation. The proposed models had to provide a real-valued anomaly score for each validation audio segment, and their performance was evaluated by ranking the Area Under ROC Curve (AUC) and Area Under Partial ROC Curve (pAUC) obtained across different devices[5].
The 10 best performing submissions in 2020 applied Deep Learning (DL), treating the developement dataset as training data. Some used different forms of data augmentation and additions from ex-

[1]DCASE website: `http://dcase.community`
[2]Online Resources:
Code: `https://github.com/andres-fr/dcase2021_umaps`
Webpage: `https://ai4s.surrey.ac.uk/2021/dcase_uads`

ternal datasets such as AudioSet[6] and Fraunhofer's IDMT-ISA-EE dataset[2]. For inference, most submissions directly applied the trained DL models, often via multi-task ensembles. The most popular alternative was to apply K-Nearest Neighbors (KNN) to learned embeddings of the training set[7, 8]. Most best performing models incorporated the information of the specific device upon training and evaluation. An exception was [9], which treated the data for all devices jointly. In general, a broad variety of models and training schemes achieved scores over 90%.

The 2021 UAD-S edition also combined two datasets (MIMII DUE[10] and ToyAdmos2[11]), extended in several aspects. Particularly, for each device, the 2021 dataset includes 7 *devices*, 6 *sections* and 2 *domains*, totalling 84 splits. Each device has 6 sections, which are balanced partitions of the data for evaluation purposes. Each section presents 2 domains: *source* and *target*, which differ in aspects like operating speed, machine load and environmental noise. As in 2020, the training data does not contain any anomalous sounds. The training data is also highly imbalanced: in all sections, only ∼0.3% of the training samples are on the *target* domain. Test data is balanced in terms of devices, splits and domains. The evaluation procedure is similar to 2020, but this time an overall *score* is given as the harmonic mean across all AUC and pAUC scores[12].

Out of 27 submissions for 2021, the autoencoder (AE) baseline ranked 21st with a *score* of ∼56.4% on the evaluation set. The 2021 winners[13] (∼66.8%) propose a particularly heterogeneous ensemble, combining different "*complementary*" representations, objectives and models, rather than "*relying on well-known domain adaptation techniques*". A related concept is the contrastive loss applied by [14], (9th place, ∼61%). Second place was achieved by [15] (∼65%) with a simpler setup based on applying non-parametric inference methods (Local Outlier Factor (LOF) and KNN) to trained embeddings. We note that, while it was observed that Representation Learning (RepL)-based methods generally underperform reconstruction-based ones for UAD[16], this does not seem to be the tendency here: reconstruction objectives are barely present in the top ranks, possibly due to sensitivity to domain shifts, and the emphasis is on representations, e.g. the importance of spectrogram hyperparameters noted by [13] and the implications and effectiveness of different embeddings analyzed by [17] (3rd place, ∼64.2%) which propose to use AdaCos[18]. An exception is [19] (4th place, ∼63.75%), which did propose a reconstruction-based method that compensates domain shift conditions.

Like in 2020, a variety of DL-related approaches were adopted, but the scores were substantially lower in the 2021 edition. Keeping in mind the small differences in the evaluation procedure, we argue that the emphasis on RepL-based methods, the relative success of non-parametric inference and the difficulty directly addressing domain shifts via well-known techniques point at the complexity of the task and the relevance of an adequate data representation, independently of the choice of model and training scheme. Therefore, in this exploratory work we propose to inspect the UAD-S data distribution itself. The goal is to gain further insights in order to facilitate diagnosis and developement of further approaches.

3. INSPECTING REPRESENTATIONS WITH UMAP

Generally, direct exploration of high-dimensional data like that encountered in the discussed approaches is difficult. Fortunately, when data is organized in lower-dimensional structures it can be possible to retain some of its structure while projecting the data onto as few as 2 dimensions, allowing for informative visual inspection.

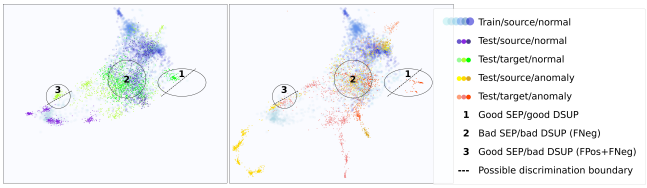

Figure 1: Excerpt from the device UMAP plot for *pump* with annotations in black. For example, region **1** presents good SEP and DSUP, since there is a simple boundary that clearly separates anomalies from normals and training data, and train/normal supports test/normal. *FPos* stands for false positives, and *FNeg* for false negatives. Each dot corresponds to 5 stacked log-mel frames. Color shades correspond to dataset *sections*. Training data is shown on both sides. Zoom to ∼1000% for detail.

UMAP[20] is a non-linear projection technique that has been shown to surpass alternatives like Principal Component Analysis (PCA) and t-SNE[21] in terms of speed, stability against reparametrizations and "*meaningfulness*" when applied on biological data[22]. Nevertheless, dimensionality reduction usually entails information loss, and artifacts arise: dense clusters may appear spread out and well-separated structures may collide when projected. For this reason, we restrict ourselves to the **assumption** that *if two regions appear separable on the UMAP projection, they are also separable on the original representation*. Crucially, *the opposite is not necessarily true*: two regions appearing mixed could be due to a projection artifact. With this in mind, we focus on two UAD-S properties:

- **Separability (SEP)**: In a projection with "good" SEP, a simple boundary can be drawn between anomalous and normal data with small error.
- **Discriminative Support (DSUP)**: If the training data provides set support for all normal data, and is separable from anomalous data, that set support can be directly used to discriminate anomalies. We consider that to be "good" DSUP.

Thus, by our assumption, if a given UMAP projection presents good SEP and DSUP, we infer that the corresponding high-dimensional representation has a simple boundary to separate normal from anomalous data (i.e. good SEP), and can be expressed using proximity to the training data (good DSUP). We argue that this is beneficial for the UAD-S task. Figure 1 illustrates this.

The concept of SEP could be quantified at a local level via e.g. Support Vector Machines (SVMs)[23]: given a set of data vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, labeled with -1 or 1 as $(y_1, \ldots, y_N)$, the SVM objective is to find the hyperplane parametrized by β that creates the biggest margin between both classes, allowing for some error ε, via the following objective[24, ch. 12]:

$$\min \|\beta\| \quad \text{s.t.} \begin{cases} y_i(x_i^T \beta + \beta_o) \geq 1 - \xi_i \\ \xi_i \geq 0 \\ \sum_i \xi_i \leq \varepsilon \end{cases} \qquad \forall i \in \{1, \ldots, N\} \tag{1}$$

Good SEP would be then achieved when a simple boundary separates normal and anomalous data with low ε. DSUP could be similarly quantified by comparing training and anomalous data, provided training data supports all normal data. But here we propose a complementary approach: to **qualitatively** assess SEP and DSUP via visual inspection of UMAPs. To that end, we render a series of dual plots, showing **anomalous test data on the right and normal test data on the left**.

4. UMAP FOR DCASE 2021

For our data sources, we merged the *Development* and *Additional Training* datasets[12] from DCASE 2021, task 2. To illustrate the role of external datasets, we also incorporated the 10-second cut variant of Fraunhofer's IDMT-ISA-EE dataset[2], and a custom subset of AudioSet consisting of 10-second segments from ∼40k unique videos. All audio files were converted to mono 16kHz, and $(-1, 1)$ normalization was applied. Based on high-performing systems from 2020, we computed amplitude spectrograms via the square modulus of Short-Term Fourier Transforms (STFTs) with 1024 samples per window and 50% overlap for all datasets. We then converted amplitudes to dB, yielding the log-STFTs. From the STFT spectrograms, we also computed 128-bin melgrams[25] and converted them to dB, yielding the log-mels. We ended up with ∼300k frames per *source* split and 927 frames per *target* split, totaling ∼13 million for *source* and 39k for *target*. Our AudioSet subset had then ∼12.2 million frames, and Fraunhofer ∼223k. We used `librosa`[26] for the above audio computations. We also computed 512-dimensional Look, Listen and Learn (L3) embeddings with a hop size of 0.1 seconds using `openl3`[27]. L3 embeddings encode longer-term relationships and this results in less frames (e.g. ∼3.5 million for our AudioSet). STFTs from environmental AudioSet videos were used for L3 audio training.

To encode temporal relationships, we stacked consecutive frames. In this work we explored 3 stack sizes: 1, 5 and 10. We computed a set of 2D UMAP projections for each of the 3 computed representations and 3 stack sizes. To get resolution at different scales, we computed one UMAP per device plus a global UMAP, totalling 72 UMAPs. Due to hardware limitations, the full datasets couldn't be processed and random samples were taken: For the per-device projections with stack size 1 and 5, we took 20k random samples for each validation split, and a maximum of 10k for every other split (recall that *target* training splits have just 927 frames). For the global projections with stack size 1 and 5, we took 2k samples per validation split, a maximum of 1k per training split, 50k for AudioSet and 50k for Fraunhofer. We also computed the stack size 10 UMAPs with no external data. For any given representation and stack size, we developed 3 kinds of scatter plots to enable different levels of detail: Global plots like Figure 2 are based on the global UMAPs and show the full dataset, coloring the different devices. Device plots like Figure 3 are based on the per-device UMAPs and color the different sections and domains. Section plots like Figure 5 are based on the per-device UMAPs, but they show a single section and color the specific audio files.

5. DISCUSSION

In general, SEP patterns across different devices and sections could be observed, but regions with both good SEP and DSUP were very hard to find, the best example we found has been already presented in Figure 1. Distinctively anomalous patterns are also scarce and do not appear to follow any obvious repeating patterns. We also observe that data from the *target* domain generally overlaps with the *source* domain.

The difference between ToyAdmos2 and MIMII DUE datasets can be seen at multiple scales: the ToyCar and ToyTrain clusters are clearly distinguishable from all other devices (see Figure 2), and the internal structure for the ToyAdmos2 devices is also apparently simpler than for MIMII DUE devices (compare Figures 3 and 4).

Another distinctive feature is the "horn" shape formed by the AudioSet data samples (e.g. Figures 5 and 4). Since energy spectrograms are non-negative, they are confined to the first quadrant, which is a cone with a vertex on the zero-energy point. By checking the energies, we have observed that the lowest-energy samples are highly concentrated on all observed "horns" (a few outliers get projected elsewhere). This indicates that, accounting for the logarithmic conversion to dB, the representations conserve the conic geometry and the observed "horn" shape likely corresponds to the tip of the cone, giving a sense of origin that can aid interpretation.

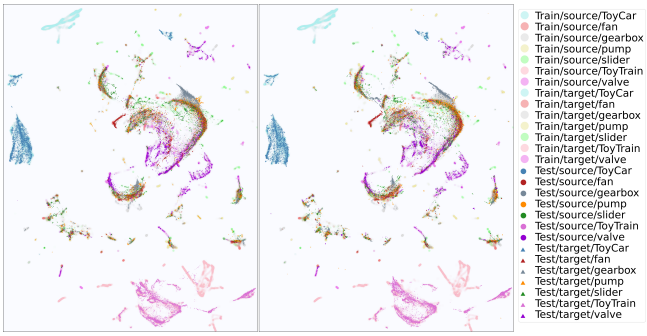

Figure 2: Global UMAP plot, sampled from the DCASE 2021 dataset. Each dot corresponds to 10 stacked log-STFT frames. Smaller dots correspond to anomalies on the right, and normal data on the left. Training data is shown on both sides. Zoom to ∼1000% for detail.

Interestingly, in all L3 device plots for *fan*, the training data appears almost completely separated from the test data (see e.g. Figure 4: the "shadows" have almost no test data on them). This means that a single L3 frame is enough to distinguish the *fan* training data from the test data fairly well. This should not be confused with a different phenomenon: some "shadow" clusters lack any overlying test data (e.g. the dark blue ones in Figure 3), but that is likely because those regions correspond to evaluation splits for which the challenge organizers did not release the test data. This is likely the case if the behavior is consistent across all representations.

Another particularity is that the AudioSet cluster appears to be smaller on the L3 representations, and the non-AudioSet data appears more scattered. This may be due to the fact that the L3 embeddings were trained on AudioSet and achieve a more compact representation there.

In the following we highlight several modelling **hypotheses** based on the above observations and the literature. We refer to our online resources for extensive results and code.

1. **Mixing ToyAdmos2 and MIMII DUE data may hinder performance**: Trivially distinguishable categories may lead to inefficient boundaries for anomaly discrimination. This was already proposed in [28].
2. **Temporal context and pretraining regulate a tradeoff between SEP and DSUP**: Generally, we observe that longer stack sizes provide better SEP. This makes sense because given enough length all audio files can be uniquely identified. But we also observed that this tends to scatter data apart and worsen DSUP. With pretrained embeddings, the observed tendency of concentrating the pretrained domain and scattering the rest may also entail a similar tradeoff. An ensemble with different tradeoff configurations may be beneficial. The complementary and contrastive approaches discussed in Section 2 may implicitly leverage this fact.

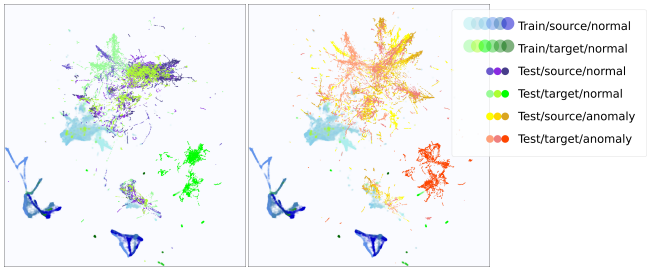

Figure 3: Device UMAP plot for *ToyCar*. Each dot corresponds to a single L3 frame. Color shades correspond to dataset *sections*. Training data is shown on both sides. Zoom to ∼1000% for detail.

3. **Normalization is a dominating factor**: If we interpret the data in Figure 5 as a cone with the vertex at the tip of the AudioSet "horn", renormalizing a frame would shift that frame roughly along the cone axis, which can greatly impact SEP and DSUP. The importance of proper normalization is supported by top-performing approaches like [9] and [17].
4. **Incorporating domain-related priors may help performance**: Bad DSUP only means that the the training data support can't be directly used for discrimination, but other kinds of prior knowledge still could be used to leverage the existing SEP. The 2021 dataset provides *domain*-related labels describing the domain shifts in the training data that could be used as priors. To the best of our knowledge, none of the participants made use of it, and could be a beneficial addition.

Lastly, we are aware of several methodological shortcomings:

1. We are only observing a subsample of the data, so extreme outliers are likely to be missed. Taking them into account may be crucial for successful analysis and detection.
2. As discussed in Section 3, data projections can only be used to confirm SEP and DSUP, not to discard them.
3. Qualitative, visual inspection may also be subject to perceptual biases, e.g. by color strength or shape consistency. Furthermore, plotting anomalous and normal data on different sides hinders the visual detection of slight differences.
4. Encoding temporal relations by stacking successive frames can lead to suboptimal representations due to e.g. conditional relations among frames, normalization and weighting issues.

Points 1 and 2 can be tackled by applying quantitative methods to the non-projected data, since the artifacts and size restrictions are imposed by the UMAP step. To overcome any issues related to high data volume and dimensionality, LOF and/or KNN-based methods like the ones used in [29, 7, 15] can be explored. Interactive exploration of the plots can help identifying small differences and overcoming perceptual biases. Representations that encode broader temporal context and other kinds of context can be explored to replace the frame stacks. Particularly, giving more weight to anomalous frames (or even ignoring very common frames) may help improving SEP and DSUP.

6. CONCLUSION AND FUTURE WORK

In this paper we performed an analysis of fixed and learned UAD-S data representations, based on the visual inspection of UMAPs and assessment of *separability* (SEP) and *discriminative support* (DSUP). In line with the difficulties encountered at the DCASE challenge, we did not find consistently good SEP and DSUP in any of the observed representations. The representations helped to expose potential issues in connection with the literature, and ways to address them. Despite the discussed methodological shortcomings, we defend that visual UMAP inspection can complement well other quantitative forms of analysis, and we hope that the software we provide can become a useful tool in the context of UAD-S.
The analysis could be enhanced with interactive plots providing sonification (to better understand the topology by hearing it), and highlighting corresponding datapoints across different representations. Analysis of further representations and techniques like X-vectors and the Teager-Kaiser energy operator[13] may also be of interest, as well as the impact of different embedding objectives on SEP and DSUP. Another possible extension could be to visualize the actual predictions of a system, extending this analyisis framework to supervised scenarios.

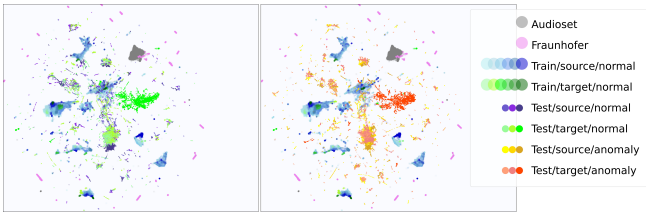

Figure 4: Device UMAP plot for *fan*. Each dot corresponds to a single L3 frame. Color shades correspond to dataset *sections*. Training data is shown on both sides. Zoom to ∼1000% for detail.

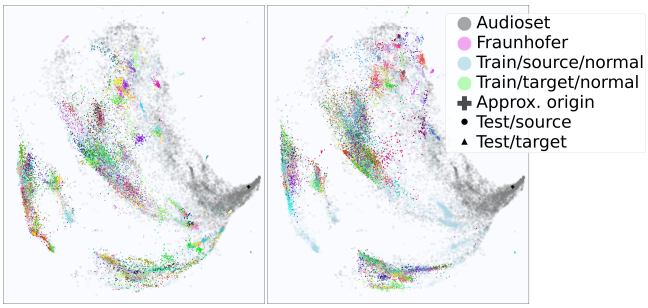

Figure 5: UMAP plot for *valve*, section 0. Each dot corresponds to a single log-mel frame. Smaller dots correspond anomalies on the right, and normal data on the left, and are colored by audio filename. After ignoring the 500 lowest-energy frames, the cross signals the average position of the following 100 ones. Zoom to ∼1000% for detail.

7. ACKNOWLEDGMENTS

The authors would like to thank Helen Cooper for the support throughout the research process. This work was supported by grant EP/T019751/1 from the Engineering and Physical Sciences Research Council (EPSRC) and made use of time on Tier 2 HPC facility JADE2, funded by EPSRC grant EP/T022205/1.

8. REFERENCES

[1] T. P. Carvalho, F. A. A. M. N. Soares, R. Vita, R. da P. Francisco, J. P. Basto, and S. G. S. Alcalá, "A systematic literature review of machine learning methods applied to predictive maintenance," *Computers & Industrial Engineering*, 2019.

[2] S. Grollmisch, J. Abeßer, J. Liebetrau, and H. Lukashevich, "Sounding industry: Challenges and datasets for industrial sound analysis," in *EUSIPCO 2019*.

[3] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for Anomalous Sound Detection," in *WASPAA 2019*.

[4] 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 *DCASE 2019 Proceedings*.

[5] 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 *DCASE2020 Proceedings*, July 2020.

[6] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*.

[7] Q. Zhou, "ArcFace based sound MobileNets for DCASE2020 task 2," DCASE2020 Challenge, Tech. Rep., July 2020.

[8] Y. Sakamoto and N. Miyamoto, "Anomaly calculation for each components of sound data and its integration for DCASE 2020 challenge task2," DCASE2020 Challenge, Tech. Rep., July 2020.

[9] J. Lopez, L. Hong, P. Lopez-Meyer, L. Nachman, G. Stemmer, and J. Huang, "A speaker recognition approach to anomaly detection," DCASE2020 Challenge, Tech. Rep., July 2020.

[10] R. Tanabe, H. Purohit, K. Dohi, T. Endo, Y. Nikaido, T. Nakamura, and Y. Kawaguchi, "MIMII DUE: Sound dataset for malfunctioning industrial machine investigation and inspection with domain shifts due to changes in operational and environmental conditions," *arXiv:2006.05822, 1–4*, 2021.

[11] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, M. Yasuda, and S. Saito, "ToyADMOS2: Another dataset of miniature-machine operating sounds for anomalous sound detection under domain shift conditions," *arXiv:2106.02369*, 2021.

[12] Y. Kawaguchi, K. Imoto, Y. Koizumi, N. Harada, D. Niizumi, K. Dohi, R. Tanabe, H. Purohit, and T. Endo, "Description and discussion on DCASE 2021 challenge task 2: Unsupervised anomalous sound detection for machine condition monitoring under domain shifted conditions," *arXiv:2106.04492, 1–5*, 2021.

[13] J. Lopez, G. Stemmer, and P. Lopez-Meyer, "Ensemble of complementary anomaly detectors under domain shifted conditions," DCASE2021 Challenge, Tech. Rep., July 2021.

[14] X. Cai, H. Dinkel, Z. Yan, Y. Wang, J. Zhang, and Y. Wang, "The small rice camera ready submission to the DCASE2021," DCASE2021 Challenge, Tech. Rep., July 2021.

[15] K. Morita, T. Yano, and K. Tran, "Anomalous sound detection using CNN-based features by self supervised learning," DCASE2021 Challenge, Tech. Rep., July 2021.

[16] S. Li, K. Tian, and R. Wang, "Unsupervised heart abnormality detection based on phonocardiogram analysis with beta variational auto-encoders," in *ICASSP 2021*.

[17] K. Wilkinghoff, "Utilizing sub-cluster AdaCos for anomalous sound detection under domain shifted conditions," DCASE2021 Challenge, Tech. Rep., July 2021.

[18] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] I. Kuroyanagi, T. Hayashi, Y. Adachi, T. Yoshimura, K. Takeda, and T. Toda, "Anomalous sound detection with ensemble of autoencoder and binary classification approaches," DCASE2021 Challenge, Tech. Rep., July 2021.

[20] L. McInnes and J. Healy, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv:1802.03426*, 2018.

[21] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[22] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. Kwok, L. G. Ng, F. Ginhoux, and E. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," *Nature Biotechnology*, vol. 37, 01 2019.

[23] B. Boser, I. Guyon, and V. Vapnik, "A training algorithm for optimal margin classifier," *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, vol. 5, 08 1996.

[24] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[25] S. S. Stevens and J. Volkmann, "The relation of pitch to frequency: A revised scale," *American Journal of Psychology*, vol. 53, p. 329, 1940.

[26] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, vol. 8, 2015.

[27] J. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019*.

[28] P. Primus, "Reframing unsupervised machine condition monitoring as a supervised classification task with outlier-exposed classifiers," DCASE2020 Challenge, Tech. Rep., July 2020.

[29] K. Durkota, L. Michael, L. Martin, and T. Jan, "Neuron-net: Siamese network for anomaly detection," DCASE2020 Challenge, Tech. Rep., July 2020.