

Disentangling physical parameters for anomalous sound detection under domain shifts

Kota Dohi
Hitachi Ltd.
Tokyo, Japan
kota.dohi_gr@hitachi.com

Takashi Endo
Hitachi Ltd.
Tokyo, Japan
takashi.endo.qf@hitachi.com

Yohei Kawaguchi
Hitachi Ltd.
Tokyo, Japan
yohei.kawaguchi.xk@hitachi.com

Abstract—To develop a sound-monitoring system for machines, a method for detecting anomalous sound under domain shifts is proposed. A domain shift occurs when a machine’s physical parameters change. Because a domain shift changes the distribution of normal sound data, conventional unsupervised anomaly detection methods can output false positives. To solve this problem, the proposed method constrains some latent variables of a normalizing flows (NF) model to represent physical parameters, which enables disentanglement of the factors of domain shifts and learning of a latent space that is invariant with respect to these domain shifts. Anomaly scores calculated from this domain-shift-invariant latent space are unaffected by such shifts, which reduces false positives and improves the detection performance. Experiments were conducted with sound data from a slide rail under different operation velocities. The results show that the proposed method disentangled the velocity to obtain a latent space that was invariant with respect to domain shifts, which improved the AUC by 13.2% for Glow with a single block and 2.6% for Glow with multiple blocks.

Index Terms—Machine health monitoring, Anomalous sound detection, Anomaly detection, Semi-supervised disentanglement learning, Normalizing flows

I. INTRODUCTION

As the number of skilled maintenance workers decreases worldwide, the demand for automatic sound-monitoring system has been increasing. Such systems aim to detect anomalous states of a machine from its sound.

Because anomalous sound data can rarely be obtained in practice, unsupervised anomaly detection methods are often adopted for these systems [1], [2]. Neural generative models such as a variational autoencoder (VAE) [3] and a normalizing flows (NF) [4], [5] are the most commonly used methods for unsupervised anomaly detection because of their high detection performance. These methods try to detect data with different distributions from normal data, without using anomalous data. However, not only a machine’s anomalous states but also changes in its physical parameters (domain shifts) or aging can affect the distributions of the normal data, which induces false positives when using unsupervised methods. Aging causes changes in data distributions over a long period of time, and these changes can be handled by continual learning or incremental learning [6]. On the other hand, domain shifts, which are the focus in this paper, can induce sudden, huge differences in data distributions, because physical parameters can change within a short period of time. Moreover, because

these physical parameters are often numerical values, it is impossible to collect a sufficient amount of data for all possible parameters. Accordingly, we need an unsupervised method that can handle domain shifts while requiring sound data with only a few sets of physical parameters.

In this paper, we develop an unsupervised anomalous sound detection method that can handle domain shifts due to changes in physical parameters. Our idea is to disentangle the factors of domain shifts and perform anomaly detection by using a space that is invariant with respect to these shifts. Specifically, we propose to constrain a neural generative model so that some of the latent variables represent factors of domain shifts and other latent variables represent components that are invariant with respect to domain shifts. As a result, the anomaly scores calculated using the latter latent variables are not affected by domain shifts but only by other variation factors such as a machine’s anomalous state, which can lead to fewer false positives.

II. PROBLEM STATEMENT

Anomalous sound detection is the task of identifying whether a machine is normal or anomalous according to an anomaly score calculated by a trained model from the machine’s sound. Each piece of input sound data is determined as anomalous if its anomaly score exceeds a threshold value. We consider unsupervised anomalous sound detection, in which only normal sound is available for training. We also assume that a machine’s physical parameters are only available during training. This assumption is realistic because sensors to measure the physical parameters may not be available in real-world operation, depending on environmental conditions. This problem setting is similar to that of DCASE 2021 Challenge Task 2 [7], in which machines have up to three different numerical and physical parameters that cause domain shifts.

III. RELATION TO PRIOR WORK

A. Semi-supervised disentanglement learning

Learning disentangled representations has been at the core of representation learning research [8]. Unsupervised disentanglement learning methods, in which a VAE with regularizers is commonly used to encourage disentanglement, have mainly been investigated [9]. However, it has been pointed out that

unsupervised disentanglement is impossible without inductive biases [10]. Semi-supervised disentanglement learning methods, on the other hand, explicitly use a few labeled pieces of data to disentangle factors of variation [11], [12]. Locatello et al. [11] trained a VAE with an added loss term to incorporate label information during training. Esser et al. [13] used an NF to learn transformations between latent spaces and semantic concepts. Esling et al. [14] used an NF to disentangle categorical tag information from a latent space. Our proposed method is a semi-supervised disentanglement learning method that uses an NF to disentangle numerical and physical parameters without an additional loss term.

B. Anomalous sound detection under domain shifts

For DCASE2021 Challenge Task 2, we published the MIMII DUE dataset [15], the first dataset for anomalous sound detection under domain shifts. In this dataset, we changed physical parameters between the source and target domain to induce domain shifts. The source domain data and a few samples from the target domain data were available during training. The top-ranked approaches in the challenge first used autoencoder-based methods or classifiers to extract embeddings from the data and then used likelihood-based methods like a Gaussian mixture model (GMM) to calculate anomaly scores from the embeddings. Kuroyanagi et al. [16] proposed to calculate anomaly scores by training a GMM for each domain on the autoencoder’s reconstruction errors. Sakamoto et al. [17] attained the highest scores for the target data by estimating the mean of the target data under an assumption that the mean changes between the source and target domains. Wilkinghoff [18] trained a classifier that discriminates each set of physical parameters to obtain embeddings. Our proposed method does not require the target domain data for training, as it explicitly uses numerical and physical parameters to obtain a domain-shift-invariant latent space that does not change between the source and target domains.

IV. CONVENTIONAL APPROACH

A. Semi-supervised disentanglement learning using VAE

Locatello et al. [11] proposed a semi-supervised disentanglement learning method to disentangle the factors of variations represented by a few labeled data instances, denoted as \mathbf{y} . They modified the loss function of VAE to incorporate supervision:

$$L = -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + \beta(D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}))) + \gamma \mathbb{E}_{\mathbf{x}, \mathbf{y}}[R(q_\phi(\mathbf{y}|\mathbf{x}))], \quad (1)$$

where \mathbf{x} denotes the input data, \mathbf{z} is the latent variables, β is a hyperparameter introduced in [9], γ is another hyperparameter, and $R(\cdot)$ is a function to induce supervised disentanglement.

B. Unsupervised anomaly detection using NF

Among neural generative models for unsupervised anomalous sound detection, an NF evaluates the exact likelihood of the input data and has shown better detection performance than other models, including a VAE [19].

The NF models a series of invertible transformations $f = f_1 \circ f_2 \circ \dots \circ f_K$ between an input data distribution $p(\mathbf{x})$ and a known distribution $p(\mathbf{z})$. The log likelihood of the input data can be calculated as

$$\log p(\mathbf{x}) = \log p(\mathbf{z}_0) + \sum_{i=1}^K \log \left| \det \left(\frac{d\mathbf{z}_i}{d\mathbf{z}_{i-1}} \right) \right|, \quad (2)$$

$$\mathbf{z}_i = f_i(\mathbf{z}_{i-1}), \quad (3)$$

where \mathbf{z}_0 is a latent vector that follows a known distribution such as the standard isotropic gaussian $N(0, 1)$, and z_i ($i = 1, 2, \dots, K$) is an intermediate latent vector. The anomaly score can be calculated as the negative log likelihood (NLL) of the input data [20]–[22],

$$a(\mathbf{x}) = -\log p(\mathbf{x}). \quad (4)$$

The NF has mainly been used as an unsupervised method, which fails to handle distribution changes due to domain shifts. Specifically, when a machine’s physical parameters change and the distribution of its normal sound data changes, unsupervised methods can output high anomaly scores, leading to false positives.

V. PROPOSED APPROACH

A. Learning of domain-shift-invariant latent space for anomaly detection using NF

To handle distribution changes due to domain shifts, we propose to disentangle the factors of domain shifts and construct a domain-shift-invariant latent space for anomaly score calculation. For this purpose, we use an NF because of its high expressive power.

First, we train an NF model to obtain domain-shift-invariant representations. As long as the latent variables follow an isotropic Gaussian, if we let some latent variables \mathbf{z}_d represent a factor of domain shifts, then the latent space constructed by other latent variables \mathbf{z}_c should be invariant with respect to that factor. To make \mathbf{z}_d represent a numerical and physical parameter v that causes domain shifts, we constrain \mathbf{z}_d to follow a Gaussian distribution $N(kv, 1)$, where k is a hyperparameter to induce stable training of the model. If a set of sound data with different values of the parameter v is available, then the model can learn to map input data into a latent space that shifts linearly with v . This forces some latent variables \mathbf{z}_d to represent the physical parameter v while making other latent variables \mathbf{z}_c invariant to that parameter.

Using the obtained domain shift-invariant latent space, we then calculate anomaly scores that are unaffected by domain shifts. The likelihood of the latent variables, $p(\mathbf{z}_0)$, can be factorized as

$$\log p(\mathbf{z}_0) = \log p(\mathbf{z}_c) + \log p(\mathbf{z}_d). \quad (5)$$

From (2) and (5), the likelihood of input data \mathbf{x} can be written as:

$$\log p(\mathbf{x}) = \log p(\mathbf{z}_c) + \log p(\mathbf{z}_d) + \sum_{i=1}^K \log \left| \det \left(\frac{d\mathbf{z}_i}{d\mathbf{z}_{i-1}} \right) \right|, \quad (6)$$

Because the latent variables \mathbf{z}_d are constrained to represent a factor v of domain shifts, they cannot be used for calculating domain shift-invariant anomaly scores. Also, the third term of (6) can be domain shift-dependent. Accordingly, the anomaly score in (4) can be calculated using only the first term of (6):

$$a(\mathbf{z}_c) = -\log p(\mathbf{z}_c). \quad (7)$$

B. Multi-scale architecture in NF for learning disentangled representations

A multi-scale architecture in an NF was first introduced in [23] and has since been commonly used in other NF models [24].

To apply our method in a multi-scale architecture, we modify only the last block of the architecture. Let c, h, w denote the channel size, height, and width of a feature vector, respectively, and assume that the architecture consists of N blocks. In the i th block ($i = 1, 2, \dots, N - 1$), an input \mathbf{x}_i with dimensions (c, h, w) is squeezed to give a feature with dimensions $(4c, h/2, w/2)$. After some flow transformations, half the channels are used as the input \mathbf{x}_{i+1} to the next block, while the other channels, \mathbf{z}_{ic} , are factored out; here, \mathbf{z}_{ic} is constrained to follow $N(0, 1)$. In the last block, we constrain half the channels, \mathbf{z}_d , to follow $N(0, 1)$ and the others, \mathbf{z}_{Nc} , to follow $N(kv, 1)$. Because only the latent variables in the last block are constrained to represent a factor of domain shifts, the model has to propagate domain-shift-dependent components to the last block. As a result, the latent variables factored out at each block can represent domain-shift-invariant components at each different scale. The shift-invariant latent space \mathbf{z}_c is obtained by concatenating the \mathbf{z}_{ic} from all the blocks:

$$\mathbf{z}_c = (\mathbf{z}_{1c}, \dots, \mathbf{z}_{Nc}). \quad (8)$$

The anomaly score is then calculated using (7).

VI. EXPERIMENTS

A. Dataset

We prepared two slide rails with the same machine ID, which operated with a designated velocity of 50–750 mm/s and a distance of 100–500 mm. Figure 1 shows examples of log-mel spectrograms of the sound with different operation velocities.

We first recorded normal sound data with 15 different velocities (50, 100, ..., and 750 mm/s) and three different distances (100, 250, and 500 mm), giving 45 different physical parameter sets in total. Each parameter set had 10 sound clips, with each clip consisting of a 10-second single-channel 16-kHz recording. We used data with different velocities for the training and test data: seven of the 15 velocities (100, 200, 300, 400, 500, 600, and 700 mm/s) were used for the training data, while other velocities were used for evaluating the model’s ability to disentangle velocities that were not included in the training data.

To evaluate the anomaly detection performance, we also recorded pairs of normal and anomalous sound data. Both the normal and anomalous sound data had 15 different velocities

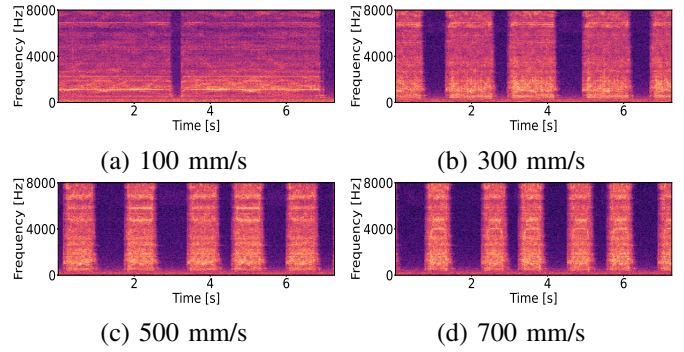


Fig. 1. Examples of log-mel spectrograms with different operation velocities (mm/s) and an operation distance of 500 mm.

and a fixed distance of 500 mm, with 10 sound clips for each velocity. The normal sound data was recorded using the same slide rail as the one used for the training data. The anomalous sound data was recorded using the other slide rail.

B. Experimental conditions

We used Glow [24] for the NF model because it is often chosen in out-of-distribution detection and anomaly detection tasks [19], [25]. We also used a VAE with the loss function given in (1) to compare the disentanglement performance with that of our proposed method.

To obtain input features, we applied the same procedure for Glow and the VAE. Each frame of the log-mel spectrograms was first computed with a length of 1024, a hop size of 512, and 128 mel bins. At least 313 frames were generated for each sound clip, and 64 consecutive frames with 48 overlappings were concatenated to generate the input features.

We prepared two Glow models with and without the multi-scale architecture. The Glow model with the multi-scale architecture (multi-scale Glow) had three blocks with five flow steps in each block, and each flow step had three CNN layers with 32 hidden layers. Therefore, \mathbf{z}_0 had 16 channels and \mathbf{z}_d had 8 channels. For the Glow model without the multi-scale architecture (single-scale Glow), the only difference from the multi-scale Glow was it had just one block. When using the proposed method, we constrained half the channels of the last block to follow $N(kv, 1)$, as described in Sec. V-B. In contrast, the single-scale Glow and the multi-scale Glow without the proposed method were trained by constraining all the latent variables to follow $N(0, 1)$.

The VAE model had 10 linear layers with 128 dimensions, except for the fifth layer, which had eight dimensions. The first dimension of the latent variables was constrained to represent the velocity v , and $R(\cdot)$ in (1) was calculated by taking the mean squared error between the value of the first latent variable and the actual velocity. For the hyperparameters we set $\beta = 1$ and $\gamma = 0.01$.

All models were trained for 1000 epochs by using the Adam optimizer [26] with a learning rate of 0.0001 and a batch size of 128.

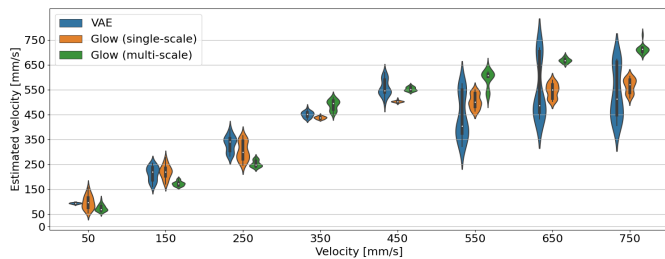


Fig. 2. Means of the latent variables for unseen velocities.

TABLE I

AUCs (IN %) OF SOUND DATA WITH THE SEEN VELOCITIES (VELOCITIES IN THE TRAINING DATA), UNSEEN VELOCITIES (VELOCITIES ONLY IN THE TEST DATA), AND ALL VELOCITIES.

Method	Seen velocities	Unseen velocities	All velocities
VAE using reconstruction error	52.9	57.5	55.2
VAE using KLD	47.0	52.4	49.9
VAE with loss in (1)	64.8	60.3	62.3
Single-scale Glow	89.1	72.7	78.6
Single-scale Glow using proposed method	97.7	87.3	91.8
Multi-scale Glow	100.0	86.0	91.9
Multi-scale Glow using proposed method	100.0	89.9	94.5

C. Results

We first estimated the unseen velocities from the trained models to investigate whether these velocities could be disentangled. For both the single-scale and multi-scale Glow, the estimated velocity was calculated by taking the mean of \mathbf{z}_d in the last block. For the VAE, the first dimension of the latent variables was used. The velocities estimated from input features in the same sound clip were averaged to give one estimation for each clip. Figure 2 shows the estimation for each unseen velocity. The multi-scale Glow showed the best performance, especially for the lower velocities of 50–250 mm/s and the higher velocities of 650 and 750 mm/s. Though the estimates tended to be larger for the medium velocities of 350–550 mm/s, they still showed positive correlations with the actual velocities. On the other hand, the VAE failed to estimate the higher velocities. This result shows that the Glow models were better at disentangling the factors of variations with their higher expressive power. The multi-scale Glow gave better estimation results than the single-scale Glow, especially for higher velocities. This may be because the multi-scale Glow has more learnable parameters and the multi-scale architecture enables extraction of representations at different scales.

We then calculated the anomaly scores for each sound clip. To evaluate the detection performance of each model, we calculated the area under the receiver operating characteristic curve (AUC) for the seen velocities, unseen velocities, and all velocities. We used the same model for all of the velocities. For the VAE, we calculated three different scores: the recon-

struction error (the first term in (1)) from a model trained using a conventional loss (first and second terms in (1)), the Kullback-Leibler divergence (KLD, second term in (1)) from the same model, and the KLD from a model trained using the loss in (1). For Glow, we used single-scale and multi-scale versions with and without the proposed method. Table I lists the results, which indicate that the proposed method improved the AUC by 13.2% for the single-scale Glow and 2.6% for the multi-scale Glow. In addition, the Glow models outperformed all the VAE models, even though VAE with the loss in (1) outperformed the other VAEs with the conventional loss. The proposed method showed greater improvement in the single-scale Glow than in the multi-scale Glow. This may be because the number of learnable parameters in the single-scale Glow was not enough to learn the distribution of the normal data, which made the effect of using the domain-shift-invariant latent space more evident.

Figures 3 and 4 show the anomaly scores for data using the single-scale Glow with the conventional method and the proposed method, respectively. In Fig. 3, the anomaly scores of the normal sound data with unseen velocities, especially for 50, 250, 450, and 750 mm/s, were higher than those of the normal and even the anomalous sound data with seen velocities. Because the normal data with unseen velocities cannot be used for training, the anomaly detector may detect normal samples with these velocities as anomalous samples. On the other hand, in Fig. 4, the normal sound data with unseen velocities showed about the same scores as with seen velocities, except for 50 mm/s. This result indicates that the proposed methods can lower the anomaly scores of normal sound data with unseen velocities by disentangling the operation velocity. For 50 mm/s, because the distribution of the normal data can significantly change around this lower velocity, the anomaly score for the normal data was higher with this velocity than with the other velocities. The anomaly scores of the anomalous sound data became higher for several velocities with the proposed method. As the model is trained to disentangle velocities by using the normal data, it may have disentangled components that did not represent the velocity of the anomalous sound data, thus raising the anomaly scores.

VII. CONCLUSION

We proposed an anomalous sound detection method that can handle domain shifts. The proposed method uses an NF model to disentangle the numerical and physical parameters of a machine, which gives a domain-shift-invariant latent space. Experimental results demonstrated that the proposed method disentangles the factors of domain shifts better than a VAE does, thus enabling improvement in the anomaly detection performance for data with unseen physical parameters.

REFERENCES

- [1] 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, pp. 271–275.

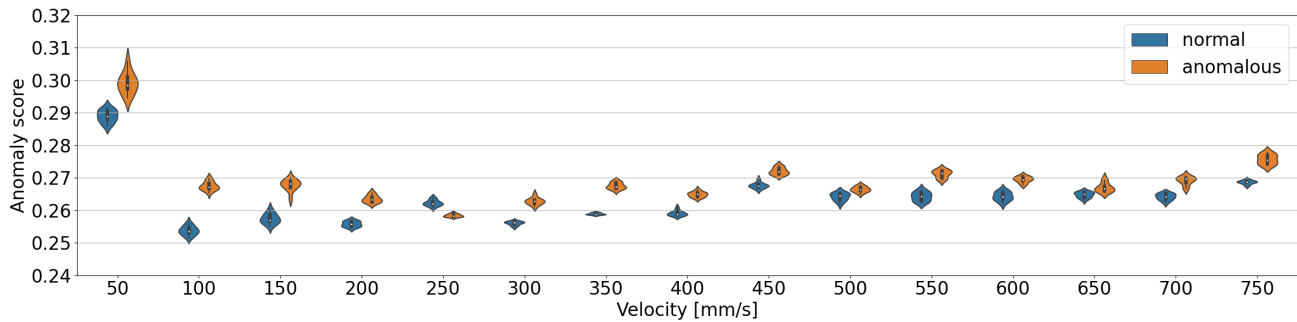


Fig. 3. Anomaly scores of sound data with the single-scale Glow and the conventional method.

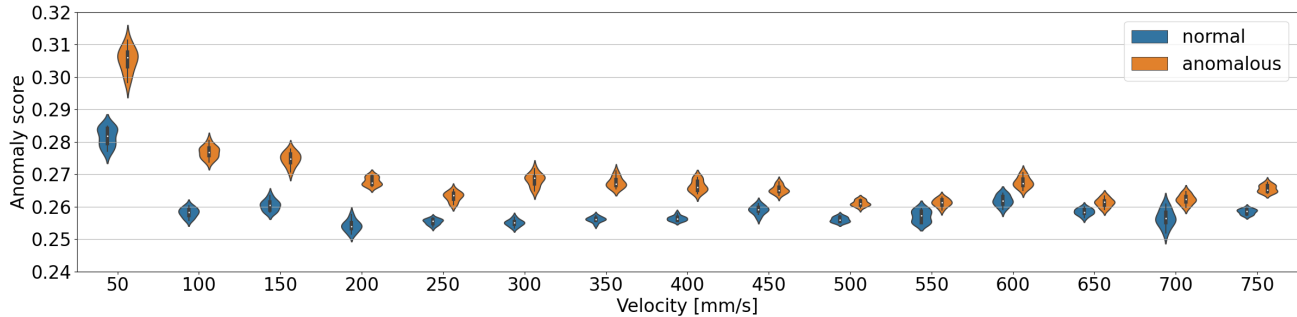


Fig. 4. Anomaly scores of sound data with the single-scale Glow and the proposed method.

- [2] Y. Koizumi, S. Saito, H. Uematsu, and N. Harada, "Optimizing acoustic feature extractor for anomalous sound detection based on neyman-pearson lemma," in *EUSIPCO*, 2017, pp. 698–702.
- [3] D. P. Kingma and Max Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.
- [4] E. Tabak and C. Turner, "A family of nonparametric density estimation algorithms," *Communications on Pure and Applied Mathematics*, vol. 66, pp. 145–164, 2013.
- [5] L. Dinh, D. Krueger, and Y. Bengio, "NICE: Non-linear independent components estimation," arXiv:1410.8516, 2015.
- [6] F. Wiewel and B. Yang, "Continual learning for anomaly detection with variational autoencoder," in *ICASSP*, 2019, pp. 3837–3841.
- [7] 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," in *DCASE2021 Workshop*, 2021.
- [8] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
- [9] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," 2017.
- [10] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," in *ICML*, 2019, pp. 4114–4124.
- [11] F. Locatello, M. Tschannen, S. Bauer, G. Ratsch, B. Schölkopf, and O. Bachem, "Disentangling factors of variation using few labels," in *ICLR*, 2020.
- [12] W. Nie, Z. Wang, A. Patel, and R. Baraniuk, "An improved semi-supervised VAE for learning disentangled representations," arXiv:2006.07460, 2020.
- [13] P. Esser, R. Rombach, and B. Ommer, "A disentangling invertible interpretation network for explaining latent representations," in *CVPR*, 2020.
- [14] P. Esling, N. Masuda, A. Bardet, R. Despres, and A. Chemla-Romeu-Santos, "Flow synthesizer: Universal audio synthesizer control with normalizing flows," *Applied Sciences*, vol. 10, no. 1, pp. 302, 2020.
- [15] 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," in *WASPAA*, 2021.
- [16] I. Kuroyanagi, T. Hayashi, Y. Adachi, T. Yoshimura, K. Takeda, and T. Toda, "Anomalous sound detection with ensemble of autoencoder and binary classification approaches," Tech. Rep., DCASE2021 Challenge, 2021.
- [17] Y. Sakamoto and N. Miyamoto, "Combine mahalanobis distance, interpolation auto encoder and classification approach for anomaly detection," Tech. Rep., DCASE2021 Challenge, 2021.
- [18] K. Wilkinghoff, "Utilizing sub-cluster adacos for anomalous sound detection under domain shifted conditions," Tech. Rep., DCASE2021 Challenge, 2021.
- [19] K. Dohi, T. Endo, H. Purohit, R. Tanabe, and Y. Kawaguchi, "Flow-based self-supervised density estimation for anomalous sound detection," in *ICASSP*, 2021.
- [20] M. Schmidt and M. Simic, "Normalizing flows for novelty detection in industrial time series data," in *ICML workshop on Invertible Neural Networks and Normalizing Flows*, 2019.
- [21] A. Ryzhikov, M. Borisyak, A. Ustyuzhanin, and D. Derkach, "Normalizing flows for deep anomaly detection," arXiv:1912.09323, 2019.
- [22] M. L. D. Dias, C. L. C. Mattos, T. L. C. D. Silva, J. Macêdo, and W. C. P. Silva, "Anomaly detection in trajectory data with normalizing flows," in *IJCNN*, 2020.
- [23] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP," in *ICLR*, 2017.
- [24] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *NeurIPS*, 2018.
- [25] V. Haunschmid and P. Praber, "Anomalous sound detection with masked autoregressive flows and machine type dependent postprocessing," Tech. Rep., DCASE2020 Challenge, 2020.
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.