MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Local Density-Based Anomaly Score Normalization for Domain Generalization

Wilkinghoff, Kevin; Yang, Haici; Ebbers, Janek; Germain, François G; Wichern, Gordon; Le Roux, Jonathan

TR2026-010 January 07, 2026

Abstract

State-of-the-art anomalous sound detection (ASD) systems in domain-shifted conditions rely on projecting audio signals into an embedding space and using distance-based outlier detection to compute anomaly scores. One of the major difficulties to overcome is the so-called domain mismatch between the anomaly score distributions of a source domain and a target domain that differ acoustically and in terms of the amount of training data provided. A decision threshold that is optimal for one domain may be highly sub-optimal for the other domain and vice versa. This significantly degrades the performance when only using a single decision threshold, as is required when generalizing to multiple data domains that are possibly unseen during training while still using the same trained ASD system as in the source domain. To reduce this mismatch between the domains, we pro- pose a simple local-density-based anomaly score normalization scheme. In experiments conducted on several ASD datasets, we show that the proposed normalization scheme consistently improves performance for various types of embedding-based ASD systems and yields better results than existing anomaly score normalization approaches.

IEEE Transactions on Audio, Speech and Language Processing 2026

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Local Density-Based Anomaly Score Normalization for Domain Generalization

Kevin Wilkinghoff, *Member, IEEE,* Haici Yang, Janek Ebbers, François G. Germain, *Member, IEEE,* Gordon Wichern, *Member, IEEE,* Jonathan Le Roux, *Fellow, IEEE*

***Abstract*—State-of-the-art anomalous sound detection (ASD) systems in domain-shifted conditions rely on projecting audio signals into an embedding space and using distance-based outlier detection to compute anomaly scores. One of the major difficulties to overcome is the so-called domain mismatch between the anomaly score distributions of a source domain and a target domain that differ acoustically and in terms of the amount of training data provided. A decision threshold that is optimal for one domain may be highly sub-optimal for the other domain and vice versa. This significantly degrades the performance when only using a single decision threshold, as is required when generalizing to multiple data domains that are possibly unseen during training while still using the same trained ASD system as in the source domain. To reduce this mismatch between the domains, we propose a simple local-density-based anomaly score normalization scheme. In experiments conducted on several ASD datasets, we show that the proposed normalization scheme consistently improves performance for various types of embedding-based ASD systems and yields better results than existing anomaly score normalization approaches.**

***Index Terms*—Anomalous Sound Detection, Domain Generalization, Domain Shift, Score Normalization**

I. Introduction

ANOMALY detection is the task of distinguishing between normal and anomalous data or inliers and outliers [1]. In recent years, research in anomalous sound detection (ASD) has been strongly promoted by an acoustic machine condition monitoring task belonging to the annual DCASE challenge [2]–[7]. This anomaly detection task presents several difficulties. First, only normal data are assumed to be available for training. This is motivated by the fact that anomalies typically occur only rarely, and that they are very costly to obtain on purpose, as this implies damaging possibly expensive machines or waiting for them to break down; moreover, it is challenging to capture the entire diversity of all possible anomalies with a finite set of training samples. Second, audio recordings in real-world factories tend to be very noisy. As a result, the embeddings obtained from those recordings need to be simultaneously very sensitive to changes in the target machine sound and insensitive to other acoustic events and background noise contained in the recordings, in order to reliably detect anomalies which, in comparison, may be very subtle. Last but not least, embeddings should be easily adaptable to shifts in the acoustic environment or changes in the acoustic sources (i.e., machines), so-called domain shifts. Ideally, users should only need to provide a very small number of samples to define how normal recordings sound like in data domains unseen during training, and a trained ASD model should still provide good results. Such an ability is referred to as domain generalization (DG) [8], [9]. Note that using target domain data for training the system was allowed in the DCASE challenge because it is impossible to tell if participants use the reference samples of the target domain for training or not. Still, in practice it is much more favorable that systems do not need to be re-trained for every possible domain shift and thus a priori knowledge about target domains encountered during testing in the form of domain-specific training samples ideally should not be utilized for training. An overview of methods for handling domain shifts for ASD related to DCASE can be found in [10].

State-of-the-art systems for ASD are based on projecting audio signals into a relatively low-dimensional embedding space and applying general outlier detection algorithms to these embeddings afterwards [11]. As only normal data is available for training, training a simple binary classifier to distinguish between normal and anomalous data is impossible. Therefore, the main difficulty when developing such ASD systems is to decide on a suitable loss function to train the embedding model that does not rely on anomalous data. The currently best-performing embeddings are obtained by utilizing auxiliary classification tasks based on meta information such as machine types, machine IDs, or settings [12]–[17], or on self-supervised learning (SSL) [18]–[22]. This auxiliary classification approach is also called outlier exposure [23] as samples belonging to other classes are used as proxy outliers [13]. Compared to one-class models such as autoencoders [24]–[28] that treat all signal components as equally important, models trained with an auxiliary classification task learn to closely monitor target signals and ignore background noise as well as other signals as long as they do not contain useful information to solve the classification task [29].

Kevin Wilkinghoff was with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA. He is now with Aalborg University and Pioneer Centre for Artificial Intelligence, Aalborg, Denmark (e-mail: kevin.wilkinghoff@ieee.org).
Haici Yang was an intern at MERL. She is now with Dolby Laboratories, Atlanta, GA, USA (e-mail: hy17@iu.edu).
Janek Ebbers was with MERL. He is now with Amazon, Aachen, Germany (e-mail: ebbers@nt.upb.de).
François G. Germain was with MERL (e-mail: fgermain@alumni.stanford.edu).
Gordon Wichern and Jonathan Le Roux are with MERL (e-mail: wichern@merl.com; leroux@merl.com).
Manuscript received April 19, 2021; revised August 16, 2021.

This work, which extends our prior work on normalizing embedding-based anomaly scores for DG [30], makes several new contributions. First and foremost, the evaluation of our proposed normalization approach is extended from a single to multiple state-of-the-art embeddings based on pre-trained models. Furthermore, additional existing normalization approaches are reviewed, discussed, and experimentally compared to our normalization approach. Last but not least, we provide the source code[1] for the conducted experiments.

The remaining parts of this paper are organized as follows. In Section II, related literature for computing embedding-based anomaly scores and normalizing them is discussed. In Section III, the proposed normalization approach is motivated, presented, and illustrated. The effectiveness of the proposed normalization scheme is experimentally evaluated in Section V using the setup from Section IV with several state-of-the-art embeddings on multiple datasets. In addition, a few ablation studies are carried out. The paper is concluded with a summary and possible extensions for future work in Section VI.

II. Related work

Anomaly scores for embedding-based ASD can be computed in several ways. The general idea is to project the data into an embedding space obtained by training a neural network and to apply general outlier detection approaches to these embeddings [31]. The underlying assumption is that anomalous test samples should substantially differ from normal test and training samples in the embedding space. Among these approaches for computing anomaly scores are distance-based approaches such as k-nearest neighbors (k-NN) [20], [32], [33], and (global) density-based approaches based on Gaussian mixture models (GMMs) [14], [34], [35] or the Mahalanobis distance [33]. In [36], distance-based approaches were shown to lead to better performance than density-based approaches in domain-shifted conditions because it is difficult to accurately estimate the density of the distribution in sparsely-represented target domains. The choice of the distance function depends on the loss function of the embedding model. For commonly used angular margin losses such as ArcFace [37], AdaCos [38], sub-cluster AdaCos [14], or AdaProj [39], the cosine distance is the most natural choice to measure the distance between samples. For other loss functions that do not act on the unit sphere, other distances such as the Euclidean distance can be used instead. Still, distance-based approaches are sensitive to data domains with different densities, which degrades the performance (see Section III-A). Note that this negative effect can even occur when only data belonging to the source domain are encountered if the data distribution consists of clusters with varying densities. Our proposed normalization approach is less sensitive to distributions with clusters of different densities as it also takes *local* density estimates into account.

There are several existing outlier detection approaches based on local density estimates. Their most prominent representative is the local outlier factor (LOF) [40], which compares the local density of a test sample to that of its nearest neighbors from a reference set. There are multiple variants and extensions of LOF. In [41], the neighborhoods are defined by balls with a certain radius instead of taking the nearest neighbors of each sample. Connectivity-based outlier factor (COF) [42] focuses on effectively handling low-density regions by introducing paths between nearest neighbors and utilizing decreasingly weighted lengths of the piece-wise path edges. Other variants of LOF aim to normalize scores to increase interpretability and to simplify the setting of a decision threshold. Local distance based outlier factor (LDOF) [43] takes the mean distance to the K nearest neighbors and normalizes it with the average pairwise distance between the K nearest neighbors. Local outlier probabilities (LoOP) [44] tries to normalize outlier scores such that different outlier score distributions are similarly scaled and can be directly interpreted as probabilities. This is achieved by introducing a so-called probabilistic distance of a query sample to a reference set. This probabilistic distance allows errors by defining spheres around the reference sample that contain all elements of a reference set only with a certain probability. Using the expected value of this distance, a probabilistic LOF is derived, whose standard deviation is used to normalize the resulting outlier scores. The anomaly scores of LOF-based outlier detection methods take the local density into account but are based on the distance in an embedding space of a test sample to the reference samples that are closest to it according to that same distance. Since different densities have a strong effect on the magnitude of the distance, the performance of these methods is still substantially degraded in domain-shifted conditions. In contrast, our proposed normalization scheme first alters the anomaly scores based on the local density and only then selects the closest neighbors. This reduces unwanted effects caused by clusters of reference samples with different densities and thus leads to anomaly scores that are more robust to domain shifts.

Normalizing scores is a well-investigated topic for various related applications. Apart from specific score normalization approaches for LOF, there are also works on normalizing outlier scores of arbitrary models such that the scores are contained in $[0, 1]$ and can be interpreted as probabilities [45]. Another example is to apply an additive similarity normalization [46], [47], based on [48], by normalizing individual distance-based outlier scores with the mean outlier score of the K nearest neighbors. Speaker verification is an additional application where score normalization approaches are frequently applied, under the name of score calibration [49], [50]. In contrast to our presented approach, these normalization approaches are not specifically designed for embedding-based ASD in domain-shifted conditions. However, there are also several works that reduce the domain mismatch by normalizing the scores in domain-shifted conditions. An example is a domain-wise standardization of the anomaly score distributions with the goal of aligning them [51]. Another example is cross-domain similarity local scaling (CSLS), which uses two different additive terms derived from samples of the source and target domains [52]. As we will show in Section V-A, our proposed anomaly score normalization approach leads to slightly better performance than the existing approaches while not requiring domain labels or specific training for each target

[1] https://github.com/merlresearch/anomaly-score-normalization

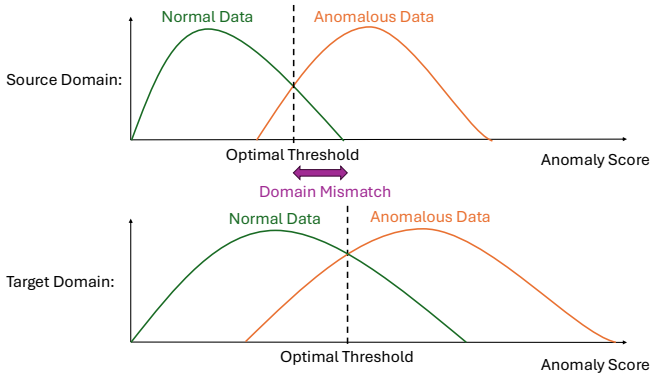

Fig. 1. Illustration of the domain mismatch between the anomaly scores of a source domain and a target domain. Normal and anomalous samples are usually less well separated in the target domain than in the source domain, which decreases domain-independent performance over the performance obtained for the source domain alone. Furthermore, the optimal decision thresholds for separating the scores belonging to the normal and anomalous data of different data domains differ substantially, which significantly decreases performance when using only a single threshold for both domains. Figure taken from [10].

domain, and allowing for independent evaluation of each test sample.

III. Methodology

A. Motivation

Domain shifts significantly decrease the performance of a trained ASD system because the distributions of the anomaly scores before and after the domain shift, i.e., in the source and target domains, are not necessarily well-aligned. This makes it difficult to separate normal and anomalous samples in both domains with a single decision threshold. This misalignment, referred to as domain mismatch, is illustrated in Fig. 1. Its main cause is that embedding models usually try to distribute the data into very compact clusters but are only trained with source domain data. Note that even if target domain samples are used for training, the effect is typically negligible due to the strong imbalance of the number of available training samples between the source and target domains. If the same embedding model is used to project data belonging to the target domain into the embedding space, the resulting clusters may have densities completely different from those of the source domain data. Therefore, the anomaly scores may also be scaled differently. Moreover, usually only a few reference samples are provided for the target domain, making it difficult to accurately estimate the distribution for density-based outlier detection approaches and leading to higher distances in general for distance-based approaches. Here, clusters with different densities correspond to sub-classes of the normal data. The main idea in this work is to reduce the domain mismatch caused by differently scaled anomaly scores. This is achieved by normalizing the scores so that the distances between samples from both domains are more uniformly distributed.

B. Distance-based anomaly score approaches

In this section, the proposed anomaly score normalization approach will be presented. To this end, we first introduce the notation and a simple distance-based baseline approach. Then, two variants of the score normalization scheme will be defined.

1) Baseline approach: Let us now properly define the baseline approach. Let $\mathcal{X}_{\text{test}}$ denote the set of test samples and $\mathcal{X}_{\text{ref}}$ denote a reference set of normal training samples that can belong to any or multiple data domains. Following best practices, all available training samples are used as reference samples. Then, the baseline approach for calculating an anomaly score based only on the nearest neighbor in the reference set is defined as

$$\begin{aligned}\mathcal{A}_{\cos}^{\text{NN}}(x, \mathcal{X}_{\text{ref}}) &:= \min_{y \in \mathcal{X}_{\text{ref}}} \mathcal{A}_{\cos}(x, y) \\ &:= \min_{y \in \mathcal{X}_{\text{ref}}} 0.5 \cdot \left(1 - \frac{\langle x, y \rangle}{\|x\|_2 \|y\|_2}\right) \in [0, 1].\end{aligned}$$

Note that, on the unit sphere, this equation simplifies to

$$\mathcal{A}_{\cos}^{\text{NN}}(x, \mathcal{X}_{\text{ref}}) = \min_{y \in \mathcal{X}_{\text{ref}}} 0.5 \cdot (1 - \langle x, y \rangle).$$

Samples that are outliers and thus not very representative of the underlying distribution of normal samples cause unwanted effects for this distance-based anomaly score calculation approach. These can be reduced by using k-NN instead of only measuring the distance to the closest neighbor [20], [33], or applying k-means to the reference set before computing the distance [36], which also reduces computational cost at inference time.

2) Proposed normalization approach: As mentioned in Section II, a distance-based baseline approach outperforms density-based approaches in domain-shifted conditions. However, to achieve good performance with a distance-based approach, the anomaly scores, i.e., the distances of test samples to their closest reference samples, need to follow a similar distribution for source and target domains. In most cases, this assumption is not true, causing a domain mismatch as explained in Section III-A. To reduce this mismatch, we propose two different approaches for normalizing the anomaly scores. Both approaches are based on the local density of the reference samples and differ only in how the local density is defined. For the first approach, the local density is based on the K nearest neighbors within the reference set. The local density of the second approach is defined by global weighted ranking pooling (GWRP) [53], which was also used in [54]. GWRP calculates the density with all reference samples but uses an exponentially decreasing weight based on the distance-based ranking of the reference samples to put higher emphasis on closer samples and thus somehow enforce a locality constraint. The idea for both approaches is to increase the scores measured against reference samples within regions with high density of embeddings while reducing the scores measured against samples within regions with low density.

Using the notation introduced for the baseline approach in Section III-B1, let $y \in X_{\text{ref}}$ denote an element of X_{ref}, and

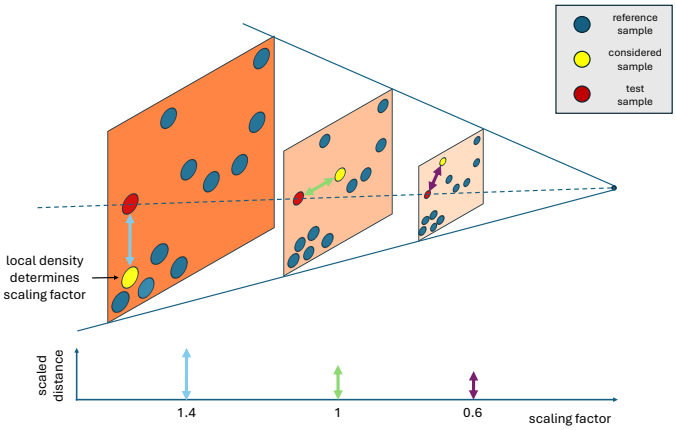

Fig. 2. Illustration of the impact of the ratio-based normalization approach on the selection of the reference points (in blue) that are to be considered the nearest for a given test point (in red). For three different considered points (in yellow) with similar distance to the test sample in the original embedding space, the scale at which they are compared to other points is shown. For one of them, the scaling factor is 1.4 because the point is in a dense neighborhood; for another one, the scaling factor is 1; and for the third sample in the sparse area, the scaling factor is 0.6. When assessing the distance between the test point and any of the considered samples, the distances should be computed in the corresponding rescaled planes. In the end, the reference point with the smallest scaled distance is selected. Here, the point in the sparse area with 0.6 scaling factor gets selected, despite the fact that all points initially had a similar distance to the test sample. For illustration purposes, planes are depicted here, although the normalization approach involves cosine distances on a sphere. For the difference-based normalization approach, reference samples are shifted based on their local densities instead of re-scaling the embedding space, which has a similar effect on the normalized distances but is more difficult to illustrate.

$y_k \neq y$ denote the k-th closest sample in $\mathcal{X}_{\text{ref}}$ to y. Then, the normalized anomaly scores are defined as

$$\mathcal{A}^{\text{K-NN}}_{\text{scaled}}(x, \mathcal{X}_{\text{ref}} \,|\, K) := \min_{y \in \mathcal{X}_{\text{ref}}} \frac{\mathcal{A}_{\cos}(x, y)}{\sum_{k=1}^{K} \mathcal{A}_{\cos}(y, y_k)} \in \mathbb{R}_+,$$

$$\mathcal{A}^{\text{GWRP}}_{\text{scaled}}(x, \mathcal{X}_{\text{ref}} \,|\, r) := \min_{y \in \mathcal{X}_{\text{ref}}} \frac{\mathcal{A}_{\cos}(x, y)}{\sum_{k=1}^{|\mathcal{X}_{\text{ref}}|-1} \mathcal{A}_{\cos}(y, y_k) \cdot r^{k-1}} \in \mathbb{R}_+,$$

where the hyperparameters $K \in \mathbb{N}^+$ and $r \in [0, 1]$ denote the number of closest samples to consider and the weight factor, respectively. In our experiments, we also consider a closely-related variant where the ratio is replaced with a difference, i.e., the normalization term is subtracted instead. It shall be emphasized that the normalization constants for all reference samples do not depend on the test sample and thus can be pre-computed, which ensures that there is no computational overhead at inference.

C. Discussion of the normalization effect

As stated before, the main goal of applying the normalization is to reduce the domain mismatch by having a more uniformly distributed embedding space across multiple domains and possibly existing sub-classes. Effectively, this is achieved by reducing the distance of reference samples with low local density to a given test sample compared to the distance to reference samples with high density. Therefore, if two reference samples are similarly close in the original embedding space, we favor the sample which is more isolated, i.e., the one that likely belongs to the target domain. Still, the distance to the reference sample itself is taken into account, ensuring that after normalization the closest neighbor is a reasonable sample and not just a random outlier. If we interpret the normalization as distorting the geometric distances from the test sample, this results in pushing reference samples with dense local neighborhoods away, while pulling samples with less dense local neighborhoods closer. This is illustrated in Fig. 2. As a result, the proposed normalization allows one to use a single decision threshold for samples in dense and sparse regions. These parameters control the degree of locality, ranging from $K = 1$ or $r = 0$, where the local density is based only on the nearest neighbor, to a global density $K = |X_{\text{ref}}|-1$ or $r = 1$, for which all other reference samples are considered. Note that for these edge cases the parameterizations with K and r are the same.

IV. Experimental Setup

A. Datasets

For the experimental evaluations of this work, the following five datasets were used: 1) the DCASE2020 ASD dataset [2] based on MIMII [55] and ToyADMOS [56], 2) the DCASE2022 ASD dataset [4] based on MIMII-DG [57] and ToyADMOS2 [58], 3) the DCASE2023 ASD dataset [5] based on MIMII-DG [57] and ToyADMOS2+ [59], 4) the DCASE2024 ASD dataset [6] based on [57], ToyADMOS2# [60], and additional samples recorded with the same setup as presented in IMAD-DS [61], and 5) the DCASE2025 ASD dataset [6] based on [57], ToyADMOS2025 [62], and additional samples recorded with the same setup as presented in IMAD-DS [61]. All these datasets focus on semi-supervised ASD for acoustic machine condition monitoring and consist of a development set and an evaluation set containing recordings of machines with real factory background noise. Each development and evaluation set is divided into a training split containing only normal data and a test split containing a mix of normal and anomalous data. A summary of the datasets can be found in Table I and more details can be found in the corresponding references. The main differences between the datasets will now be discussed.

DCASE2020: The DCASE 2020 ASD dataset consists of recordings belonging to six different machine types, namely "fan", "pump", "slider", "valve" from MIMII [55], and "ToyCar" and "ToyConveyor" from ToyAdmos [56]. For each machine type, there are six to seven specific machines in total, of which three to four belong to the development set and the remaining ones belong to the evaluation set. There are approximately 1000 normal training samples and around 400 test samples for each individual machine. Each recording has a length of 10 s and a sampling rate of 16 kHz. In contrast to the other two datasets, the DCASE2020 ASD dataset does not contain any data in domain shifted conditions, i.e., it essentially only consists of a source domain.

DCASE2022: The DCASE2022 ASD dataset explicitly features the DG problem for ASD. This means that 990 normal

TABLE I
OVERVIEW OF THE CONSIDERED DCASE ASD DATASETS.

	# machine types			# sections (per machine type)					# recordings (per section)			
									source domain		target domain	
Name	total	dev. set	eval. set	total	dev. set	eval. set	# ACT classes	split	normal	anomalous	normal	anomalous
DCASE2020 [2]	6	6	6	6-7	3-4	3	41	train	≤ 1000	0	0	0
								test	≤ 400	≤ 200	0	0
DCASE2022 [4]	7	7	7	6	3	3	242	train	990	0	10	0
								test	50	50	50	50
DCASE2023 [5]	14	7	7	1	1	1	167	train	990	0	10	0
								test	50	50	50	50
DCASE2024 [6]	16	7	9	1	1	1	141	train	990	0	10	0
								test	50	50	50	50
DCASE2025 [7]	14	7	7	1	1	1	172	train	990	0	10	0
								test	50	50	50	50

training samples belonging to a source domain and only 10 normal training samples belonging to a target domain are provided for each of the machines. The machine-specific test splits each consist of 200 samples that can belong to any of the two domains and may be normal or anomalous. In contrast to the DCASE2020 ASD dataset, the duration of individual recordings is not fixed but ranges between 6 s and 18 s. Moreover, additional meta information, referred to as attribute information, about specific machine settings or the acoustic environment are provided for each recording.

DCASE2023: The DCASE2023 ASD dataset increases the difficulty of the ASD task with the following two modifications, which are jointly referred to by the term "first-shot problem". The first modification is that the development set and the evaluation set contain mutually exclusive machine types. This ensures that ASD systems need to work well for arbitrary machine types as participants of the challenge cannot fine-tune their ASD systems to perform well for specific machine types. The second modification is that, for some machine types, only recordings of a single machine are available. This is more realistic but also substantially degrades the performance of discriminative embedding models as less information needs to be captured within the embeddings to obtain correct classification results.

DCASE2024: Apart from exchanging some of the machine types contained in the dataset, the DCASE2024 ASD dataset has the following key differences compared to the DCASE2023 ASD dataset. For some of the machine types, no attribute information is provided. This means that the embedding model needs to be trained without utilizing any additional meta information for some machine types, which substantially simplifies the imposed auxiliary classification tasks used for training the models and leads to less informative embeddings. Furthermore, some machines have an exclusive background noise, which means that embedding models trained to identify these machines may only be monitoring the noise. As the background noise does not carry any relevant information for detecting anomalous machine sounds, this strongly degrades the performance when using such an embedding model.

DCASE2025: The main modification of the DCASE 2025 ASD dataset is that the dataset also contains supplemental data for each machine type, which may either contain clean recordings of the target machines or only background noise. However, to simplify the evaluations in this work, we restrained from specific adaptations of the ASD systems and simply ignored the supplemental data. Thus, the main purpose of using the dataset in the experiments of this work is to have additional evaluations with recordings belonging to different machine types than the ones contained in previous versions of the dataset.

B. Evaluation metrics

To evaluate the performance of the ASD systems, we followed the official evaluation metrics for ASD experiments with each individual dataset [2], [5], [6]. For the DCASE2020 dataset, the area under the receiver operating characteristic curve (AUC-ROC) and partial area under the receiver operating characteristic curve (pAUC) [63] with $p = 0.1$ are computed for each section, and then the arithmetic mean is taken over all results. For the DCASE2023 and DCASE2024 datasets, the section-specific AUC-ROCs are computed individually for each domain by considering only the normal test samples belonging to the source or target domain but using all anomalous test samples regardless of the domain they belong to. The pAUCs are computed using all test samples belonging to any domain. Finally, the harmonic mean is taken over both AUC-ROCs and the pAUC of all sections.

C. Embedding models

For the experiments conducted in this work, four different embedding models were used. The first one (Direct-ACT) is trained with an auxiliary classification task based on meta information and SSL. The other three (OpenL3-raw, BEATs-raw, and EAT-raw) are pre-trained general-purpose models that are used without further training. In addition, we used an ensemble of ten Direct-ACT models to evaluate the normalization approach. In the following, the models are described and their implementation details are provided.

Direct-ACT: This embedding model is directly trained on the data using an auxiliary classification task and does not depend on any pre-trained models. More specifically, this model is based on [22] and consists of two feature branches. The first branch utilizes the magnitude of the full spectrum, i.e., the Fourier transform of the entire signal. The second branch computes the magnitude of the short-time Fourier transform (STFT) with a Hann window of size 1024 and a step size of 512, and applies temporal mean normalization to remove constant frequency information and thus make both features more different from each other. A different convolutional neural network is applied to each feature branch, mapping each of the features into a 256-dimensional embedding space. After that, both resulting embeddings are concatenated to obtain a single embedding. The auxiliary classification task (ACT) used to train the embedding model is defined based on the provided meta information, with one vanilla sub-task on the concatenated embeddings and an additional SSL sub-task on augmented embeddings. In the vanilla sub-task, the model has to discriminate, based on the concatenated feature embedding, between different values of the provided meta information, more specifically the machine types, machine IDs and additional parameter settings, or information about the acoustic environment. The additional SSL sub-task uses feature exchange [22], which randomly exchanges the embeddings of the two feature branches between samples and asks the model to predict whether the embeddings belong to the same sample or not, on top of predicting the meta information associated with each sample. For the entire dataset, a single embedding model is trained for 10, 5, and 15 epochs on the DCASE2020, DCASE2023, and DCASE2024 dataset, respectively, using Adam [64] with a batch size of 64. These particular hyperparameters were chosen by optimizing the performance on the corresponding development sets. The loss function is the AdaProj loss [39], which projects the data into class-specific linear subspaces on the unit sphere and ensures an angular margin between the classes. For data augmentation, mixup [65] with a uniformly distributed mixing coefficient is applied to the waveforms. Note that the only differences of this ASD system from the original system presented in [22] are that statistics exchange [21] is not used and that the sub-cluster AdaCos loss [14] is replaced with the AdaProj loss [39] with a sub-space dimension of 32.

OpenL3-raw: To show that the proposed normalization scheme does not depend on the choice of the embeddings and can also be utilized for pre-trained embeddings, the following ASD system based on OpenL3 embeddings [66], which is an open-source model for Look, Listen, and Learn embeddings [67], [68], is used for additional evaluations. First, all waveforms are divided into chunks using a sliding window with a length of $1\,\mathrm{s}$ and a hop size of $0.1\,\mathrm{s}$. Then, OpenL3 embeddings with a dimension of 512 are extracted from these chunks by computing mel spectrograms with 128 mel bins and using the model pre-trained on the environmental sound subset. Finally, the (temporal) mean over all embeddings belonging to the chunks is taken to obtain a single embedding for each recording that serves as an input feature for a task-specific embedding model. To calculate anomaly scores based on these embeddings, the mean squared error (MSE) is used instead of the cosine distance.

BEATs-raw: To provide more evidence for the claim that the presented normalization approach does not depend on the input features, additional experiments with BEATs embeddings [69] were conducted. In recent works on ASD [51], [70], systems based on these embeddings performed remarkably well. For the experiments conducted in this work, the official BEATs model pre-trained for three iterations on Audioset [71] without additional fine-tuning was used. To obtain a vector-sized embedding for each recording that serves as an input feature, the temporal mean of the patch embeddings is flattened to preserve frequency and channel information as proposed in [51]. This results in a single BEATs embedding with a dimension of 6144 for each audio recording. Again, the cosine distance is replaced with the MSE when computing anomaly scores. We also tried to train ACT models by replacing the mel spectrogram input with BEATs and openL3 embeddings, but found that just using the off-the-shelf embeddings resulted in superior performance, a result consistent with [51]. We also tried to fine-tune BEATs on the ASD task to replicate [70], but again the raw embeddings provided the best performance.

EAT-raw: Similar to BEATs, we also utilized EAT embeddings [72] as used by several other ASD systems [73], [74]. More concretely, we used the official checkpoint of the large EAT model pre-trained for 20 epochs on Audioset [71] without further fine-tuning. To obtain vector-sized embeddings, the temporal mean of the patch embeddings is concatenated with the extracted CLS embedding, resulting in embeddings with a dimension of 6912. As with the other pre-trained models, the MSE is used to compute anomaly scores.

Direct-ACT (Ensemble): As many state-of-the-art ASD systems are ensembles consisting of multiple models, we also utilized an ensemble for additional evaluations. This ensemble model is obtained by averaging the resulting anomaly scores of ten independently trained direct-ACT models with the architecture described above.

V. Experimental results

A. Comparison of normalization approaches

As a first experiment, different DG approaches are compared to our proposed normalization approach and a simple nearest neighbor baseline approach. More concretely, we compared four alternatives to our proposed approach: 1) K-means with $K = 16$ on the reference samples belonging to the source domain [36], 2) synthetic minority over-sampling technique (SMOTE) [75] to balance the number of samples for both domains by randomly interpolating between 4 reference samples of the target domain to synthesize additional target domain samples, 3) LOF [40] to compute local density-based anomaly scores, and 4) a domain-specific standardization of test score distributions [51]. All experiments are conducted on the DCASE2022, DCASE2023, DCAE2024, and DCASE2025 ASD datasets with all five ASD models, i.e., Direct-ACT, OpenL3-raw, BEATs-raw, EAT-raw, and Direct-ACT (Ensemble). The results can be found in Table II and the following observations can be made.

TABLE II
HARMONIC MEANS OF ALL AUCS AND PAUCS OBTAINED WITH DIFFERENT ANOMALY SCORE CALCULATION APPROACHES AND EMBEDDING MODELS. ALL INCLUDED VALUES ARE THE HARMONIC MEANS OVER THE PERFORMANCE METRICS OF ALL DEVELOPMENT AND EVALUATION SETS OF THE DCASE2022, DCASE2023, DCASE2024, AND DCASE2025 ASD DATASETS. FOR NON-DETERMINISTIC MODELS, MEANS OVER TEN INDEPENDENT TRIALS CORRESPONDING TO TEN TRAINED EMBEDDING MODELS ARE SHOWN. TO ALLOW FOR BETTER COMPARISON, THE SAME TEN TRAINED EMBEDDING MODELS WERE USED FOR ALL EVALUATIONS. TO OBTAIN MEANS OF THE SOURCE DOMAIN AS REFERENCE SAMPLES, K-MEANS WITH $K = 16$ WAS APPLIED. FOR SMOTE, FOUR NEIGHBORS WERE USED TO SYNTHESIZE SAMPLES. FOR LOF AND BOTH VARIANTS OF THE PROPOSED NORMALIZATION APPROACH, $K = 1$ WAS USED. HIGHEST NUMBERS IN EACH COLUMN AND BLOCK FOR EACH EMBEDDING MODEL ARE IN BOLD.

DG approach	domain	Direct-ACT	OpenL3-raw	BEATs-raw	EAT-raw	Direct-ACT (ensemble)	average
-	source	**66.0%**	**62.6%**	**65.0%**	**64.0%**	**67.2%**	**65.0%**
using source means [36]	source	62.4%	59.8%	63.2%	61.4%	63.6%	62.1%
SMOTE [75]	source	65.6%	60.6%	63.7%	63.1%	66.7%	63.9%
LOF [40]	source	58.9%	58.1%	60.0%	58.2%	62.1%	59.5%
standardization [51]	source	63.6%	59.0%	61.5%	60.2%	65.0%	61.9%
normalization (difference)	source	62.3%	59.3%	63.4%	62.2%	64.8%	62.4%
normalization (ratio)	source	60.6%	57.8%	62.0%	61.6%	63.1%	61.0%
-	target	52.5%	51.7%	54.2%	53.1%	52.6%	52.8%
using source means [36]	target	58.7%	52.3%	53.4%	53.8%	59.6%	55.6%
SMOTE [75]	target	55.6%	55.0%	57.5%	55.7%	56.0%	56.0%
LOF [40]	target	59.5%	56.5%	58.4%	56.6%	62.2%	58.6%
standardization [51]	target	60.0%	56.2%	59.4%	57.6%	61.2%	58.9%
normalization (difference)	target	60.9%	57.5%	**61.2%**	59.3%	63.4%	60.5%
normalization (ratio)	target	**61.5%**	**58.0%**	**61.2%**	**59.4%**	**64.1%**	**60.8%**
-	mixed	58.0%	56.7%	59.3%	58.5%	58.3%	58.2%
using source means [36]	mixed	61.1%	56.3%	58.2%	58.1%	62.3%	59.2%
SMOTE [75]	mixed	60.3%	57.9%	61.1%	60.0%	60.9%	60.0%
LOF [40]	mixed	59.9%	57.7%	59.8%	58.0%	62.9%	59.7%
standardization [51]	mixed	**62.7%**	58.5%	61.5%	60.0%	64.1%	61.4%
normalization (difference)	mixed	62.3%	**59.1%**	**62.8%**	**61.3%**	**64.5%**	**62.0%**
normalization (ratio)	mixed	61.8%	58.6%	62.0%	60.9%	64.3%	61.5%

TABLE III
QUALITATIVE COMPARISON OF DIFFERENT DG APPROACHES. APART FROM THE EFFECTIVENESS IN TERMS OF PERFORMANCE, IT IS SHOWN WHETHER DOMAIN LABELS OF THE REFERENCE SAMPLES ARE REQUIRED, WHETHER THE APPROACH ONLY ADAPTS TO SPECIFIC TARGET DOMAINS, AND WHETHER EACH TEST SAMPLE IS EVALUATED INDEPENDENTLY FROM OTHERS. THE PERFORMANCE PROVIDED FOR THE EFFECTIVENESS IS THE AVERAGE MIXED-DOMAIN PERFORMANCES OVER ALL DATASETS AND EMBEDDING MODELS AS CONTAINED IN TABLE II.

DG approach	requires domain labels	domain shift-specific	independent test samples	effectiveness (performance)
-	no	no	yes	low (58.2%)
using source means [36]	yes	no	yes	low/medium (59.2%)
SMOTE [75]	yes	yes	yes	medium (60.0%)
LOF [40]	no	no	yes	medium (59.7%)
standardization [51]	yes	yes	no	high (61.4%)
proposed approach (difference)	no	no	yes	high (62.0%)
proposed approach (ratio)	no	no	yes	high (61.5%)

First, it can be seen that the performance in the target domain is much worse than in the source domain when no DG approach is applied, which verifies the motivation for this work. Second, all approaches improve the performance in the target domain while reducing the performance in the source domain, although to different extents. Applying K-means to the reference samples of the source domain leads to moderate performance gains for Direct-ACT and the ensemble model, but actually reduces the performance for OpenL3-raw, BEATs-raw, and EAT-raw. Among all DG approaches, SMOTE achieves the best performance in the source domain but only leads to minor improvements in the target domain. Still, consistent improvements can be observed. LOF, on the other hand, leads to the worst performance in the source domain while achieving significant performance gains in the target domain. Overall, the performance of LOF in the mixed domain is similar to the one achieved with SMOTE. A domain-wise standardization of the anomaly scores and the difference-based and ratio-based variants of our proposed anomaly score normalization approach lead to the highest performance gains in the target domain while not reducing the performance in the source domain too much. As a result, these three approaches all achieve similar and highest overall performance with no clear winner in terms of pure performance. As a minor observation, it can be seen that our normalization approach also increases the performance gains obtained with additive ensembles. The most likely reason for this is that the anomaly scores are scaled more similarly after normalizing them.

Apart from comparing the effectiveness of the considered approaches in terms of performance, we also investigated other advantages and disadvantages of individual approaches in Table III. From that, it can be seen that our approach offers several advantages. In contrast to using K-means in the source domain, balancing with SMOTE, or using a domain-specific standardization of anomaly scores, our approach does not require domain labels for the reference samples. This is a clear

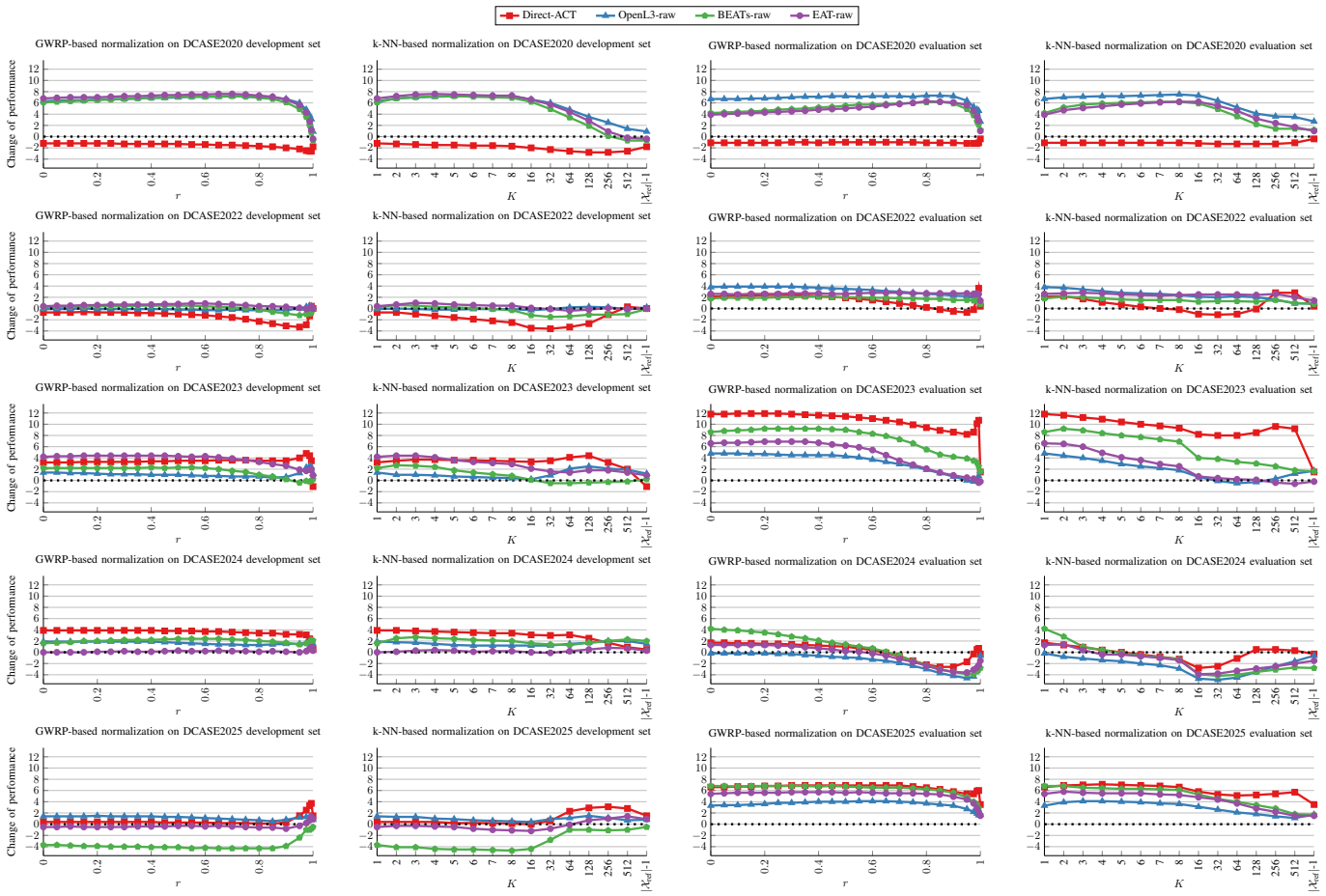

Fig. 3. Performance change for the ratio-based score normalization using different values of the GWRP constant r and number K for k-NN. The models Direct-ACT, OpenL3-raw, BEATs-raw, and EAT-raw are evaluated on the DCASE2020, DCASE2022, DCASE2023, DCASE2024, and DCASE2025 datasets. For Direct-ACT, mean results over ten independent trials are shown. Similar trends can be seen in the plots corresponding to the difference-based normalization.

advantage, since precisely defining domains and obtaining labels is difficult in real-world applications. Moreover, SMOTE and standardizing the score distributions are adaptations to specific domain shifts, which require modifying the system for every new domain shift that occurs. Last but not least, domain-specific standardization (as implemented in [74]), which is the only approach resulting in similar performance improvements as our proposed approach, estimates first- and second-order statistics of the anomaly scores of all test samples to modify the scores. Therefore, test samples cannot be independently evaluated, which is a strong restriction for real-world applications. We also tried to estimate these statistics using the anomaly scores of the training samples, but this only degraded the performance instead of improving it. In summary, our proposed approach yields the highest performance while not suffering from any of the mentioned disadvantages, and thus is a favorable choice among all DG approaches.

B. Sensitivity analysis with respect to the hyperparameters

As additional ablation studies, we investigated tuning the hyperparameters r of GWRP and K of k-NN for the respective parameterizations of the proposed normalization approach. The performance changes on the DCASE2020, DCASE2022, DCASE2023, DCASE2024, and DCASE2025 datasets for the embedding models Direct-ACT, OpenL3-raw, BEATs-raw and EAT-raw when normalizing the scores are depicted in Fig. 3. The following observations can be made.

First and most importantly, the proposed approach improves the performance for most datasets if local density estimates are used. There are a few exceptions to this, for example the performance obtained with Direct-ACT on the development set of the DCASE2020 and DCASE2022 datasets, which slightly degrades. However, since the DCASE2020 dataset does not contain any domain shifts, the normalization approach does not need to be applied. Interestingly, the performance gains for OpenL3-raw, BEATs-raw, and EAT-raw on this dataset are still substantial, which indicates that the proposed normalization approach is beneficial for off-the-shelf embeddings when not using any fine-tuning. For the DCASE2022 dataset, the performance degradation with direct-ACT can be explained by a slightly larger decrease in performance on the source domain compared to the improvement achieved on the target domain, resulting in a marginal overall decline. Other exceptions are the performance obtained with OpenL3-raw on the evaluation set of the DCASE2024 dataset and with BEATs-raw on the development set of the DCASE2025 dataset, which significantly degrades, for an unknown reason.

TABLE IV
HARMONIC MEANS OF ALL AUCS AND PAUCS OBTAINED WITH DIFFERENT ASD SYSTEMS ON A REPRESENTATIVE GROUP OF ASD DATASETS USED AS BENCHMARKS IN SEVERAL RECENT PEER-REVIEWED WORKS. FOR ALL OF OUR PROPOSED ASD SYSTEMS, A RATIO-BASED ANOMALY SCORE NORMALIZATION WITH $K = 1$ WAS USED. WHENEVER APPLICABLE, MEANS OF ALL INDEPENDENT TRIALS ARE SHOWN. HIGHEST NUMBERS IN EACH COLUMN ARE IN BOLD.

		DCASE2020 dataset [2] (no domain shifts)			DCASE2023 dataset [5] (first-shot DG)			DCASE2024 dataset [6] (first-shot DG with less meta data)		
ASD system	trials	dev. set	eval. set	arithm. mean	dev. set	eval. set	harm. mean	dev. set	eval. set	harm. mean
Direct-ACT	10	90.7%	90.2%	90.5%	68.4%	68.0%	68.2%	62.0%	54.7%	58.1%
OpenL3-raw	1	75.2%	77.5%	76.4%	60.5%	63.8%	62.1%	56.6%	55.7%	56.1%
BEATs-raw	1	81.5%	82.2%	81.9%	64.8%	67.6%	66.2%	58.1%	62.4%	60.2%
EAT-raw	1	79.3%	79.9%	79.6%	65.0%	66.1%	65.5%	56.8%	58.9%	57.8%
Direct-ACT (ensemble)	1	**94.2%**	93.3%	**93.8%**	**71.3%**	72.4%	**71.8%**	**65.2%**	56.5%	60.5%
Koizumi et al. [2]	1	66.6%	70.0%	68.3%	–	–	–	–	–	–
Wilkinghoff [14] (single model)	1	90.7%	92.8%	91.8%	–	–	–	–	–	–
Wilkinghoff [14] (ensemble)	1	–	94.1%	–	–	–	–	–	–	–
Liu et al. [76]	1	89.4%	–	–	–	–	–	–	–	–
Harada et al. [77]	1	–	–	–	56.9%	61.1%	58.9%	55.4%	56.5%	55.9%
Wilkinghoff [36]	5	–	–	–	62.8%	63.0%	62.9%	–	–	–
Hou et al. [78]	1	88.8%	92.0%	90.4%	–	–	–	–	–	–
Wilkinghoff [22] (single model)	5	–	–	–	64.2%	66.6%	65.4%	–	–	–
Wilkinghoff [22] (ensemble)	5	–	–	–	–	70.9%	–	–	–	–
Han et al. [79]	1	–	–	–	64.3%	–	–	–	–	–
Zhang et al. [80]	1	–	–	–	–	71.3%	–	–	–	–
Jiang et al. [70]	1	90.9%	**94.3%**	92.6%	64.2%	**74.2%**	68.8%	–	–	–
Wilkinghoff [39]	10	–	–	–	62.9%	64.5%	63.7%	–	–	–
Saengthong et al. [51]	1	74.7%	–	–	–	73.8%	–	–	–	–
Yin et al. [81]	1	–	–	–	68.1%	–	–	–	–	–
Fujimura et al. [82]	5	–	–	–	67.2%	68.8%	67.6%	–	–	62.0%
Yin et al. [83]	1	–	–	–	–	–	–	–	**67.1%**	–
Jiang et al. [70] presented in [84]	5	–	–	–	–	–	–	62.5%	65.6%	64.0%
Jiang et al. [84]	5	–	–	–	–	–	–	64.1%	66.0%	**65.0%**
Fujimura et al. [73]	4	90.4%	93.5%	91.9%	64.0%	72.0%	67.8%	59.9%	61.5%	60.7%

The second main observation is that optimal hyperparameter values are strongly dataset-dependent and partly embedding-dependent. Although there are several cases where increasing the hyperparameter slightly improves the performance, e.g., for OpenL3-raw, BEATs-raw, and EAT-raw on the DCASE2020 dataset or for all embedding models on the DCASE2025 development set, these improvements are not consistent on all datasets. This is particularly evident on the evaluation set of the DCASE2024 dataset, where performance drops rapidly when increasing the value of K and r. In contrast, very small values lead to consistent improvements in performance. Thus, we recommend being conservative and only using a single sample to define the local density, i.e., $K = 1$ or $r = 0$ without additional prior knowledge. Note that these numbers are different from the results presented in [30], as the results presented here are based on the official performance metric of the DCASE Challenge, while a simplified performance measure was used in [30].

C. Comparison to the state of the art

In Table IV, the proposed anomaly score normalization approach is evaluated on a representative group of ASD datasets, namely the DCASE2020, DCASE2023, and DCASE2024 datasets, and the resulting performance is compared to the state of the art. For this comparison, we consider the five systems direct-ACT, OpenL3-raw, BEATs-raw, EAT-raw, and direct-ACT (ensemble) as described in Section IV-C, all with a local-density-based normalization of the anomaly scores using k-NN with $K = 1$. While the direct-ACT model with the proposed normalization stays slightly behind the state-of-the-art performance, the ensemble robustly outperforms the state-of-the-art systems on the DCASE2020 and DCASE2023 datasets. On the DCASE2024 dataset, our performance is better on the development set but still worse than the state of the art and comparable to the baseline system [77] on the evaluation set. The main reason for this poorer performance is that some machine types contained in the DCASE2024 dataset have specific noise conditions. For these machine types, an ACT-based system only needs to monitor the noise and thus completely fails in detecting subtle changes in target machine sounds. Another reason is that, for some machine types, no attribute information is provided, which further degrades the performance. Both imposed difficulties need to be addressed to be able to achieve state-of-the-art performance. Additional evidence for this claim is that the performance obtained with the simple BEATs-raw model is only slightly worse than the direct-ACT ensemble on the DCASE2024 dataset while being much worse on the DCASE2020 and DCASE2023 datasets. Note that BEATs and AnoPatch make extensive use of SSL via masking patches while the direct-ACT model only uses SSL via feature exchange, similarly to OpenL3. The difference in performance can thus be seen as evidence for the importance of using SSL to learn suitable representations for ASD.

VI. CONCLUSIONS AND FUTURE WORK

In this work, a simple yet highly effective anomaly score normalization approach for DG was presented. The approach was extensively evaluated on the DCASE2020, DCASE2022,

DCASE2023, DCASE2024, and DCASE2025 ASD datasets using state-of-the-art embedding models, from a model based on directly training with the discriminative angular margin loss AdaProj, to models based on pre-trained embeddings, such as OpenL3, BEATs, and EAT. The proposed normalization approach was shown to consistently and significantly improve the performance in domain-shifted conditions and outperforms other existing anomaly score calculation and normalization approaches while not using any domain labels or adapting to specific target domains, and treating each test sample independently. As a result, an ensemble-based ASD system was presented that utilizes this normalization approach and achieves state-of-the-art performance on the DCASE2020 and DCASE2023 dataset. For future work, we plan to evaluate the proposed normalization approach for other applications and different data modalities in addition to audio. Furthermore, we plan to explore synergies with other normalization approaches, as also done in [74]. Last but not least, the proposed ASD system will be adapted to provide state-of-the-art performance on the DCASE2024 dataset by fine-tuning BEATs with an ACT, similarly to AnoPatch [70].

REFERENCES

[1] C. C. Aggarwal, *Outlier Analysis*. Springer, 2013.

[2] Y. Koizumi *et al.*, "Description and discussion on DCASE2020 Challenge Task2: Unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2020, pp. 81–85.

[3] Y. Kawaguchi *et al.*, "Description and discussion on DCASE 2021 Challenge Task 2: Unsupervised anomalous detection for machine condition monitoring under domain shifted conditions," in *Proc. DCASE*, 2021, pp. 186–190.

[4] K. Dohi *et al.*, "Description and discussion on DCASE 2022 Challenge Task 2: Unsupervised anomalous sound detection for machine condition monitoring applying domain generalization techniques," in *Proc. DCASE*, 2022.

[5] ——, "Description and discussion on DCASE 2023 Challenge Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2023, pp. 31–35.

[6] T. Nishida *et al.*, "Description and discussion on DCASE 2024 Challenge Task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2024, pp. 111–115.

[7] ——, "Description and discussion on DCASE 2025 challenge task 2: First-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2025.

[8] J. Wang, C. Lan, C. Liu, Y. Ouyang, and T. Qin, "Generalizing to unseen domains: A survey on domain generalization," in *Proc. IJCAI*, 2021, pp. 4627–4635.

[9] J. Wang *et al.*, "Generalizing to unseen domains: A survey on domain generalization," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 8, pp. 8052–8072, 2023.

[10] K. Wilkinghoff, T. Fujimura, K. Imoto, J. Le Roux, Z.-H. Tan, and T. Toda, "Handling domain shifts for anomalous sound detection: A review of DCASE-related work," in *Proc. DCASE*, 2025.

[11] K. Wilkinghoff, "Audio embeddings for semi-supervised anomalous sound detection," Ph.D. dissertation, University of Bonn, 2024.

[12] R. Giri, S. V. Tenneti, F. Cheng, K. Helwani, U. Isik, and A. Krishnaswamy, "Self-supervised classification for detecting anomalous sounds," in *Proc. DCASE*, 2020, pp. 46–50.

[13] P. Primus, V. Haunschmid, P. Praher, and G. Widmer, "Anomalous sound detection as a simple binary classification problem with careful selection of proxy outlier examples," in *Proc. DCASE*, 2020, pp. 170–174.

[14] K. Wilkinghoff, "Sub-cluster AdaCos: Learning representations for anomalous sound detection," in *Proc. IJCNN*, 2021.

[15] ——, "Combining multiple distributions based on sub-cluster adacos for anomalous sound detection under domain shifted conditions," in *Proc. DCASE*, 2021, pp. 55–59.

[16] S. Venkatesh, G. Wichern, A. S. Subramanian, and J. Le Roux, "Improved domain generalization via disentangled multi-task learning in unsupervised anomalous sound detection," in *Proc. DCASE*, 2022.

[17] F. G. Germain, G. Wichern, and J. Le Roux, "Hyperbolic unsupervised anomalous sound detection," in *Proc. WASPAA*, 2023, pp. 1–5.

[18] T. Inoue *et al.*, "Detection of anomalous sounds for machine condition monitoring using classification confidence," in *Proc. DCASE*, 2020, pp. 66–70.

[19] J. A. Lopez, H. Lu, P. Lopez-Meyer, L. Nachman, G. Stemmer, and J. Huang, "A speaker recognition approach to anomaly detection," in *Proc. DCASE*, 2020, pp. 96–99.

[20] I. Nejjar, J. Meunier-Pion, G. Frusque, and O. Fink, "DG-Mix: Domain generalization for anomalous sound detection based on self-supervised learning," in *Proc. DCASE*, 2022.

[21] H. Chen *et al.*, "An effective anomalous sound detection method based on representation learning with simulated anomalies," in *Proc. ICASSP*, 2023.

[22] K. Wilkinghoff, "Self-supervised learning for anomalous sound detection," in *Proc. ICASSP*, 2024, pp. 276–280.

[23] D. Hendrycks, M. Mazeika, and T. G. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. ICLR*, 2019.

[24] 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 Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 212–224, 2019.

[25] K. Suefusa, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, "Anomalous sound detection based on interpolation deep neural network," in *Proc. ICASSP*, 2020, pp. 271–275.

[26] R. Giri, F. Cheng, K. Helwani, S. V. Tenneti, U. Isik, and A. Krishnaswamy, "Group masked autoencoder based density estimator for audio anomaly detection," in *Proc. DCASE*, 2020, pp. 51–55.

[27] S. Kapka, "ID-conditioned auto-encoder for unsupervised anomaly detection," in *Proc. DCASE*, 2020, pp. 71–75.

[28] G. Wichern, A. Chakrabarty, Z. Wang, and J. Le Roux, "Anomalous sound detection using attentive neural processes," in *Proc. WASPAA*, 2021, pp. 186–190.

[29] K. Wilkinghoff and F. Kurth, "Why do angular margin losses work well for semi-supervised anomalous sound detection?" *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 608–622, 2024.

[30] K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Keeping the balance: Anomaly score calculation for domain generalization," in *Proc. ICASSP*, 2025.

[31] E. Schubert, A. Zimek, and H. Kriegel, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data Min. Knowl. Discov.*, vol. 28, no. 1, pp. 190–237, 2014.

[32] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," in *Proc. PKDD*, 2002.

[33] Y. Deng *et al.*, "Ensemble of multiple anomalous sound detectors," in *Proc. DCASE*, 2022.

[34] H. Purohit, R. Tanabe, T. Endo, K. Suefusa, Y. Nikaido, and Y. Kawaguchi, "Deep autoencoding GMM-based unsupervised anomaly detection in acoustic signals and its hyper-parameter optimization," in *Proc. DCASE*, 2020, pp. 175–179.

[35] I. Kuroyanagi, T. Hayashi, K. Takeda, and T. Toda, "Improvement of serial approach to anomalous sound detection by incorporating two binary cross-entropies for outlier exposure," in *Proc. EUSIPCO*, 2022, pp. 294–298.

[36] K. Wilkinghoff, "Design choices for learning embeddings from auxiliary tasks for domain generalization in anomalous sound detection," in *Proc. ICASSP*, 2023.

[37] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.

[38] X. Zhang, R. Zhao, Y. Qiao, X. Wang, and H. Li, "AdaCos: Adaptively scaling cosine logits for effectively learning deep face representations," in *Proc. CVPR*, 2019, pp. 10 823–10 832.

[39] K. Wilkinghoff, "AdaProj: Adaptively scaled angular margin subspace projections for anomalous sound detection with auxiliary classification tasks," in *Proc. DCASE*, 2024, pp. 186–190.

[40] M. M. Breunig, H. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," in *Proc. SIGMOD*, 2000, pp. 93–104.

[41] S. Papadimitriou, H. Kitagawa, P. B. Gibbons, and C. Faloutsos, "LOCI: Fast outlier detection using the local correlation integral," in *Proc. ICDE*, 2003, pp. 315–326.

[42] J. Tang, Z. Chen, A. W. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Proc. PAKDD*, 2002.

[43] K. Zhang, M. Hutter, and H. Jin, "A new local distance-based outlier detection approach for scattered real-world data," in *Proc. PAKDD*, 2009.

[44] H. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "LoOP: local outlier probabilities," in *Proc. CIKM*, 2009, pp. 1649–1652.

[45] ——, "Interpreting and unifying outlier scores," in *Proc. SDM*, 2011, pp. 13–24.

[46] E. Pizzi, S. D. Roy, S. N. Ravindra, P. Goyal, and M. Douze, "A self-supervised descriptor for image copy detection," in *Proc. CVPR*, 2022, pp. 14 512–14 522.

[47] D. Bralios *et al.*, "Generation or replication: Auscultating audio latent diffusion models," in *Proc. ICASSP*, 2024, pp. 1156–1160.

[48] H. Jégou, M. Douze, and C. Schmid, "Exploiting descriptor distances for precise image search," INRIA, Tech. Rep., 2011.

[49] S. Cumani, P. D. Batzu, D. Colibro, C. Vair, P. Laface, and V. Vasilakakis, "Comparison of speaker recognition approaches for real applications," in *Proc. Interspeech*, 2011, pp. 2365–2368.

[50] Z. N. Karam, W. M. Campbell, and N. Dehak, "Towards reduced false-alarms using cohorts," in *Proc. ICASSP*, 2011, pp. 4512–4515.

[51] P. Saengthong and T. Shinozaki, "Deep generic representations for domain-generalized anomalous sound detection," in *Proc. ICASSP*, 2025.

[52] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in *Proc. ICLR*, 2018.

[53] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Proc. ECCV*, 2016.

[54] J. Guan, Y. Liu, Q. Zhu, T. Zheng, J. Han, and W. Wang, "Time-weighted frequency domain audio representation with GMM estimator for anomalous sound detection," in *Proc. ICASSP*, 2023.

[55] H. Purohit *et al.*, "MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection," in *Proc. DCASE*, 2019, pp. 209–213.

[56] Y. Koizumi, S. Saito, H. Uematsu, N. Harada, and K. Imoto, "ToyADMOS: A dataset of miniature-machine operating sounds for anomalous sound detection," in *Proc. WASPAA*, 2019, pp. 313–317.

[57] K. Dohi *et al.*, "MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task," in *Proc. DCASE*, 2022, pp. 31–35.

[58] 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," in *Proc. DCASE*, 2021.

[59] N. Harada, D. Niizumi, D. Takeuchi, Y. Ohishi, and M. Yasuda, "ToyADMOS2+: New Toyadmos data and benchmark results of the first-shot anomalous sound event detection baseline," in *Proc. DCASE*, 2023.

[60] D. Niizumi, N. Harada, Y. Ohishi, D. Takeuchi, and M. Yasuda, "ToyADMOS2#: Yet another dataset for the DCASE2024 challenge task 2 first-shot anomalous sound detection," in *Proc. DCASE*, 2024.

[61] D. Albertini, F. Augusti, K. Esmer, A. Bernardini, and R. Sannino, "IMAD-DS: A dataset for industrial multi-sensor anomaly detection under domain shift conditions," in *Proc. DCASE*, 2024.

[62] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "ToyADMOS2025: The evaluation dataset for the DCASE2025T2 first-shot unsupervised anomalous sound detection for machine condition monitoring," in *Proc. DCASE*, 2025, pp. 230–234.

[63] D. K. McClish, "Analyzing a portion of the ROC curve," *Medical decision making*, vol. 9, no. 3, pp. 190–195, 1989.

[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[65] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. ICLR*, 2018.

[66] A. Cramer, H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *Proc. ICASSP*, 2019, pp. 3852–3856.

[67] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proc. ICCV*, 2017, pp. 609–617.

[68] ——, "Objects that sound," in *Proc. ECCV*, 2018.

[69] S. Chen *et al.*, "BEATs: Audio pre-training with acoustic tokenizers," in *Proc. ICML*, 2023.

[70] A. Jiang *et al.*, "AnoPatch: Towards better consistency in machine anomalous sound detection," in *Proc. Interspeech*, 2024, pp. 107–111.

[71] J. F. Gemmeke *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. ICASSP*, 2017, pp. 776–780.

[72] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: self-supervised pre-training with efficient audio transformer," in *Proc. IJCAI*, 2024, pp. 3807–3815.

[73] T. Fujimura, K. Wilkinghoff, K. Imoto, and T. Toda, "ASDKit: A toolkit for comprehensive evaluation of anomalous sound detection methods," in *Proc. DCASE*, 2025.

[74] P. Saengthong and T. Shinozaki, "GENREP for first-shot unsupervised anomalous sound detection of DCASE 2025 challenge," DCASE2025 Challenge, Tech. Rep., 2025.

[75] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[76] Y. Liu, J. Guan, Q. Zhu, and W. Wang, "Anomalous sound detection using spectral-temporal information fusion," in *Proc. ICASSP*, 2022, pp. 816–820.

[77] N. Harada, D. Niizumi, Y. Ohishi, D. Takeuchi, and M. Yasuda, "First-shot anomaly sound detection for machine condition monitoring: A domain generalization baseline," in *Proc. EUSIPCO*, 2023.

[78] Q. Hou, A. Jiang, W. Zhang, P. Fan, and J. Liu, "Decoupling detectors for scalable anomaly detection in AIoT systems with multiple machines," in *Proc. GLOBECOM*, 2023, pp. 5937–5942.

[79] B. Han *et al.*, "Exploring large scale pre-trained models for robust machine anomalous sound detection," in *Proc. ICASSP*, 2024, pp. 1326–1330.

[80] Y. Zhang, J. Liu, Y. Tian, H. Liu, and M. Li, "A dual-path framework with frequency-and-time excited network for anomalous sound detection," in *Proc. ICASSP*, 2024, pp. 1266–1270.

[81] J. Yin, W. Zhang, M. Zhang, and Y. Gao, "Self-supervised augmented diffusion model for anomalous sound detection," in *Proc. APSIPA*, 2024.

[82] T. Fujimura, I. Kuroyanagi, and T. Toda, "Improvements of discriminative feature space training for anomalous sound detection in unlabeled conditions," in *Proc. ICASSP*, 2025.

[83] J. Yin, Y. Gao, W. Zhang, T. Wang, and M. Zhang, "Diffusion augmentation sub-center modeling for unsupervised anomalous sound detection with partially attribute-unavailable conditions," in *Proc. ICASSP*, 2025.

[84] A. Jiang *et al.*, "Adaptive prototype learning for anomalous sound detection with partially known attributes," in *Proc. ICASSP*, 2025.