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

Keeping the Balance: Anomaly Score Calculation for Domain Generalization

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

TR2025-030 March 08, 2025

Abstract

Emitted sounds may drastically change when using different microphones, when properties of the sound sources change, or when recording in different acoustic environments. Ideally, anomalous sound detection (ASD) systems should be able to generalize well to unseen target domains by only providing a few target domain samples to define how normal data samples sound like, without needing to re-train or modify the system. In contrast with the source domain, for which many normal training samples are available, accurately estimating the underlying distribution of normal data after a domain shift based on very few samples is challenging. This usually leads to a mismatch between the corresponding anomaly scores of source and target domains and significantly reduces performance. In this work, we propose a framework for re-scaling anomaly scores based on the ratio between the cosine distance of a test sample to a normal reference sample and the distances to this sample's next-closest neighbors in the reference set. In experimental evaluations, it is shown that the re-scaled anomaly scores reduce the domain mismatch for multiple domains. As a result, we obtain new state-of-the-art performances on the DCASE2020 and DCASE2023 ASD datasets

IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2025

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

Keeping the Balance: Anomaly Score Calculation for Domain Generalization

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

Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

***Abstract*—Emitted sounds may drastically change when using different microphones, when properties of the sound sources change, or when recording in different acoustic environments. Ideally, anomalous sound detection (ASD) systems should be able to generalize well to unseen target domains by only providing a few target domain samples to define how normal data samples sound like, without needing to re-train or modify the system. In contrast with the source domain, for which many normal training samples are available, accurately estimating the underlying distribution of normal data after a domain shift based on very few samples is challenging. This usually leads to a mismatch between the corresponding anomaly scores of source and target domains and significantly reduces performance. In this work, we propose a framework for re-scaling anomaly scores based on the ratio between the cosine distance of a test sample to a normal reference sample and the distances to this sample's next-closest neighbors in the reference set. In experimental evaluations, it is shown that the re-scaled anomaly scores reduce the domain mismatch for multiple domains. As a result, we obtain new state-of-the-art performances on the DCASE2020 and DCASE2023 ASD datasets.**

***Index Terms*—anomalous sound detection, domain generalization, machine condition monitoring**

I. Introduction

The goal of semi-supervised anomalous sound detection (ASD) is to detect anomalous sounds by using normal training data only, as anomalous data is rare and thus costly to obtain. The main difficulties of the task as set up in the DCASE challenge for acoustic machine condition monitoring [1]–[5] are that the target sounds are not isolated but occur within complex acoustic scenes that may be very noisy, and anomalies are usually just subtle differences away from normal sounds. ASD systems need to be sensitive to the anomalies but insensitive to unrelated sound events and noise. Furthermore, there is no inherent property highlighting anomalous data as anomalies are entirely application-dependent: sound events that are normal for one application may be anomalous for another and vice versa. On top of that, raw audio data is very high-dimensional, making a direct comparison between different recordings difficult. To overcome all these issues, the main idea is to learn to project audio data into an embedding space where normal data is clustered and anomalies substantially differ from the normal data [6]. The state-of-the-art approach to train such an embedding model is to use an auxiliary classification task based on meta information or self-supervised learning (SSL). This helps to be less sensitive to the noise by closely monitoring the target sounds in an acoustic scene [7]. Ideally, the system should be able to generalize well to unseen domain shifts of the data caused by modifying any properties of the target sounds, the acoustic environment, or the recording equipment. In practice, users of ASD systems do not want to re-train or tune their systems but only provide a few recordings in the new conditions to show the system how normal data should sound like.

After projecting the data into an embedding space, one can measure the distance between embeddings belonging to different recordings or estimate the distributions of normal training data to distinguish between normal and anomalous sounds. For individual domains, the performance improves with more accurate estimates of the distributions of the embeddings [8], e.g., by using a Gaussian mixture model (GMM) [9]–[11] or the Mahalanobis distance [12]. However, for target domains with only a few training samples, it is impossible to estimate the underlying distribution, leading to very different distributions of the anomaly scores of the source and target domains (domain mismatch), which degrades performance when using a single decision threshold. In this multi-domain case, using a simple distance-based nearest-neighbor approach leads to anomaly scores that are more similar and thus also to better performance than when trying to estimate the distribution [13].

In this work, we introduce a novel and highly effective re-scaling approach of cosine-distance-based anomaly scores; in multiple experiments, we investigate different design choices as well as the robustness of hyperparameter settings, and show that the proposed approach significantly and consistently improves resulting ASD performance by reducing the domain mismatch of anomaly scores; based on this approach, we obtain a new state-of-the-art performance on the DCASE2020 and DCASE2023 ASD datasets.

II. Anomaly score calculation

The goal of anomaly detection systems is to map a sample x to an anomaly score $\mathcal{A}(x) \in \mathbb{R}$ such that normal samples have a low anomaly score and anomalous samples have a high score. Then, a decision threshold can be applied to detect anomalous samples. In this section, we will first discuss multiple methods for computing an anomaly score, then move on to the description of the proposed approach. As most state-of-the-art ASD systems first map the audio recordings (or derived features) into an embedding space by using a neural network, we will equate the sample x with its embedding in the embedding space $\mathcal{X}$.

A. Baseline approach

When trying to detect anomalous samples among a set of test samples $\mathcal{X}_{\text{test}}$ in an embedding space, one of the simplest approaches is to just compare the distance of each test sample to a set of normal reference samples $\mathcal{X}_{\text{ref}}$, such as the training set $\mathcal{X}_{\text{train}}$. Here, the assumption is that normal test samples will have a much smaller distance than anomalous samples to the closest normal training sample. Angular margin losses tuned through training on auxiliary classification tasks are known to lead to good ASD performance. In this case, the embedding space is on the unit sphere and thus a distance between two samples can be calculated using the cosine distance. Then, for example, the anomaly score of test sample $x \in \mathcal{X}_{\text{test}}$ can be defined as the distance to its nearest neighbor in a set $\mathcal{X}_{\text{ref}}$ as

$$\begin{aligned}\mathcal{A}^{\text{NN}}_{\cos}(x, \mathcal{X}_{\text{ref}}) &:= \min_{y \in \mathcal{X}_{\text{ref}}} \mathcal{A}_{\cos}(x, y) \\ &:= \min_{y \in \mathcal{X}_{\text{ref}}} \frac{1}{2}(1 - \langle x, y \rangle) \in [0, 1],\end{aligned}$$

where $\|x\|_2 = \|y\|_2 = 1$ for elements x, y on the unit sphere. One can also easily apply this to multiple domains by just considering the distance to all samples in the combined source and target domains.

In the case of multiple data domains, this simple cosine distance based approach is known to lead to better performance than more sophisticated approaches based on estimating the distribution of the normal embeddings, such as GMMs [13]. The reason for this performance degradation is that, while it is certainly possible to improve the performance on the source domain [10], it is impossible to accurately estimate the distribution in the target domain when only a few target-domain samples are available and the dimension of the embedding space has a magnitude of several hundreds. In contrast, the simple nearest neighbor approach does not require any training or distribution estimation and thus works equally well (or bad) in both domains. Possible extensions of this approach are to not only measure the distance to a single nearest neighbor but to multiple closest ones (i.e., k-nearest neighbors (k-NN)) [12], [14] or to apply k-means to the source domain samples first and then to measure the distance to these means and the original target domain samples [13]. Both approaches lead to slight improvements in performance as they are more robust to per-sample noise.

B. Proposed re-scaling approach

Although the previously discussed baseline approach works reasonably well to detect anomalies, it has one major problem: namely, the approach still assumes that the distances to the normal training samples behave similarly for the source and target domains. However, one would assume that the target domain samples have less dense distributions, i.e., are more scattered, and thus have higher distances between themselves than source domain samples, particularly as the target domain samples are not used for training the embedding models. Moreover, the distributions of normal samples can look very irregular in high dimensions with several differently scaled overlapping clusters corresponding to sub-classes of the normal data spaces, even for the source domain alone. This leads to a strong mismatch between the distributions of the anomaly scores belonging to different domains or sub-classes, and trying to detect anomalous samples with a single decision threshold is highly sub-optimal. Our goal here is to develop an approach for re-scaling the anomaly scores that reduces the mismatch between the scores by being less dependent on the absolute distances to specific reference samples. For speaker verification, score re-scaling approaches are known as score calibration [15], [16] and also help to improve the performance.

We propose two such approaches, the first one based on k-NN and the second on global weighted ranking pooling (GWRP) [17] as also used in [18]. Both approaches rely on the idea of comparing the distance between the test sample and a reference sample with the distances between that reference sample and its closest neighbors. For a given $x \in \mathcal{X}_{\text{test}}$, we denote by $y \in X_{\text{ref}}$ a reference sample and by y_k the k-th closest sample to that reference sample in $\mathcal{X}_{\text{ref}}$, for $k = 1, \dots, K$. Our proposed re-scaled anomaly scores are then defined as

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

$$\mathcal{A}^{\text{GWRP}}_{\text{scaled}}(x, \mathcal{X}_{\text{ref}} \,|\, r) := \min_{y \in \mathcal{X}_{\text{ref}}} \frac{\mathcal{A}_{\cos}(x, y)}{\sum_{k=1}^{|\mathcal{X}_{\text{ref}}|-1} \mathcal{A}_{\cos}(y, y_k) \cdot r^{k-1}} \in \mathbb{R}_+, \tag{2}$$

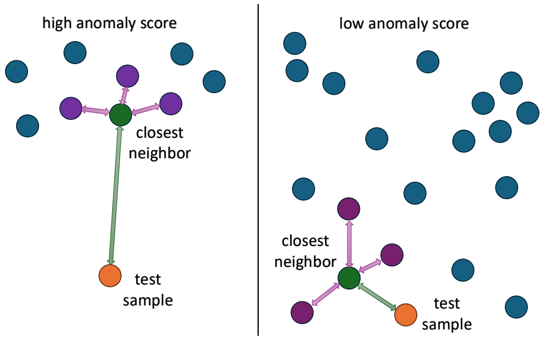

Fig. 1. Illustration of the proposed anomaly score re-scaling approach $\mathcal{A}^{\text{k-NN}}_{\text{scaled}}$ for $K=3$. On the left, the distance to the closest neighbor and the distances to this neighbor's closest neighbors are very different, leading to a high re-scaled anomaly score. On the right, all distances are similar, leading to a small re-scaled anomaly score.

where $K \in \mathbb{N}^+$ denotes the number of next closest samples to consider and $r \in [0, 1]$ the weight factor, both acting as hyperparameter.

As illustrated in Fig. 1, the intuitive idea of the proposed re-scaling is to balance the distances to samples belonging to clusters with different densities, which are likely to correspond to different domains or sub-classes. The distance between a test sample and a given reference sample is normalized based on the local density of that reference sample, expressed as the mean distance to its local neighborhood. The reference sample with lowest renormalized distance to the test sample is used to defined the anomaly score. Thus, if two reference samples are similarly close, we favor the sample that is more isolated, i.e., the one that probably belongs to the target domain. Essentially, high-density samples are interpreted as further apart from other samples, while low-density ones are interpreted as closer, compared to what they would be according to the unnormalized distances. It shall be noted that the normalization factors for each reference sample can be pre-computed as they only depend on the reference set and thus there are no losses in efficiency during inference.

k-NN and GWRP start from the same points (when $K = 1$ and $r = 0$), and work from different directions. k-NN directly picks a limited number of neighbors to consider, while giving them equal weights. GWRP, on the contrary, considers every data point in the domain, but emphasizes closer ones by imposing exponentially decreasing weights. In the end, they intersect again when K in k-NN equals to the number of data points ($k = N$), and $r = 1$.

It shall be emphasized that the proposed approach has several advantages. First, the anomaly scores are not directly related to the absolute distances between different samples as only relations between distances are considered. Thus, the domain mismatch between the distributions of anomaly scores is significantly reduced. Furthermore, no training or estimation of a distribution, which is difficult in high-dimensional spaces, is needed to compute an anomaly score because the scores effectively only depend on the distance to the local neighborhood of the closest reference sample of a test sample.

C. Relation to a local outlier factor-based approach

In this section, we discuss local outlier factor (LOF) [21] in more detail as it is in some ways related to our proposed approach. LOF detects anomaly outliers by assessing how isolated a data point is with respect to a pre-determined number K of nearest neighbors. It calculates a notion of local density for the test sample x and each of its K nearest neighbors. The outlier factor that is used as an anomaly score is defined as the ratio of the local density of x to those of x's

TABLE I
HARMONIC MEANS OF ALL AUCS AND PAUCS OBTAINED WITH DIFFERENT RE-SCALING APPROACHES OF THE ANOMALY SCORES. MEAN AND STANDARD DEVIATION OVER TEN INDEPENDENT TRIALS CORRESPONDING TO TEN TRAINED EMBEDDING MODELS ARE SHOWN. TO ALLOW FOR BETTER COMPARISON, THE SAME TEN TRAINED EMBEDDING MODELS ARE USED FOR ALL EVALUATIONS. HIGHEST NUMBERS IN EACH COLUMN ARE IN BOLD.

			DCASE2023 development set [4], [19], [20]				DCASE2023 evaluation set [4], [19], [20]			
			single model			ensemble	single model			ensemble
re-scaling	parameter	reference samples $\mathcal{X}_{\text{ref}}$	source domain	target domain	both domains	both domains	source domain	target domain	both domains	both domains
-	-	source means and target samples	68.6 ± 1.2%	66.3 ± 1.0%	67.4 ± 0.6%	68.7%	68.1 ± 1.4%	62.7 ± 1.0%	65.2 ± 0.6%	67.7%
-	-	all samples	68.6 ± 1.1%	64.9 ± 1.0%	66.7 ± 0.8%	67.9%	72.3 ± 1.6%	59.0 ± 0.9%	65.0 ± 0.8%	66.8%
LOF	$K = 8$	all samples	65.4 ± 0.9%	58.4 ± 1.6%	61.7 ± 1.1%	63.6%	73.7 ± 1.2%	55.1 ± 1.9%	63.0 ± 1.4%	64.1%
LOF	$K = 16$	all samples	65.1 ± 0.9%	58.3 ± 1.7%	61.5 ± 1.2%	62.8%	71.9 ± 1.5%	53.3 ± 2.2%	61.2 ± 1.4%	63.2%
k-NN	$K = 8$	source means and target samples	67.5 ± 1.3%	68.5 ± 1.3%	68.0 ± 0.7%	69.6%	65.3 ± 2.0%	62.0 ± 1.5%	63.6 ± 1.3%	66.6%
k-NN	$K = 16$	source means and target samples	67.6 ± 1.1%	68.5 ± 1.4%	68.0 ± 0.7%	69.6%	66.5 ± 1.1%	**63.4 ± 1.1%**	64.9 ± 0.8%	67.5%
k-NN	$K = 8$	all samples	**68.7 ± 1.2%**	68.2 ± 1.0%	**68.4 ± 0.8%**	**72.0%**	**73.9 ± 1.7%**	62.1 ± 1.6%	67.4 ± 1.3%	71.2%
k-NN	$K = 16$	all samples	68.2 ± 1.5%	68.5 ± 0.8%	68.3 ± 0.8%	71.7%	73.8 ± 1.6%	62.7 ± 1.4%	**67.8 ± 1.2%**	71.2%
GWRP	$r = 0.85$	source means and target samples	68.1 ± 1.2%	68.3 ± 1.6%	68.2 ± 0.8%	70.0%	67.0 ± 1.4%	62.8 ± 1.1%	64.8 ± 0.9%	67.3%
GWRP	$r = 0.90$	source means and target samples	67.6 ± 1.3%	68.3 ± 1.5%	67.9 ± 0.8%	69.7%	67.0 ± 1.0%	63.2 ± 1.0%	65.0 ± 0.7%	67.4%
GWRP	$r = 0.95$	source means and target samples	67.7 ± 1.4%	67.9 ± 1.6%	67.8 ± 0.8%	69.5%	67.1 ± 1.3%	63.1 ± 0.8%	65.0 ± 0.5%	67.4%
GWRP	$r = 0.85$	all samples	68.5 ± 1.2%	68.3 ± 1.0%	**68.4 ± 0.8%**	71.8%	**73.9 ± 1.7%**	62.4 ± 1.4%	67.7 ± 1.2%	**71.3%**
GWRP	$r = 0.90$	all samples	68.2 ± 1.4%	68.5 ± 0.9%	68.3 ± 0.8%	71.6%	73.8 ± 1.7%	62.6 ± 1.5%	**67.8 ± 1.2%**	71.2%
GWRP	$r = 0.95$	all samples	68.1 ± 1.6%	**68.6 ± 0.9%**	68.3 ± 0.8%	71.5%	73.5 ± 1.5%	62.9 ± 1.5%	**67.8 ± 1.3%**	71.1%

K neighbors. Thus, there is a clear connection between our method and LOF. Both methods' decision boundaries take into account the distance with target points' neighbors, and do not require any implicit or explicit clustering. However, LOF computation relies on identifying the closest reference neighbors to the target sample before density estimation. Hence, closer (in unnormalized distance) but high-density neighbors would end up included in the computation. Conversely, our proposed method identifies the closest reference sample after normalization. Hence, a far-away (in unnormalized distance) but low-density reference could end up being selected for anomaly scoring. Additionally, we only consider a single reference sample while LOF averages over many, risking the undesirable scenario where samples from both source and target domains are combined. Furthermore, our approach does not involve computing the local density of a given test sample, which for the case where the test sample is anomalous may not be well defined. Finally, in previous works on ASD, LOF was used to compute anomaly scores [11], [12], but mostly in large ensembles or as a cherry-picked computation method for individual machine types, without any noticeable gains in overall performance.

III. EXPERIMENTAL EVALUATIONS

A. ASD system design

For all experimental evaluations in this work, we used a modified version of the state-of-the-art ASD system presented in [22] to map the audio signals into an embedding space. The system consists of two feature branches based on the magnitude of the entire spectrum, i.e., the discrete Fourier transform of the audio signal, and the magnitude short-time Fourier transform (STFT) with a Hann window of size 1024 and a step size of 512 samples. The STFT features are further processed by subtracting the temporal mean to remove stationary frequency information and make both feature branches more different while also denoising the data. Next, a different convolutional neural network is applied to each feature branch to map the features into a 256-dimensional embedding space. Both embeddings are concatenated to obtain a single embedding. The model is trained for 5 epochs with a batch size of 64 using Adam [23]. The training objective is an auxiliary classification task based on identifying meta information such as machine types, machine IDs, and additional parameter settings. Here, each combination of provided values is considered as a separate class. Furthermore, feature exchange [22], an SSL approach that randomly exchanges the embeddings of both feature branches between two training samples and asks the model to distinguish between original and modified embeddings, is used as an additional training objective. Note that a single model is used for the entire dataset and only normal data belonging to the source domain is used for training the embedding model. The ensembles used for the experimental evaluations in this work are obtained by averaging the anomaly scores of ten individual models. More details about the system can be found in [13] and [22].

In contrast to the original version of the system, no statistics exchange [24] was used to train the model as we noticed only marginal or even no performance gains. Furthermore, the sub-cluster AdaCos loss [10] was replaced with the AdaProj loss [25] to slightly improve the performance. The AdaProj loss is an angular margin loss that learns entire class-specific linear sub-spaces on the unit sphere instead of trying to project a sample as close as possible to its corresponding class centers.

B. Experimental results

1) Comparison of re-scaling approaches: First, we compared different design choices for re-scaling the anomaly scores. The results can be found in Table I. We also verified that using LOF leads to worse performance, and that applying k-means in the source domain to obtain reference samples, as proposed in [13], improves the performance when not re-scaling the scores. Interestingly, when re-scaling the scores, the opposite is true: using means as reference samples leads to much worse results than directly using the training samples. For the evaluation set, the performance slightly degrades when using the means. A possible explanation is that individual samples are not close enough to individual means and thus measuring the distance is less meaningful. Furthermore, different means may also represent different sub-classes of the source domain data corresponding to minor domain shifts within the source domain, resulting in anomaly score distributions that are not necessarily well-aligned. This is also supported by the fact that most of the performance gains for the evaluation set are obtained on the source domain. We also experimented with domain-dependent score standardization approaches similar to the one presented in [33], but in our experiments this did not lead to any improvements in performance.

2) Effect of varying the constant r of GWRP: Figure 2 shows the influence of the value of the constant r on the GWRP performance. In general, the model favors higher values of r, and the advantage of higher r (> 0.6) is more obvious on the evaluation set than on the development set. Particularly, the performance curve reaches its peak within the range of 0.85 to 0.95, thus we propose to make $r = 0.9$ the default parameter setting. When all the data points have equal contribution, the performance is prone to bias caused by distant and/or outlier points possibly belonging to very different clusters. Reflected

TABLE II
HARMONIC MEANS OF ALL AUCS AND PAUCS OBTAINED WITH DIFFERENT ASD SYSTEMS. WHENEVER APPLICABLE, MEANS OF ALL INDEPENDENT TRIALS ARE SHOWN. HIGHEST NUMBERS IN EACH COLUMN ARE HIGHLIGHTED WITH BOLD LETTERS.

		DCASE2020 dataset [1], [26], [27]			DCASE2023 dataset [4], [19], [20]			DCASE2024 dataset [5], [19], [20]		
ASD system	number of trials	dev. set	eval. set	arithm. mean	dev. set	eval. set	harm. mean	dev. set	eval. set	harm. mean
baseline (single model)	10	92.1%	91.5%	91.8%	67.4%	65.2%	66.3%	60.5%	55.1%	57.7%
baseline (ensemble)	1	93.3%	92.4%	92.9%	68.7%	67.7%	68.2%	61.6%	56.2%	58.8%
proposed approach (single model)	10	90.6%	89.9%	90.3%	68.3%	67.8%	68.0%	60.6%	56.6%	58.5%
proposed approach (ensemble)	1	**93.6%**	92.7%	**93.2%**	**71.7%**	71.2%	**71.4%**	**63.4%**	**58.6%**	**60.9%**
Wilkinghoff [10] (single model)	1	90.7%	92.8%	91.8%	–	–	–	–	–	–
Wilkinghoff [10] (ensemble)	1	–	94.1%	–	–	–	–	–	–	–
Liu et al. [28]	1	89.4%	–	–	–	–	–	–	–	–
Wilkinghoff [13]	5	–	–	–	62.8%	63.0%	62.9%	–	–	–
Hou et al. [29]	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. [30]	1	–	–	–	64.3%	–	–	–	–	–
Zhang et al. [31]	1	–	–	–	–	71.3%	–	–	–	–
Jiang et al. [32]	1	90.9%	**94.3%**	92.6%	64.2%	**74.2%**	68.8%	–	–	–
Wilkinghoff [25]	10	–	–	–	62.9%	64.5%	63.7%	–	–	–
Saengthong et al. [33]	1	74.7%	–	–	–	73.8%	–	–	–	–

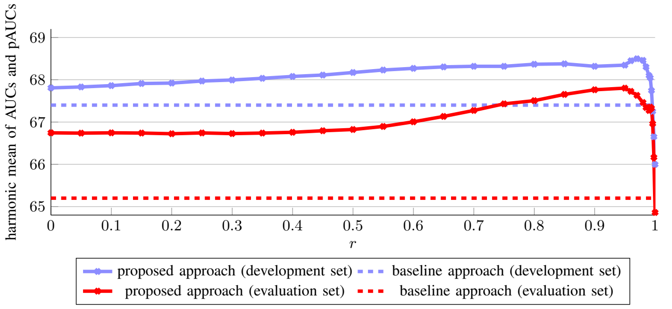

Fig. 2. Effect of varying the GWRP constant r on the performance when re-scaling the anomaly scores with the proposed approach. Mean results over ten independent trials are shown.

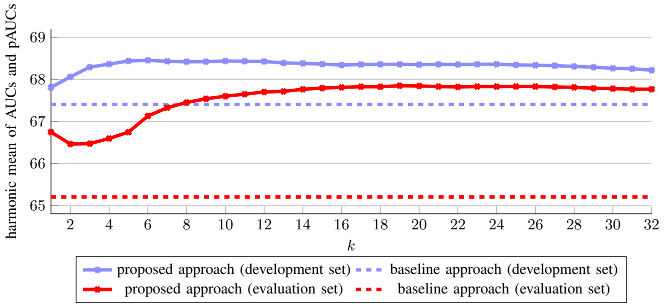

Fig. 3. Effect of varying the number K for k-NN on the performance when re-scaling the anomaly scores with the proposed approach. Mean results over ten independent trials are shown.

in Fig. 2, the performance at $r = 1$ is much worse than the baseline, thus a value of $r < 1$ should be ensured. Another observation to be made is that simply using the second closest neighbor alone ($r = 0$) leads to strong performance gains.

3) Effect of varying the number K of k-NN: In Fig. 3, we show the performance as a function of K. On the evaluation set, the performance of k-NN drops for the first few values of K, and starts reaching the peak from around $K = 16$, followed by a long tail that eventually meets the ending point of GWRP ($r = 1$). Compared to GWRP, k-NN shows a higher performance upper bound. Although GWRP and k-NN start and end at the same points, k-NN has a much flatter peak than that of GWRP, making it easier for hyperparameter tuning. Our experiment runs on a dataset containing 1000 datapoints ($N = 1000$). In this case, $[16, 32]$ is easily a safe range for K. Considering performance, efficiency, and hyper-parameter tuning, our findings suggest k-NN as the better choice for most use cases.

4) Comparison to the state of the art: As a final experiment, we verify the effectiveness of the proposed re-scaling approach on the DCASE2020 and the DCASE2024 ASD datasets [1], [5], and compare its performance to the state-of-the-art systems on the DCASE2020 and DCASE2023 datasets [1], [4]. For all experiments, no parameters have been changed or tuned except for the number of training epochs. For the DCASE2020 dataset and the DCASE2024 dataset, 15 epochs and 10 epochs were used, respectively. Otherwise, the ASD system was used and evaluated as presented above, with (or without) a k-NN re-scaling using $K = 16$ next nearest neighbors.

The results can be found in Table II and the following observations can be made. Most importantly, it can be seen that the re-scaling consistently improves the resulting performance in domain-shifted conditions, especially in the target domain (cf. Table I), for both the single models and the ensembles. Even for the challenging DCASE2024 ASD dataset, for which the baseline system does not perform well, the performance still substantially improved. When evaluating the effect on the performance on a dataset without domain shifts, i.e., the DCASE2020 dataset, one can observe that the performance of the individual models degrades but, interestingly, the ensemble still performs similarly well regardless of whether the re-scaling approach is applied or not. Overall, the performance we obtained is very similar on the development and evaluation splits of individual datasets, indicating that not much fine-tuning of hyperparameters is needed to obtain a good performance. Last but not least, the experimental results show that the presented system outperforms the state-of-the-art systems on both the DCASE2020 and DCASE2023 ASD datasets.

IV. CONCLUSIONS

In this work, a simple yet effective re-scaling approach to calibrate anomaly scores of different data domains was presented. This is achieved by computing the ratio between the distance of a test sample to a given reference sample and the distances of this reference sample to its nearest neighbors, and using the minimum of these normalized distances over the reference set as an anomaly score. In experiments conducted on multiple datasets in domain-shifted conditions, this re-scaling approach was shown to significantly improve the performance. As a result, our proposed system achieves a new state-of-the-art performance on the DCASE2020 and DCASE2023 ASD datasets.

References

[1] Y. Koizumi, Y. Kawaguchi, K. Imoto, T. Nakamura, Y. Nikaido, R. Tanabe, H. Purohit, K. Suefusa, T. Endo, M. Yasuda *et al.*, “Description and discussion on DCASE2020 challenge task2: Unsupervised anomalous sound detection for machine condition monitoring,” in *Proc. DCASE*, 2020, pp. 81–85.

[2] 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 detection for machine condition monitoring under domain shifted conditions,” in *Proc. DCASE*, 2021, pp. 186–190.

[3] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto *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.

[4] K. Dohi, K. Imoto, N. Harada, D. Niizumi, Y. Koizumi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, and Y. Kawaguchi, “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.

[5] T. Nishida, N. Harada, D. Niizumi, D. Albertini, R. Sannino, S. Pradolini, F. Augusti, K. Imoto, K. Dohi, H. Purohit *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.

[6] Kevin Wilkinghoff, “Audio embeddings for semi-supervised anomalous sound detection,” Ph.D. dissertation, University of Bonn, 2024.

[7] 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.

[8] R. Müller, F. Ritz, S. Illium, and C. Linnhoff-Popien, “Acoustic anomaly detection for machine sounds based on image transfer learning,” in *Proc. ICAART*, 2021, pp. 49–56.

[9] 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.

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

[11] 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.

[12] Y. Deng, A. Jiang, Y. Duan, J. Ma, X. Chen, J. Liu, P. Fan, C. Lu, and W. Zhang, “Ensemble of multiple anomalous sound detectors,” in *Proc. DCASE*, 2022.

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

[14] 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.

[15] 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.

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

[17] A. Kolesnikov and C. H. Lampert, “Seed, expand and constrain: Three principles for weakly-supervised image segmentation,” in *Proc. ECCV*, ser. Lecture Notes in Computer Science, vol. 9908, 2016, pp. 695–711.

[18] 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.

[19] 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.

[20] K. Dohi, T. Nishida, H. Purohit, R. Tanabe, T. Endo, M. Yamamoto, Y. Nikaido, and Y. Kawaguchi, “MIMII DG: Sound dataset for malfunctioning industrial machine investigation and inspection for domain generalization task,” in *Proc. DCASE*, 2022, pp. 31–35.

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

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

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

[24] H. Chen, Y. Song, Z. Zhuo, Y. Zhou, Y. Li, H. Xue, and I. McLoughlin, “An effective anomalous sound detection method based on representation learning with simulated anomalies,” in *Proc. ICASSP*, 2023.

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

[26] 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.

[27] H. Purohit, R. Tanabe, T. Ichige, T. Endo, Y. Nikaido, K. Suefusa, and Y. Kawaguchi, “MIMII dataset: Sound dataset for malfunctioning industrial machine investigation and inspection,” in *Proc. DCASE*, 2019, pp. 209–213.

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

[29] 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.

[30] B. Han, Z. Lv, A. Jiang, W. Huang, Z. Chen, Y. Deng, J. Ding, C. Lu, W. Zhang, P. Fan *et al.*, “Exploring large scale pre-trained models for robust machine anomalous sound detection,” in *Proc. ICASSP*, 2024, pp. 1326–1330.

[31] 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.

[32] A. Jiang, B. Han, Z. Lv, Y. Deng, W.-Q. Zhang, X. Chen, Y. Qian, J. Liu, and P. Fan, “AnoPatch: Towards better consistency in machine anomalous sound detection,” in *Proc. Interspeech*, 2024, pp. 107–111.

[33] P. Saengthong and T. Shinozaki, “Deep generic representations for domain-generalized anomalous sound detection,” 2024, arXiv:2409.05035.