TASK-ADAPTED DUAL-MICROPHONE REPRESENTATIONS WITH DOMAIN-CONDITIONED LOCAL-DENSITY FUSION

Technical Report

Jialong Lei

CRRC, CRRC QINGDAO SIFANG CO., LTD., Shandong, China, leijialong99@163.com

ABSTRACT

Dual-microphone first-shot unsupervised anomalous sound detection requires a detector to separate weak machine faults from operating-condition changes and environmental interference, while only normal sounds are available for training. We build a multi-representation detector around a parameter-efficient, task-adapted dual-microphone EAT encoder. Synchronized near and far channels share the pretrained backbone; low-rank attention updates and metadata-supervised objectives reshape intermediate statistics while preserving the transferable acoustic prior. At inference, the adapted statistics are scored against source- and target-domain normal memories using domain-conditioned local-density normalization and fused with reconstruction, CED, and M2D evidence. Controlled experiments show that covariance conditioning and task adaptation provide complementary gains: PCA whitening improves frozen EAT from 0.5834 to 0.6074, while DualMic-EAT reaches 0.6141 ± 0.0004 over two seeds (sample standard deviation 0.0004). Fixed-configuration outer leave-one-machine-out (LOMO) validation gives 0.6164, and the submitted four-branch cross-fitted fusion reaches 0.6250. A leave-one-branch-out study identifies M2D as the strongest complementary branch and motivates pruning the redundant frozen EAT anchor. The complete-development score of 0.6355 is reported only as a descriptive system-selection result.

***Index Terms*—** anomalous sound detection, domain generalization, audio transformer, parameter-efficient adaptation, local density, score fusion

1. INTRODUCTION

DCASE 2026 Task 2 extends first-shot unsupervised anomalous sound detection (UASD) to synchronized recordings captured near and far from a target machine [1]. This benchmark line builds on domain-shifted miniature and industrial machine-sound corpora, including ToyADMOS2 [2] and MIMII DG [3]. The development and evaluation machine types are disjoint, training data contain normal sounds only, and each target domain provides only ten normal training clips per machine. The task is therefore not simply a denoising problem. A competitive detector must preserve previously unseen fault cues, remain calibrated under source–target mismatch, and transfer its hyperparameters to new machine types. These requirements continue the first-shot protocol of DCASE 2023 [4], while providing a second observation of the same scene with stronger environmental interference.

Prior DCASE analyses show that large pretrained audio models and multi-model systems are effective for first-shot ASD [4, 5]. However, a generic embedding may be insensitive to weak mechanical deviations, and its Euclidean geometry can be dominated by high-variance directions unrelated to faults. The far channel adds further ambiguity. It contains more environmental interference, but it is not a noise-only reference: machine radiation, reverberation, and microphone transfer functions remain mixed in the signal. Direct subtraction can therefore attenuate both nuisance and fault-related evidence. Finally, the strongly imbalanced source and target normal memories have different local scales, making a single global distance threshold poorly conditioned [6].

We address these issues as coupled representation and calibration problems. A reconstruction autoencoder (AE) retains short-time spectral-deviation evidence; CED and M2D provide heterogeneous pretrained priors; and DualMic-EAT adds task-aware evidence through parameter-efficient adaptation. A frozen EAT anchor is retained for controlled comparison but removed from the submitted fusion after cross-fitted branch ablation. All submitted embedding branches use domain-separated normal memories and local-density normalization. Robust calibration and non-negative fusion are selected without the held-out machine in each LOMO fold. The resulting contribution is not a single larger network, but an evidence chain in which each representation has a defined role and each reported result is tied to an explicit model-selection scope.

2. SYSTEM DESIGN

2.1. Short-context reconstruction prior

The AE branch follows the official first-shot baseline [7]. At 16 kHz, a 1024-sample analysis window and 512-sample hop correspond to 64 ms and 32 ms, respectively. Each frame contains 128 log-mel coefficients; five consecutive frames form a 640-D vector covering approximately $64 + 4 \times 32 = 192$ ms. This context is long enough to expose short mechanical transitions or periodic irregularities, while still generating many training vectors per clip and avoiding premature clip-level pooling.

Figure 1 makes the architectural constraint explicit. An initial projection contracts the 640-D observation to 128 dimensions; three additional 128-to-128 transformations maintain a constant-width hidden platform before the final encoder projection reaches the eight-dimensional bottleneck. The decoder mirrors this stepped contraction to recover the input dimension. This symmetry is useful because anomaly evidence is measured in the input feature space rather than in a classifier head, while the narrow bottleneck discourages near-identity reconstruction. At scoring time, the 640-D residual is reorganized into five 128-D blocks. Source- and target-normal precision matrices provide separate domain-conditioned Mahalanobis distances, and the smaller distance is retained [8]. This models correlated, unequal-variance residuals and improves the complete-development metric from 0.5617 to 0.5809.

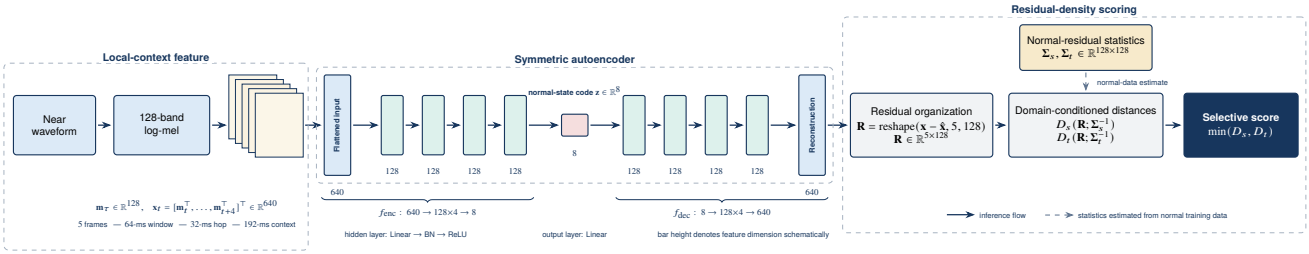

Figure 1: Near-channel AE branch. Five adjacent 128-band log-mel frames form the 640-D input. The stagewise 640–128×4–8–128×4–640 path forces normal short-context patterns through an eight-dimensional bottleneck. The 640-D reconstruction residual is reorganized into five 128-D blocks and evaluated with source- and target-normal covariance models.

2.2. Pretrained representation bank and pruning

The reconstruction view is complemented by frozen audio encoders. CED-Base [9] provides a supervised audio-tagging geometry, while M2D-AS [10] contributes a masked self-supervised representation. EAT-Base AS2M [11] is evaluated as a generic block-10 anchor, but cross-fitted ablation shows that its evidence is redundant with the adapted and masked-model branches; it is therefore excluded from the submitted fusion. For each selected block, token-wise mean and standard deviation are concatenated. The mean summarizes persistent spectral-temporal content, whereas the standard deviation retains within-clip modulation that can be lost under mean pooling.

The microphones are not forced into a shared waveform representation. CED and M2D use near-channel statistics together with the representation contrast $z_{\text{near}} - z_{\text{far}}$. This contrast exposes channel-dependent noise and propagation effects without requiring waveform-scale phase alignment. Branch definitions are fixed across machine types.

2.3. Task-adapted DualMic-EAT

Both microphone channels pass through one EAT backbone. Weight sharing keeps the near and far embeddings in a common coordinate system and avoids doubling the pretrained parameters. The input front end uses Kaldi-style filterbanks and SpecAugment [12]. Both the submitted full-development model and the outer-LOMO audit activate adapters in the QKV and output projections of transformer blocks 6–10 and extract block-10 statistics. This middle-to-late placement preserves early spectro-temporal filters as a generic acoustic front end, while permitting higher-level attention interactions to respond to machine, domain, and microphone structure. Updating every block would increase the number of free parameters and the risk of fitting seven development machine identities; adapting only the final projection would provide too little depth for channel-conditioned reorganization.

Rank-8 LoRA [13] represents each attention update as a low-rank product. The rank is a capacity-control choice rather than a claim of universal optimality: it gives the adapted blocks enough independent directions to modify attention. Together with the task heads, approximately 0.99 million of 91.4 million parameters receive gradients (about 1.1%). This compact update is important because only 7000 normal development clips are available and training lasts two epochs.

The training objective is

$$\mathcal{L} = \mathcal{L}_{\text{attr}} + \lambda_m \mathcal{L}_{\text{mach}} + \lambda_d \mathcal{L}_{\text{dom}} + \lambda_c \mathcal{L}_{\text{con}}. \tag{1}$$

Here, $\mathcal{L}_{\text{attr}}$ is an ArcFace objective [14] over machine–domain–attribute proxies and remains the principal term. Auxiliary machine and domain classifiers prevent the proxy structure from being explained by a single metadata factor. The consistency term weakly aligns synchronized channels without requiring identical embeddings. We set $\lambda_m = 0.25$, $\lambda_d = 0.15$, and $\lambda_c = 0.05$: this ordering keeps attribute discrimination dominant, uses domain prediction as an explicit regularizer, and assigns the smallest weight to cross-channel consistency because the microphones should share machine content but not their complete acoustic realization.

The 256-D ArcFace head organizes the training space, but it is not assumed to be the best anomaly representation. A compact classifier can discard within-class variation that is irrelevant to metadata prediction yet useful for unknown-fault detection. Controlled evaluation confirms this distinction: the compact head obtains 0.5590, raw block-10 near statistics 0.5926, and PCA-768-whitened statistics 0.6144. We therefore retain the block-10 token mean and standard deviation at inference. Near–far concatenation is also not automatically superior: the corresponding near/difference score is 0.6114, so the final adapted branch uses near-channel statistics after dual-channel training.

2.4. Domain-conditioned scoring and fusion

A nearest-neighbor anomaly score is sensitive to the density of its reference bank. This is problematic when the source memory contains 990 clips and the target memory contains only ten. Let $\mathcal{B}_d$ be the normal memory for domain $d \in \{s, t\}$ and $n_d(z)$ the nearest reference to query z. Following local-density normalization [15], we define

$$D_d(z) = \frac{\|z - n_d(z)\|}{\sum_{u \in \mathcal{N}_{K_d}(n_d(z))} \|n_d(z) - u\| + \epsilon}, \tag{2}$$

$$A_b(z) = \min\{D_s(z), D_t(z)\}. \tag{3}$$

The denominator estimates the normal neighborhood scale around the matched reference, excluding the reference itself. Source and target memories are scored separately before the minimum is taken; a target-like normal clip is therefore not penalized for being distant from the much larger source cluster. The submitted DualMic-EAT branch uses $(K_s, K_t) = (8, 5)$, keeping K_t below the ten target references. Euclidean distance is used for EAT-family and CED features, while cosine distance is retained for M2D. Standardization, PCA, and whitening [16] are fitted only on normal training embeddings.

Table 1: Roles of the submitted system's representation branches. "DLD" denotes domain-conditioned local-density scoring.

Branch	Primary evidence	Microphone use	Conditioning	Score
AE	Local log-mel reconstruction residual	Near	residual covariance	selective Mahalanobis
CED	Audio-tagging semantics with channel contrast	Near and near–far	none	DLD, Euclidean
M2D	Masked-modeling geometry and complementary invariance	Near and near–far	PCA-768	DLD, cosine
DualMic-EAT (Ours)	Fault-sensitive, task-adapted intermediate statistics	Shared near/far encoder	PCA-768 whitening	DLD, Euclidean

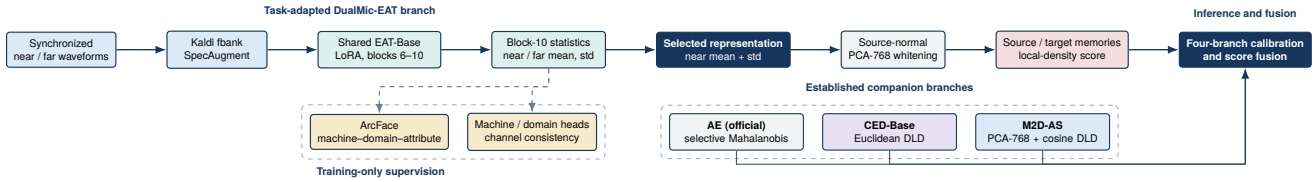

Figure 2: DualMic-EAT training and inference. The upper path is our task-adapted branch; the lower row summarizes the official AE baseline and pretrained CED/M2D companions detailed in Table 1. Full-development adapts blocks 6–10, as does each outer-LOMO retraining fold. Training heads shape the representation; inference scores whitened block-10 near statistics and fuses four branches. Frozen EAT is excluded after cross-fitted ablation.

Raw branch scales are not directly comparable. For branch b, a robust sigmoid g_b is parameterized by the training-fold median and interquartile range. Fusion is

$$A_{\text{fuse}}(z) = \sum_{b \in \mathcal{B}} w_b g_b(A_b(z)), \quad w_b \geq 0, \quad \sum_b w_b = 1. \qquad (4)$$

Weights are searched on a 0.1 simplex grid within each selection fold. The complete-development configuration used for evaluation output assigns AE/CED/M2D/DualMic weights of 0.3/0.1/0.2/0.4. The coarse grid acts as selection regularization: refining it to 0.05 raises the complete-development score but reduces cross-fitted LOMO. No statistic is estimated from evaluation test clips.

3. EXPERIMENTAL PROTOCOL

The development set contains seven machine types, each with 990 source-domain and ten target-domain normal training clips plus 200 labeled test clips. The additional training and evaluation sets contain five unseen machine types. The official score $\Omega \uparrow$ is the harmonic mean over every machine-level source-domain AUC, target-domain AUC, and pAUC; pAUC is evaluated up to a false-positive rate of 0.1. This aggregate penalizes a detector that succeeds on one machine or domain while failing on another, so component metrics are also reported.

DualMic-EAT is trained with BF16, effective batch size 40, learning rate 5×10^{-5}, and two epochs. We distinguish three evidence levels. *Complete-development* experiments isolate representation and scoring choices using all seven machine types. *Fixed-configuration outer LOMO* reinitializes and trains LoRA on six machine types, then scores the excluded type with fixed representation hyperparameters. *Cross-fitted fusion* also estimates calibration and weights only from the six non-held-out machines. The fixed DualMic audit uses the original $16/9$ density neighborhood.

Table 2: Controlled complete-development study. The best task-adapted row is marked as Ours.

Configuration	Scoring	$\Omega \uparrow$
Near AE	mean squared error	0.561682
Near AE	selective Mahalanobis	0.580897
Frozen EAT block 10	raw DLD	0.583375
Frozen EAT block 10	PCA-512W DLD	0.602619
Frozen EAT block 10	PCA-768W DLD	0.607410
DualMic-EAT head	DLD	0.558972
DualMic-EAT raw statistics	DLD	0.592583
DualMic-EAT (Ours), $s = 13711$	PCA-768W DLD	**0.614391**
DualMic-EAT, $s = 2026$	PCA-768W DLD	0.613875

The submitted $8/5$ setting, selected after an aggregate screen of the seven outer folds, raises 0.616416 to 0.618768 but is configuration-selection evidence, not an independent outer estimate. Other branch hyperparameters likewise precede cross-fitting; thus, this protocol is stronger than complete-development selection but is not fully nested.

4. RESULTS AND ANALYSIS

Table 2 separates representation learning from distance conditioning. Increasing whitened PCA from 512 to 768 components improves frozen EAT by 0.0048, and 768-component whitening is 0.0240 above raw DLD. Thus, the pretrained encoder already contains useful information, but its raw covariance makes nearest-neighbor geometry inefficient. DualMic adaptation adds a further 0.0070 over the strongest frozen EAT configuration. Across two seeds, the mean is 0.614133 with a sample standard deviation of 0.000364; both runs exceed the frozen EAT result.

Table 3 tests transfer to an excluded machine type. The fixed

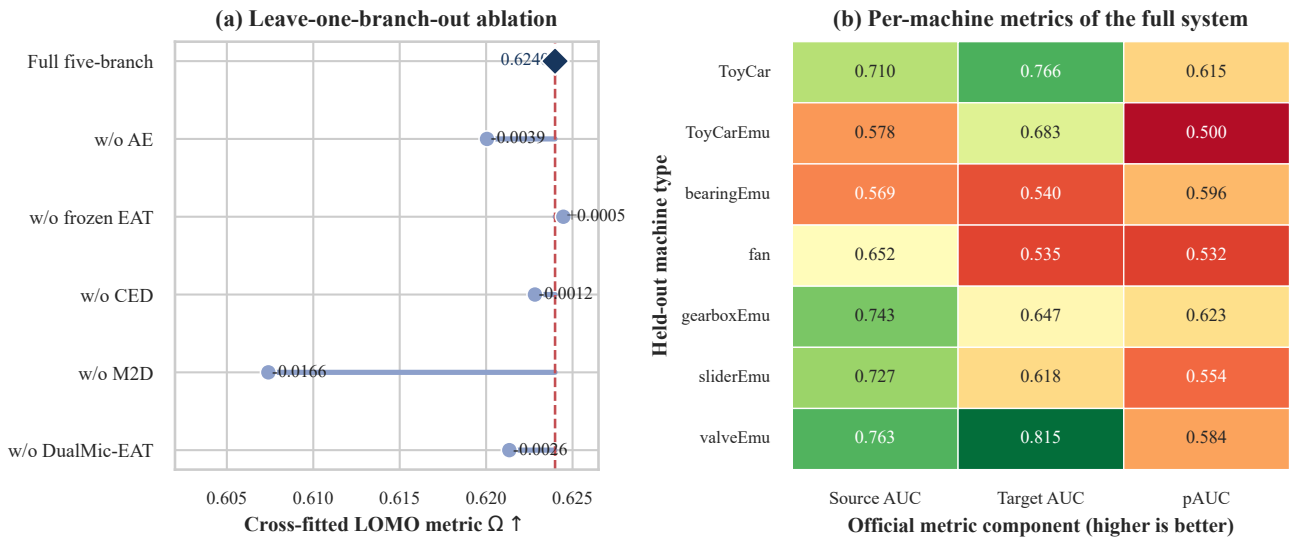

Figure 3: Cross-fitted pruning evidence. (a) Each row removes one branch from the five-branch screening system and repeats fold-specific calibration and weight selection. The positive frozen-EAT removal result motivates the submitted four-branch design. (b) Source AUC, target AUC, and pAUC of the screening system for every held-out machine.

Table 3: Generalization-aware results. Scores from different selection scopes are separated explicitly.

System and protocol	Selection scope	Ω ↑
Pre-DualMic fusion, strict nested LOMO	six machines	0.601961
DualMic-EAT, fixed LOMO (16/9)	six machines	0.616416
Five-branch fusion, cross-fitted LOMO	six machines	0.623988
Submitted four-branch (Ours), cross-fitted	**six machines**	**0.625003**
Submitted four-branch (Ours), complete dev.	seven machines	0.635498

DualMic result is 0.6164; the five-branch screen reaches 0.6240, while pruning frozen EAT and using the selected $8/5$ neighborhood raises the submitted four-branch cross-fitted fusion to 0.6250, 0.0230 above the earlier strict nested four-branch result. The 0.6355 complete-development score guides the evaluation configuration but is not an unbiased estimate of hidden-set performance.

Figure 3(a) provides an apples-to-apples component test on the candidate five-branch pool. Removing M2D causes the largest reduction (-0.0166), confirming that its masked-modeling geometry contributes information not recovered by EAT or reconstruction. AE, CED, and DualMic-EAT reduce Ω by 0.0039, 0.0012, and 0.0026 when removed. Frozen EAT is almost fully represented by the other transformer branches: removing it changes Ω by $+0.0005$. This motivates pruning the frozen EAT anchor from the submitted fusion rather than carrying a branch with negligible cross-fitted complementarity. Repeating the four-branch evaluation with a coarser 0.1 fusion grid and the smaller $8/5$ neighborhood gives the best cross-fitted result; a finer 0.05 grid falls to 0.6220, a pattern consistent with selection overfitting.

The heat map in Fig. 3(b) explains why a single scalar is insufficient. Target AUC remains strong for ToyCar and valveEmu, but is 0.540 on bearingEmu and 0.535 on fan. Low-FPR behavior is the limiting component for ToyCarEmu (pAUC 0.500), fan (0.532), and sliderEmu (0.554). Their mismatch with source AUC points to tail calibration and target-memory regularization, rather than encoder capacity alone, as the next priorities.

5. SUBMITTED SYSTEMS AND CONCLUSION

Submitted systems: **Lei_CRRC_task2_1 (NearMah)**, near-channel Mahalanobis AE; **Lei_CRRC_task2_2 (AEM2DFuse)**, AE/M2D fusion; **Lei_CRRC_task2_3 (ResMah)**, residual Mahalanobis AE; and **Lei_CRRC_task2_4 (DualDLD4)**, the pruned four-branch DualMic system. Their hyperparameters are shared across evaluation machine types.

Three conclusions emerge. Whitening confirms that representation and distance geometry must be designed jointly. Low-rank dual-microphone adaptation is most effective when supervised heads organize training but inference preserves information-rich intermediate statistics. Finally, common-protocol ablation validates M2D complementarity and the removal of redundant frozen EAT. The submitted system therefore combines task-aware and pretrained evidence while keeping controlled, held-out-machine, and complete-development results distinct.

6. REFERENCES

[1] T. Nishida, N. Harada, D. Takeuchi, D. Niizumi, K. Imoto, K. Dohi, H. Purohit, T. Endo, and Y. Kawaguchi, "Description and discussion on DCASE 2026 challenge task 2: Noise-aware unsupervised anomalous sound detection for machine condition monitoring," *arXiv preprint arXiv:2606.01578*, 2026.

[2] 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 *Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE)*, Barcelona, Spain, November 2021, pp. 1–5.

[3] 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 *Proceedings of the 7th Detection and Classification of Acoustic Scenes and Events 2022 Workshop (DCASE2022)*, Nancy, France, November 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 Workshop*, 2023.

[5] P. Saengthong and T. Shinozaki, "Deep generic representations for domain-generalized anomalous sound detection," *arXiv preprint arXiv:2409.05035*, 2024.

[6] 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," *arXiv preprint arXiv:2503.10435*, 2025.

[7] 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, pp. 191–195.

[8] P. C. Mahalanobis, "On the generalized distance in statistics," *Proceedings of the National Institute of Sciences of India*, vol. 2, no. 1, pp. 49–55, 1936.

[9] H. Dinkel, Y. Wang, Z. Yan, J. Zhang, and Y. Wang, "CED: Consistent ensemble distillation for audio tagging," *arXiv preprint arXiv:2308.11957*, 2023.

[10] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Learning representations by encouraging both networks to model the input," in *Proc. IEEE ICASSP*, 2023.

[11] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," *arXiv preprint arXiv:2401.03497*, 2024.

[12] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[14] 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.

[15] K. Wilkinghoff, H. Yang, J. Ebbers, F. G. Germain, G. Wichern, and J. Le Roux, "Local density-based anomaly score normalization for domain generalization," *arXiv preprint arXiv:2509.10951*, 2025.

[16] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer, 2002.