IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Zero-Shot Denoiser for Enhanced Acoustic Inspection: Blind Signal Separation and Text-Guided Audio Reconstruction

Koki Shoda[1], Jun Younes Louhi Kasahara[2], Qi An[1], and Atsushi Yamashita[1]

***Abstract*—Acoustic inspection is crucial for infrastructure maintenance, but its effectiveness is often hampered by environmental noise. Conventional denoising methods rely on prior knowledge or training data, limiting their practicability. This paper presents Zero-Shot Denoiser, a novel approach achieving noise reduction without pre-collected target sound samples or noise knowledge. Our method synergistically combines Blind Signal Separation (BSS) for unsupervised audio decomposition and Artifact-Resilient Attention (AR-Attention) for text-guided audio reconstruction. AR-Attention leverages pre-trained audio-language models and dual normalization to mitigate BSS artifacts and identify target sounds semantically. We introduce pseudo Signal-to-Noise Ratio, derived from the audio-language model, for automatic BSS hyperparameter optimization. In experiments using public datasets, our method, operating in a true zero-shot setting, achieved performance comparable to that of state-of-the-art supervised denoising methods, and experiments targeting hammering tests confirmed the effectiveness of our approach for real-world acoustic inspections. Our approach overcomes the limitations of data-dependent techniques and offers a versatile noise reduction solution for acoustic inspection and broader acoustic tasks.**

***Index Terms*—Acoustic Inspection, Audio-Language Model, Noise Reduction, Hammering Test, Anomaly Detection**

I. Introduction

THE safety and operational integrity of manufacturing plants and building infrastructures are critically dependent on regular inspection protocols [1], [2]. Among various inspection modalities, acoustic inspection stands out as a crucial technique, enabling the detection of internal defects and anomalies often invisible through other means. Automating these critical inspections using robotic systems, particularly using mobile robots equipped with sensor arrays [3], [4], [5], has gained significant attraction, promising increased efficiency and safety. A fundamental challenge hindering the reliable deployment of robots for acoustic inspection is the pervasive and often extreme noise present in real-world operational environments. Robust acoustic perception in noisy conditions is therefore a critical prerequisite for successful robotic automation in this domain. Consequently, achieving reliable performance necessitates effective denoising techniques.

Denoising techniques can be broadly categorized into hardware-based and software-based approaches. Hardware solutions, such as active noise cancellation devices, aim to physically isolate the target sound [6]. While offering a degree of effectiveness, these methods typically necessitate close proximity to the sound source. In expansive or intricate environments like industrial facilities or high-rise structures, this requirement translates to increased inspection time and potential risks of collision or damage to both the inspection robot and the inspected asset.

The drawbacks of hardware-based approaches, which require sensors to be positioned very close to the sound source, have driven the development of software-based denoising techniques. Software-based methods can process acoustic signals remotely, meaning they do not need to be in direct contact with the sound source. These can be further divided into filtering-based methods and Blind Signal Separation (BSS) techniques. Filtering-based methods, including deep learning models trained on specific target sounds, have demonstrated promising results [7], [8]. However, their efficacy hinges on the availability of representative samples of the target sounds. This presents a significant hurdle in anomaly detection tasks, where acquiring pre-existing samples of abnormal states is often impractical, thereby limiting their applicability in real-world scenarios.

In contrast, BSS-based approaches offer the distinct advantage of decomposing a mixture of acoustic signals into its constituent sources using only recordings from a microphone array [9]. This eliminates the need for pre-collected target sound samples; however, it also introduces several challenges. A primary concern is determining which of the separated components actually corresponds to the target sound. Various heuristics and constraints have been proposed to address this challenge [10]. For example, Wang et al. [11] employ Direction-of-Arrival (DoA) estimation to locate the target sound, assuming prior knowledge of its spatial position. Although effective in controlled settings, this DoA-based approach tends to falter when the target sound's location is either unknown or subject to change.

In addition, the overall performance of BSS is highly sensitive to several key hyperparameters, notably the number of signal decompositions and the number of bases. While numerous strategies have been developed to estimate these parameters, many of them depend on additional hyperparameters or evaluation metrics [12], [13], [14]. This dependency makes the optimization process both complex and error-prone. As a result, automatically tuning these hyperparameters without prior knowledge, such as the exact number of sources or the probability distributions underlying the source models, remains a challenging problem.

To overcome these constraints, we propose Zero-Shot Denoiser, a synergistic framework that integrates audio-language models with BSS. By effectively combining the unsupervised signal-separation capability of BSS with the generalized zero-shot recognition ability of audio-language models, our approach aims to robustly identify and reconstruct target sounds

Manuscript received: February 11 2025; Revised: 14 April 2025; Accepted: 7 June 2025. This paper was recommended for publication by Editor Hyungpil Moon upon evaluation of the Associate Editor and Reviewers' comments.

[1]K. Shoda, Q. An, and A. Yamashita are with the Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, Japan. {shoda, anqi, yamashita}@robot.t.u-tokyo.ac.jp

[2]J.Y. Louhi Kasahara is with the Department of Precision Engineering, Graduate School of Engineering, The University of Tokyo, Japan. louhi@robot.t.u-tokyo.ac.jp

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

without relying on prior samples or spatial assumptions. This significantly improves both the flexibility and reliability of acoustic noise removal in automated inspection tasks.

Our main contributions are summarized as follows:

- **Introduction of Artifact-Resilient Attention**: The sounds decomposed by BSS exhibit deviations from actual acoustic signals, including unique variations and artifacts inherent to BSS [15], [16]. Our Artifact-Resilient (AR-) Attention mechanism is designed to remain resilient to artifacts. By adaptively assigning weights based on each component's membership to a particular audio class, AR-Attention is able to reconstruct the target audio with high fidelity. Inspired by the principle of superposition in acoustics and the weighted-sum representation intrinsic to attention mechanisms [17], AR-Attention extends the zero-shot recognition capability of audio-language models [18].
- **Automatic Optimization of BSS Hyperparameters**: We observe that the membership score of the reconstructed audio to its intended class, derived from an audio-language model, can be interpreted as a proxy for noise level. We refer to this measure as pseudo-SNR. By maximizing the pseudo-SNR, we introduce a teacher-free strategy for automatically optimizing BSS hyperparameters, eliminating the need for pre-labeled data or intricate manual tuning.
- **Achieving Zero-Shot Denoising**: Unlike conventional denoising methods that rely on prior knowledge of the target's location or vast amounts of labeled data, our proposed method exploits the power of BSS and audio-language embeddings to reconstruct the target sound in a zero-shot manner. This drastically reduces the reliance on data collection and manual annotation.

II. Related Work

In this section, we provide an overview of recent research developments in source separation and denoising. The challenges inherent in source separation and denoising have motivated a wide range of research endeavors. Transformer-based models have demonstrated remarkable precision in source separation by effectively capturing long-range dependencies in audio signals [19]. Furthermore, self-supervised learning techniques have emerged as powerful tools for effective noise reduction without requiring large amounts of labeled data, thereby alleviating the burden of data collection and annotation [20]. Multi-task learning frameworks offer another promising avenue, leveraging shared representations to simultaneously address related tasks, such as musical instrument recognition and source separation, leading to improved overall performance and efficiency [21]. Similarly, multimodal approaches that integrate audio information with visual cues, such as lip movements of speakers, have shown significant promise in enhancing source separation accuracy [22].

Building upon traditional Blind Signal Separation (BSS) techniques, recent advancements have integrated deep learning methodologies to further enhance their capabilities. Supervised deep learning extensions applied to BSS models, such as Deep Multichannel Nonnegative Matrix Factorization (MNMF) [23] and Independent Deeply Learned Matrix Analysis (IDLMA) [24], have demonstrated increased accuracy in source separation tasks. However, a significant limitation of many of these approaches [19]–[24] is their reliance on pre-trained sound classes or assumptions of stable acoustic environments. As a result, their performance degrades when faced with situations where the target sound characteristics differ from those seen during training, or when the number and types of target sounds are not predetermined-conditions often encountered in dynamic or open-set environments.

In the field of image processing, the concept of zero-shot denoising is well-established [25], [26]. Image denoising in a zero-shot setting is typically facilitated by the fact that image noise generally stems from sensor imperfections, optical aberrations, or compression artifacts. Moreover, in most images, the desired signal occupies a large portion of the visual field, enabling denoising algorithms to effectively separate it from noise. However, applying this framework to audio processing presents a significant challenge. Unlike images, audio signals often contain significant amounts of noise relative to the target sound, particularly in real-world scenarios with low Signal-to-Noise Ratios (SNR). Traditional zero-shot denoising techniques, which implicitly assume that the majority of the signal corresponds to the desired content, may therefore struggle to handle audio data where the target signal is only a minor component of the overall acoustic energy [27], [28]. This fundamental difference underscores the need for a specialized approach such as the one we propose.

III. Proposed Method

A. Conceptual Overview

Our approach capitalizes on the complementary strengths of three key modules to realize a zero-shot denoising framework. Crucially, our method relies entirely on a pre-trained Audio-Language Model (ALM) [29] and involves no training or fine-tuning on any specific datasets used in our experiments. This ensures its operation in a truly zero-shot setting. First, the Blind Signal Separation (BSS) module leverages spatial diversity and the inherent statistical independence of audio sources to decompose the captured mixture into multiple candidate components, without requiring any prior information about the target sound. However, BSS alone can produce decompositions that include artifacts or overly fragmented representations of the target signal.

To address this issue, our second module, Artifact-Resilient (AR-) Attention, leverages the capabilities of the large-scale, pre-trained ALM. Specifically, it utilizes the fixed audio and text embeddings derived from this frozen ALM, along with a dual normalization strategy. This mechanism exploits the ALM's inherent zero-shot generalization ability to selectively aggregate relevant decomposed signals based on their semantic similarity to user-specified audio classes, thereby mitigating artifacts and reconstructing the target sound with high fidelity.

Finally, the pseudo-SNR generation module introduces an innovative method for automatically tuning the BSS hyperparameters at inference time for each individual input mixture. This optimization is guided by the fixed, pre-trained ALM [29]

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

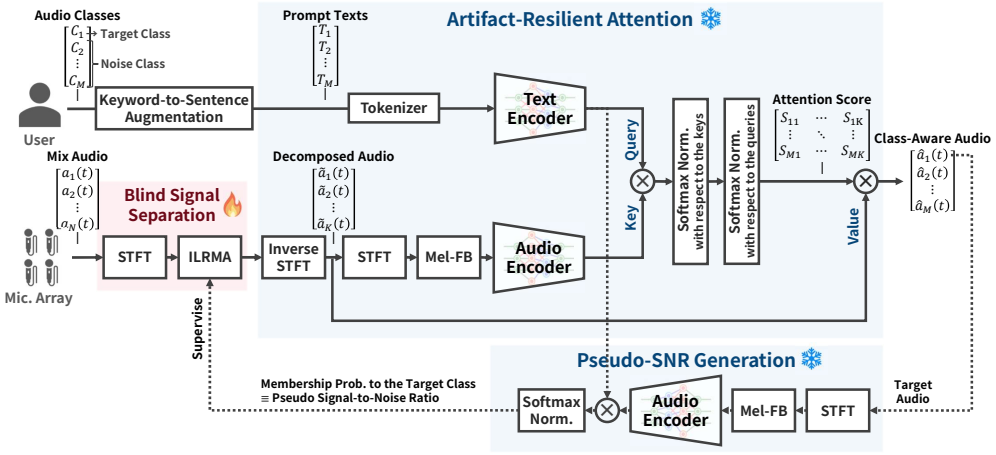

Fig. 1. Overview of the Zero-Shot Denoiser. The flame icon indicates modules that are optimized by the pseudo-SNR, while the snowflake icon indicates modules whose parameters are frozen. Here, $\otimes$ denotes the matrix multiplication, N is the number of microphones, M is the number of audio classes, K is the number of signal decompositions, STFT is the Short-Time Fourier Transform, ILRMA is Independent Low-Rank Matrix Analysis, and Mel-FB stands for the Mel-Filter Bank.

and adjusts a small set of BSS parameters to improve separation quality for that specific input. By interpreting the alignment between the reconstructed signal and its designated audio class (as measured by the ALM) as a proxy for signal clarity, we can optimize the separation process without relying on manual intervention or labeled data.

Figure 1 illustrates the overall architecture of the Zero-Shot Denoiser. In this method, a microphone array captures a mixture of noise and the target sound, which is then decomposed by BSS. Meanwhile, the user provides, in natural language, a list of noise and target sounds, referred to as audio classes. Each class is first converted into a sentence through keyword-to-sentence augmentation, then tokenized. Both the tokenized audio classes and the decomposed signals are embedded into a shared, comparable representation using the pre-trained and fixed text encoder (GPT2) [30] and audio encoder (HTSAT) [31] components of the ALM [29]. These encoders were pre-trained on large, external datasets unrelated to our experiments and are employed here without any further training or fine-tuning. Next, we compute the attention scores by applying softmax normalization, and we reconstruct class-aware audio by weighting and summing the decomposed signals according to attention scores. By simply selecting the desired class from the list, users can effectively achieve noise reduction.

B. Blind Signal Separation for Unsupervised Signal Decomposition

We employ BSS to decompose the multi-channel acoustic signals captured by a microphone array. BSS leverages spatial information, primarily inter-channel time differences, and the statistical independence of source signals to achieve high-precision signal separation. This allows for the decomposition of the mixed audio into plausible individual acoustic signals without the need for prior training or knowledge of the sound source locations. Specifically, we utilize Independent Low-Rank Matrix Analysis (ILRMA) [9], which is a state-of-the-art BSS algorithm that effectively separates mixed audio signals into their constituent sources.

In BSS using ILRMA, first, the input signals from N microphones are represented in a matrix form in the time-frequency domain using the short time Fourier transform (STFT). These mixed signals are then decomposed into K nonnegative matrices, taking into account the phase delays between the microphones (spatial information) and the independence of the sources.

However, BSS methods such as ILRMA have their own inherent challenges. The hyperparameter such as the number of signal decompositions and the number of bases significantly affect the quality of the separated signals. Furthermore, automatically identifying which of the decomposed signals corresponds to the desired target sound remains a significant challenge. To address these challenges, we introduce a novel Artifact-Resilient (AR-) Attention and pseudo-SNR generation mechanism.

C. Artifact-Resilient Attention for Target Sound Reconstruction

To robustly address the challenges of selecting critical BSS hyperparameters while enabling target sound identification, we propose AR-Attention. In AR-Attention, we compute the semantic similarity between the audio classes provided by the user in text and the multiple signals decomposed by BSS $[\tilde{a}_1(t), \tilde{a}_2(t), \ldots, \tilde{a}_K(t)]^T$, then selectively aggregate these signals based on that similarity. This approach allows us to estimate the sound source corresponding to each audio class:

$$\begin{bmatrix} \hat{a}_1(t) \\ \hat{a}_2(t) \\ \vdots \\ \hat{a}_M(t) \end{bmatrix} = \begin{bmatrix} S_{11} & \cdots & S_{1K} \\ \vdots & \ddots & \vdots \\ S_{M1} & \cdots & S_{MK} \end{bmatrix} \begin{bmatrix} \tilde{a}_1(t) \\ \tilde{a}_2(t) \\ \vdots \\ \tilde{a}_K(t) \end{bmatrix}, \tag{1}$$

where $[\hat{a}_1(t), \hat{a}_2(t), \ldots, \hat{a}_M(t)]^T$ is the estimated sound sources corresponding to each audio class, M is the number of audio classes, K is the number of signal decompositions, t is the time index, and S_{ij} denotes the attention score between the i-th audio class and the j-th decomposed audio. Even if BSS decomposes the signal into too many components, splitting the target sound across several signals, AR-Attention can integrate them into a single target sound based on their semantic similarity.

AR-Attention incorporates several key innovations to enhance its robustness and effectiveness. First, the audio en-

coder and text encoder utilize fixed weights from Contrastive Language-Audio Pretraining (CLAP) [29]. This audio-language model, pre-trained on a massive dataset and possessing strong generalization capabilities, plays a crucial role in achieving training-free operation by leveraging the CLAP weights. However, the sounds decomposed by BSS exhibit deviations from actual acoustic signals, including unique variations and artifacts inherent to BSS [15], [16]. These discrepancies can inadvertently reduce the similarity to audio classes. Since CLAP is trained on real-world sounds, these deviations can hinder proper audio embedding and make it difficult to calculate appropriate attention scores.

One representative artifact of BSS is spectral inconsistency [32]. To mitigate the deviation caused by this spectral inconsistency, we, similar to the BSS algorithm proposed by Kitamura et al. [33], perform an inverse STFT followed by another STFT. This allows the spectrograms of the decomposed signals to be projected onto a set of consistent spectrograms.

Beyond spectral inconsistency, BSS relies on assumptions such as independence, sparsity, and stationarity of the input signals for separation. Artifacts arise when the actual acoustic environment or source characteristics deviate from these assumptions, or when the algorithm converges to an approximate solution due to insufficient convergence. Furthermore, over- or under-separation can lead to the inclusion of unwanted components or the loss of necessary components, respectively, both contributing to artifacts.

To mitigate the unintended reduction in similarity between the decomposed signals and specific audio classes caused by these artifacts, normalization is applied via softmax not only along the query (audio class) dimension but also along the key (decomposed signal) dimension. This dual normalization is a key feature of AR-Attention.

The normalization in the key direction transforms the similarity scores into relative weights, indicating the degree to which each decomposed signal contributes to a specific audio class. In essence, by assuming that each signal corresponds to at least one of the given audio classes, we convert the similarities into relative weights, thereby reducing the impact of artifacts. Conversely, the normalization in the query direction ensures that the weights for all decomposed signals sum to one for each audio class, transforming them into a probability distribution across the decomposed signals.

In summary, our proposed method combines the unsupervised signal separation capabilities of BSS with the semantic understanding of audio-language models through the novel AR-Attention mechanism. This allows for zero-shot noise reduction where users can denoise audio by simply specifying the autio classes in natural language, eliminating the need for pre-collected target sound samples or knowledge of their spatial locations.

D. Pseudo-SNR Generation for Automatic Optimization of BSS Hyperparameters

To automate the hyperparameter optimization, another challenge in BSS, we introduce a novel concept called pseudo-SNR that leverages the recognition capabilities of audio-language models. The degree to which class-aware audio, reconstructed by AR-Attention, aligns with its designated audio class can be interpreted as its clarity, effectively representing the extent to which it is free from sounds belonging to other classes. Consequently, we anticipate a correlation between this class affinity of the class-aware audio and the actual Signal-to-Noise Ratio (SNR). Thus, by maximizing this pseudo-SNR, we can automatically optimize the BSS hyperparameters by Tree-structured Parzen Estimator (TPE) [34], which is a Bayesian optimization technique. This approach offers an unsupervised strategy for optimizing BSS hyperparameters, eliminating the need for pre-labeled data or intricate manual tuning.

A seemingly straightforward approach to noise reduction using pseudo-SNR would be to directly train an end-to-end deep neural network for filtering noise, using pseudo-SNR as the sole training signal. While this idea might appear attractive at first glance, it comes with critical drawbacks. In practice, pseudo-SNR is an indirect and approximate measure of audio quality, relying exclusively on it as the training objective can lead to overfitting. In other words, the network may learn to improve the pseudo-SNR metric, by exploiting its specific characteristics, without necessarily achieving SNR improvements or generalizing well to noise conditions.

In contrast, our proposed method integrates BSS-based signal decomposition with AR-Attention reconstruction. This design inherently embeds strong inductive biases derived from the physical properties of sound and the structured separation of audio components. Such biases constrain the optimization process, thereby reducing the risk of overfitting to the proxy pseudo-SNR metric. The resulting synergistic effect ensures a balanced trade-off between effective noise suppression and high-fidelity audio reconstruction, ultimately leading to a more robust noise reduction performance.

IV. Experiments

To verify the effectiveness of our proposed method, we conducted experiments using both a publicly available dataset and a real-world setting. The hyperparameters and configuration were as follows. For keyword-to-sentence augmentation, we used the prompt *'this is a sound of [class label].'*, this is the same prompt that was used in the original ALM paper's [29] evaluation of zero-shot classification. For the BSS hyperparameters, we performed Bayesian optimization with 10 trials, where the number of bases ranged from 4 to 8 and the number of decompositions ranged from 2 to 4. In optimizing the demixing filters for BSS, we employed both the Majorization-Minimization Algorithm [35] and the Iterative Projection Algorithm [36] for 100 iterations, thereby carrying out constrained local optimization efficiently. Furthermore, in AR-Attention, we fixed the weights of the audio and text encoders to those of CLAP [29], set the softmax temperature for the query direction to 30, and for the key direction to 0.01. The implementation and audio samples are publicly available at https://github.com/kokieto/ZeroShotDenoiser.git.

A. Public Dataset Evaluation

1) Setup: To evaluate the effectiveness of the proposed method across diverse sound sources we conducted experiments. As there was no suitable dataset available for evaluating

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
NUMBER OF SAMPLES FOR EACH AUDIO CLASS.

Category	Audio Class	Number of Samples
Target	mechanisms	212
Noise	music	253
	string	158
	nature	386
	speech	152
	animals	150
	keyboard	54
	wind instruments	229
	synths or electronic	69
	crowd or conversation	59
	human sounds and actions	351

noise reduction in multi-channel, multi-source environments, we created our own by combining publicly available datasets. Sound sources were obtained from BSD10k [37], and multi-channel spatial sound sources were reproduced by convolving them with a calibrated 5-channel Room Impulse Response (RIR) dataset [38]. To use various mixed sound sources, we used sounds with an effective duration of over 5 seconds after convolving with RIRs of approximately 0.4 seconds, and then trimmed them to 5-second segments, centered around the point of maximum amplitude.

Assuming anomaly detection based on machine sounds, we designated the "mechanisms" class from BSD10k (212 items) as the target sound sources. Noise sound sources were used 10 non-machine sound classes in BSD10k, totaling 1,861 items, and the number of samples for each audio class are shown in Table I. The RIR dataset contains four sound source locations and six microphone-array positions. For synthesis, we randomly selected a microphone-array position, and randomly chose two distinct sound source locations, one for the target sound and one for the noise sound, from the available 4 positions. The noise sample was then randomly selected from a pool of 1,861 samples, and the target and noise sounds were mixed at an SNR of 0 dB.

Since no zero-shot acoustic noise reduction methods have been proposed to date, we selected NB-BLSTM [39], TF-GridNet [40], NBC [41], and SpacialNet [7] as comparative baselines. These models represent the current state-of-the-art in multi-channel supervised noise reduction. We trained each model with the RAdamScheduleFree [42] optimizer and terminated training once the validation loss plateaued.

2) Results: An example of the attention score is shown in Fig. 2. In this example, when the input audio is decomposed into three signals, the SNR improvement rate was 9.4 dB. To understand the separation in this specific case, we manually listened to the resulting decomposed signals and labeled them based on our auditory perception as follows: in order, mixed low-frequency reverberation, speech audio, and a motor sound of varying volume. The target sound source, corresponding to the "mechanisms" class query, has a concentrated attention score on the index 3 signal, confirming that appropriate weighting is being performed by our AR-Attention.

To quantitatively assess the individual contributions of the key components of our proposed Zero-Shot Denoiser, we conducted an ablation study, summarized in Table II. The Scale-Invariant (SI-) SNR is defined by reference (clean) audio $a(t)$ and denoised audio $\hat{a}(t)$ as follows:

$$\text{SI-SNR} = 10\log_{10}\frac{\|\delta a(t)\|^2}{\|\delta a(t)-\hat{a}(t)\|^2}, \quad \delta = \frac{\langle \hat{a}(t), a(t)\rangle}{\|a(t)\|^2}, \tag{2}$$

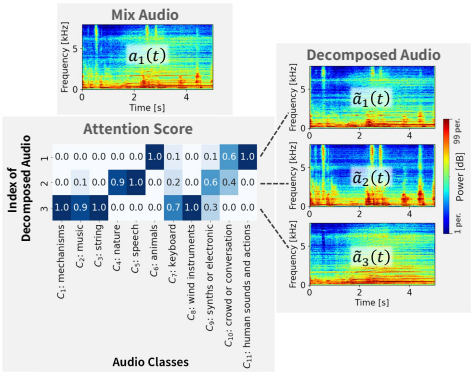

Fig. 2. Example of attention scores and decomposed signals.

where the notation $\langle \hat{a}(t), a(t)\rangle$ denotes the inner product of the $\hat{a}(t)$ and $a(t)$.

The results in Table II highlight the effectiveness of our proposed components. 'Full Method w/ BO' compared to 'w/o Dual Normalization' demonstrates the benefit of dual normalization in AR-Attention. This strategy normalizes attention scores across both audio class (query) and decomposed signal (key) dimensions, making AR-Attention more robust to BSS artifacts and enabling more accurate semantic selection and aggregation for improved SI-SNR compared to standard attention. Furthermore, comparing 'Full Method w/ BO' with 'w/o Pseudo-SNR Opt.' shows the advantage of pseudo-SNR guided optimization. Adaptively tuning BSS hyperparameters, specifically the number of bases and decomposition number, for each input using the pseudo-SNR metric allows better adaptation to specific signal characteristics, yielding superior performance over a fixed setting. Regarding the optimization strategy, a Brute-Force (BF) search that evaluated all 15 combinations achieved the highest SI-SNR of 5.21 dB. In contrast, our Bayesian Optimization (BO) approach, which used 10 evaluations, yielded comparable results at 4.90 dB but required only approximately 67% of the BF computational cost. This shows BO offers a favorable performance-efficiency trade-off for the computationally intensive pseudo-SNR evaluation, though BF remains feasible for this limited search space if computational cost is less critical.

Table III shows a comparison of the noise reduction performance of the proposed method and the comparison methods. The proposed method achieved zero-shot noise reduction performance comparable to state-of-the-art supervised noise reduction methods that requires thousands of training data. This comparison highlights the fundamental difference and trade-offs between zero-shot and supervised approaches in the context of multi-channel noise reduction tasks.

Supervised methods learn to suppress noise based on patterns observed in large training datasets. Their performance generally improves with more training data, allowing them to potentially achieve very high performance when sufficient relevant data is available. However, our zero-shot method, leveraging unsupervised BSS principles and semantic guidance from a pre-trained ALM, operates independently of task-specific training data. While its performance relies on the inherent separation capabilities of BSS and the generalization power of the ALM, which may have limits compared to

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE II
ABLATION STUDY RESULTS. THE BEST RESULT IS SHOWN IN BOLD, AND THE SECOND BEST IS UNDERLINED.

Configuration	Description	Average SI-SNR Imp. [dB]
Full Method w/ BO	Full method with AR-Attention (dual norm.) and pseudo-SNR guided Bayesian Optimization (BO).	<u>4.90</u>
Full Method w/ BF	Full method with AR-Attention (dual norm.) and pseudo-SNR guided Brute-Force (BF) search.	**5.21**
w/o Dual Normalization	AR-Attention uses standard attention (only query-dimension norm.). Pseudo-SNR Opt. with BO used.	4.60
w/o Pseudo-SNR Opt.	AR-Attention (dual norm.) used. BSS fixed to the best setting (2 decomps. and 4 bases).	4.38

TABLE III
AVERAGE SI-SNR IMPROVEMENT ON TEST DATA.

Method	Number of Training Data 800	1600	3200
NB-BLSTM [39]	-1.09	0.78	1.18
TF-GridNet [40]	1.31	4.79	**7.36**
NBC [41]	1.64	3.17	4.05
SpatialNet [7]	2.27	4.18	7.06
Prop. (No Training)	**4.90**	**4.90**	4.90

highly optimized supervised models, its key advantage lies in its applicability to scenarios where training data collection is impractical or infeasible. This is particularly relevant for acoustic inspection tasks, where target sounds (e.g., anomalies) can be diverse, rare, or unpredictable, making the creation of comprehensive supervised training sets extremely challenging. Therefore, our method offers a practical and versatile solution for noise reduction in real-world conditions where data scarcity is common.

B. Real-world Acoustic Inspection

1) Setup: To verify whether the proposed denoising is effective as a preprocessing step in acoustic inspection, an experiment was conducted on the hammering test, one of the most critical acoustic inspection tasks. In this experiment, assuming a scenario with multiple noise sources as found in an inspection site, three sound sources in total (including the target source) were arranged as shown in Fig. 3. In addition, by actually generating sound, the aim was to validate the method in a nonlinear real-world reverberation environment that could not be fully represented by RIR alone.

A concrete specimen was used in which hammering sounds served as the target signal, while six types of noises, commonly present at hammer test inspection sites, were introduced as interference. The noise comprised two categories: sounds produced by actual sources (motor, fan, machinery) and those playback sources (conversation, flowing river, traffic noise). For recording, we employed a four-channel microphone array (consisting of four Shure SM11 microphones) along with a soundboard (Roland Rubix44) to capture each sound source separately.

The distance between the sound sources and the microphones was approximately 0.6 meter, with each sound source arranged at 90 degrees intervals. The audio classes were set in natural language as seven categories: [continuous tapping, motor, fan, engine, conversation, flowing river, traffic noise]. The entire defective and healthy areas on the concrete surface were struck uniformly to avoid positional bias. As shown in the Fig. 4, the defects collected in this study were simulated to include void and delamination, and samples from both healthy and defective areas were obtained.

In our experiment we focused on delamination and voids. These defects are among the most frequently encountered in the field and have been extensively studied in previous research [43], [44]. Visible surface anomalies such as cracks can

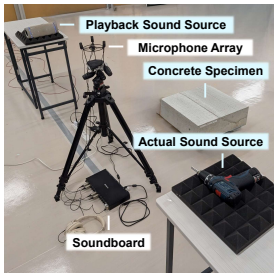

Fig. 3. Experimental setup. Sound sources are indicated in light blue caption.

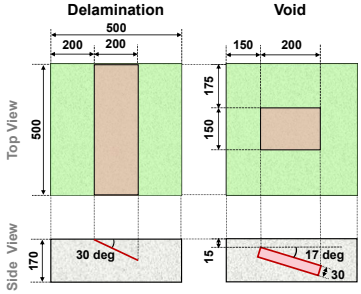

Fig. 4. Details of concrete specimen with simulated defects. Green areas represent healthy regions, while red areas indicate defective regions.

TABLE IV
AVERAGE SI-SNR IMPROVEMENT FOR EACH NOISE SOURCES.

Playback Source	Actual Source	SI-SNR Imp. [dB]
conversation	fan	4.60
conversation	machinery	5.46
conversation	motor	5.40
flowing river	fan	5.15
flowing river	machinery	5.87
flowing river	motor	5.31
traffic noise	fan	4.93
traffic noise	machinery	5.43
traffic noise	motor	4.54

be effectively detected through visual inspection or camera-based methods, but internal defects like delamination and voids require specialized techniques such as hammer tests for identification.

2) Results: The SI-SNR improvement for each noise source is shown in Table IV. In real-world environments, we confirmed that a stable noise reduction performance of approximately 5 dB can be achieved even when multiple noise sources are present.

To verify the effectiveness of the proposed method as a pre-processing step in acoustic inspection, we compared the defect discrimination accuracy for hammering sounds. Following the approach of previous studies [43], [45], features were extracted from hammer impact sounds by cropping each signal, normalizing its energy, and applying a Short-Time Fourier Transform. In the initial experiment, a linear Support Vector Machine (SVM) was used for defect discrimination of the hammering sounds. The linear SVM is often utilized in the literature to assess the intrinsic discriminative power and linear separability of features themselves, allowing for a direct comparison of feature quality [46], [47].

As shown in Table V, it is evident that the proposed method contributes to the improvement of defect discrimination accuracy. While this indicates that the proposed denoising clarifies the essential characteristics of the target signal, resulting in better features, the discrimination accuracy did not reach 100%. This observation warrants further discussion. Potential reasons include the possibility that the simple linear classifier may not be sufficient to fully capture the complex decision

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE V
AVERAGE ACCURACY OF DEFECT DISCRIMINATION ON TEST DATA.

		Number of Training Data					
Classifier	**Method**	256	512	1024	2048	4096	12288
SVM	w/ Prop.	**0.92**	**0.94**	0.95	0.96	0.97	0.98
SVM	w/o Prop.	0.89	0.92	0.93	0.94	0.95	0.96
GBDT	w/ Prop.	0.89	0.93	**0.96**	**0.98**	**0.99**	**1.00**
GBDT	w/o Prop.	0.87	0.91	0.94	0.97	0.98	0.99

boundary between healthy and defective samples, even with the improved features, or that characteristics necessary for discrimination might have been unintentionally compromised during the denoising process.

To disentangle these possibilities and evaluate the potential of the enhanced features more comprehensively, we conducted an additional experiment employing a Gradient Boosting Decision Tree (GBDT) [48] classifier. Under the identical 5-fold cross-validation protocol, the GBDT classifier achieved 100% accuracy. This subsequent result strongly supports the conclusion that our proposed denoising method effectively enhances the salient acoustic features necessary for accurate defect detection, confirming its potential to significantly contribute to the accuracy of acoustic inspection systems.

V. DISCUSSION

The experimental results on public datasets and real-world hammering tests clearly demonstrate the effectiveness of Zero-Shot Denoiser. Our method, operating in a truly zero-shot setting, achieved performance comparable to state-of-the-art supervised noise reduction methods. This overcomes the limitations of conventional noise reduction techniques that rely on target sound's location or large amounts of labeled data.

Despite these promising results, Zero-Shot Denoiser exhibits some limitations. The granularity of concepts that can be separated depends on the scope of latent knowledge acquired by the Audio-Language Model (ALM) during its pre-training. For example, while it can distinguish between general concepts like the sound of an "engine" and a "motor," individually separating sounds that lack common, descriptive textual labels, such as distinguishing motors solely by specific model numbers (e.g., "Motor XYZ7" vs. "Motor ABC3"), is challenging if these distinctions were not present or learnable in the ALM's training data. One potential direction to address finer granularity within the zero-shot paradigm is to incorporate detailed descriptive heuristics using natural language, as demonstrated in [49].

Alternatively, for scenarios where the target sound is highly specific and cannot be reliably identified using general text prompts, the proposed framework can be readily extended to a few-shot supervised separation approach. This involves collecting a small set of audio examples of the specific target sound beforehand. Instead of using a text query, the pre-trained audio encoder of the ALM is used to compute an average embedding directly from these target audio samples. This audio embedding then serves as the query vector for the AR-Attention mechanism, guiding the reconstruction of the target sound from the BSS-decomposed signals.

Another potential limitation of our approach concerns challenging acoustic environments, like the reverberant conditions found in tunnel hammering tests. While our chosen BSS algorithm (ILRMA) relies primarily on statistical independence and source characteristics over purely spatial cues, its performance can degrade under heavy reverberation or when noise sources are statistically similar and co-directional with the target. Specifically, reverberations are time-delayed and spectrally altered copies of the original hammering sound, meaning they often possess highly similar statistical characteristics and spectral structures to the direct sound itself. This inherent similarity between the target signal and its own echoes directly challenges the statistical independence assumption fundamental to many BSS algorithms, including ILRMA, making effective separation significantly more difficult. Consequently, the overall efficacy of the Zero-Shot Denoiser in such conditions fundamentally depends on the BSS module's capacity to yield adequately separated source components. A potential avenue for future work, for example, is to include BSS algorithms tailored for reverberant environments [50] within the search space explored by the automatic BSS optimization.

Furthermore, the proposed method is computationally expensive, making real-time processing challenging. In our experiments on a machine equipped with an Intel Core i7 13th generation CPU, the method required approximately 10 times the audio playback duration in computation time. In particular, the Bayesian optimization-based BSS hyperparameter search is challenging due to the need for repeated iterative optimization of demixing filters. To achieve faster processing, it is necessary to use BSS algorithms that support real-time [51] or online processing [52].

VI. CONCLUSION

This paper introduces Zero-Shot Denoiser, a novel noise reduction method for acoustic inspection that overcomes the limitations of conventional techniques by requiring neither prior knowledge of target sounds nor extensive training data. By synergistically integrating Blind Signal Separation (BSS) for unsupervised audio decomposition and Artifact-Resilient Attention (AR-Attention) for text-guided audio reconstruction using a pre-trained audio-language model, our method achieves robust noise reduction in a truly zero-shot manner. Furthermore, the proposed pseudo-SNR metric enables automatic optimization of BSS hyperparameters, significantly reducing the need for manual tuning.

In experiments using public datasets, our method, operating in a true zero-shot setting, achieved performance comparable to that of state-of-the-art supervised denoising methods, and experiments targeting hammering tests confirmed the effectiveness of our approach for real-world acoustic inspections. The proposed Zero-Shot Denoiser holds significant promise for advancing acoustic inspection and broader acoustic measurement tasks in noisy real-world environments.

VII. *ACKNOWLEDGEMENT

This work was supported in part by the SuzukiFoundation. This work was also supported in part by the World-leading Innovative Graduate Study Program in Proactive Environmental Studies (WINGS-PES), The University of Tokyo.

REFERENCES

[1] J. Liu, G. Xie, J. Wang, S. Li, C. Wang, F. Zheng, and Y. Jin, "Deep industrial image anomaly detection: A survey," *Machine Intelligence Research*, vol. 21, no. 1, pp. 104–135, 2024.

[2] J.Y. Louhi Kasahara, H. Fujii, A. Yamashita, and H. Asama, "Fuzzy clustering of spatially relevant acoustic data for defect detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2616–2623, 2018.

[3] H. Fujita, J.Y. Louhi Kasahara, S. Kanda, K. Nagatani, S. Kasahara, S. Fukumoto, S. Tamura, T. Kato, M. Korenaga, A. Sasamura, M. Hoshi, H. Asama, and A. Yamashita, "Acoustic monitoring in industrial plants with autoencoders and a mobile robot," in *Proceedings of the 20th International Conference on Ubiquitous Robots*, 2023, pp. 510–514.

[4] S. Shimizu, T. Igaue, J. Y. Louhi Kasahara, N. Yamato, S. Kasahara, H. Ito, T. Daito, S. Tamura, A. Sasamura, T. Kato, F. Nonaka, S. Kanda, K. Nagatani, H. Asama, Q. An, and A. Yamashita, "Change detection in image pairs for plant inspection using mobile robot," *International Journal of Automation Technology*, vol. 19, no. 4, 2025.

[5] K. Shoda, J. Y. Louhi Kasahara, H. Asama, Q. An, and A. Yamashita, "Defect detection with ego-noise reduction based on multimodal information in uav hammering inspection," *Advanced Robotics*, vol. 38, no. 17, pp. 1218–1230, 2024.

[6] S. Kuo, S. Mitra, and W.-S. Gan, "Active noise control system for headphone applications," *IEEE Transactions on Control Systems Technology*, vol. 14, no. 2, pp. 331–335, 2006.

[7] C. Quan and X. Li, "Spatialnet: Extensively learning spatial information for multichannel joint speech separation, denoising and dereverberation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 1310–1323, 2024.

[8] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 8135–8153, 2023.

[9] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1626–1641, 2016.

[10] J. Yin, Z. Liu, Y. Jin, D. Peng, and J. Kang, "Blind source separation and identification for speech signals," in *Proceedings of the 2017 International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC)*, 2017, pp. 398–402.

[11] L. Wang and A. Cavallaro, "A blind source separation framework for ego-noise reduction on multi-rotor drones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2523–2537, 2020.

[12] F. Zhi-guo, "Methods for estimation of the number of sources in blind source separation," *Journal of Hefei University of Technology*, 2008.

[13] J. Lu, W. Cheng, D. He, and Y. Zi, "A novel underdetermined blind source separation method with noise and unknown source number," *Journal of Sound and Vibration*, vol. 457, pp. 67–91, 2019.

[14] R. Wang and Y. Zhan, "A method of dynamic doa estimation with an unknown number of sources," in *Proceedings of the 2015 IEEE International Conference on Mechatronics and Automation*, 2015, pp. 104–109.

[15] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, "Multichannel extensions of non-negative matrix factorization with complex-valued data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 971–982, 2013.

[16] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 30, 2017.

[18] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending CLIP to image, text and audio," *Computing Research Repository*, vol. abs/2106.13043, 2021.

[19] S. Rouard, F. Massa, and A. Défossez, "Hybrid transformers for music source separation," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[20] B. Irvin, M. Stamenovic, M. Kegler, and L.-C. Yang, "Self-supervised learning for speech enhancement through synthesis," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[21] C. S. J. Doire and O. Okubadejo, "Interleaved multitask learning for audio source separation with independent databases," *Computing Research Repository*, vol. abs/1908.05182, 2019.

[22] R. Lu, Z. Duan, and C. Zhang, "Audio-visual deep clustering for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1697–1712, 2019.

[23] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[24] N. Makishima, S. Mogami, N. Takamune, D. Kitamura, H. Sumino, S. Takamichi, H. Saruwatari, and N. Ono, "Independent deeply learned matrix analysis for determined audio source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1601–1615, 2019.

[25] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," 2022. [Online]. Available: https://arxiv.org/abs/2212.00490

[26] Y. Mansour and R. Heckel, "Zero-shot noise2noise: Efficient image denoising without any data," in *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 018–14 027.

[27] X. L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 967–977, 2016.

[28] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 1085–1094, 2017.

[29] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, 2019.

[31] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 646–650.

[32] K. Yatabe, "Consistent ica: Determined bss meets spectrogram consistency," *IEEE Signal Processing Letters*, vol. 27, pp. 870–874, 2020.

[33] D. Kitamura and K. Yatabe, "Consistent independent low-rank matrix analysis for determined blind source separation," *Computing Research Repository*, vol. abs/2007.00274, 2020.

[34] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proceedings of the Advances in Neural Information Processing Systems*, vol. 24. Curran Associates, Inc., 2011.

[35] Y. Sun, P. Babu, and D. P. Palomar, "Majorization-minimization algorithms in signal processing, communications, and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 3, pp. 794–816, 2017.

[36] A. Zaeemzadeh, M. Joneidi, N. Rahnavard, and M. Shah, "Iterative projection and matching: Finding structure-preserving representatives and its application to computer vision," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5409–5418.

[37] P. Anastasopoulou, J. Torrey, X. Serra, and F. Font, "Heterogeneous sound classification with the broad sound taxonomy and dataset," 2024. [Online]. Available: https://arxiv.org/abs/2410.00980

[38] D. Di Carlo, P. Tandeitnik, C. Foy, N. Bertin, A. Deleforge, and S. Gannot, "dechorate: a calibrated room impulse response dataset for echo-aware signal processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2021, no. 1, pp. 1–15, 2021.

[39] C. Quan and X. Li, "Multi-channel narrow-band deep speech separation with full-band permutation invariant training," in *Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 541–545.

[40] Z. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Making time-frequency domain models great again for monaural speaker separation," in *Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[41] C. Quan and X. Li, "Multichannel speech separation with narrow-band conformer," in *Proceedings of the 23rd Annual Conference of the International Speech Communication Association*, 2022, pp. 5378–5382.

[42] A. Defazio, X. A. Yang, A. Khaled, K. Mishchenko, H. Mehta, and A. Cutkosky, "The road less scheduled," in *Proceedings of the Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[43] H. Fujii, A. Yamashita, and H. Asama, "Defect detection with estimation of material condition using ensemble learning for hammering test," in *Proceedings of the 2016 IEEE International Conference on Robotics and Automation*, 2016, pp. 3847–3854.

[44] J. Y. Louhi Kasahara, H. Fujii, A. Yamashita, and H. Asama, "Weakly supervised acoustic defect detection in concrete structures using clustering-based augmentation," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 6, pp. 2826–2834, 2021.

[45] K. Shoda, J. Y. Louhi Kasahara, H. Asama, Q. An, and A. Yamashita, "Clustering of hammering sounds and identification of defect clusters based on acoustic energy per impact for detection of concrete defects," in *Proceedings of the 24th International Conference on Control, Automation and Systems*, 2024, pp. 614–619.

[46] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 1597–1607.

[47] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1920–1929.

[48] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 3149–3157.

[49] H. Purohit, T. Nishida, K. Dohi, T. Endo, and Y. Kawaguchi, "Mimii-gen: Generative modeling approach for simulated evaluation of anomalous sound detection system," 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272968911

[50] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Relaxation of rank-1 spatial constraint in overdetermined blind source separation," in *Proceedings of the 23d European Signal Processing Conference*, 2015, pp. 1271–1275.

[51] R. Mukai, H. Sawada, S. Araki, and S. Makino, "Real-time blind source separation for moving speakers using blockwise ica and residual crosstalk subtraction," in *Proceedings of the 4th International Conference on Independent Component Analysis and Signal Separation*, 2003, pp. 975–980.

[52] T. Ueda, T. Nakatani, R. Ikeshita, K. Kinoshita, S. Araki, and S. Makino, "Low latency online blind source separation based on joint optimization with blind dereverberation," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 506–510.