BEAT2AASIST: BEATs Feature Splitting with Dual-Branch AASIST for Environmental Sound Deepfake Detection

Sanghyeok Chung[1], **Seungsang Oh**[1,*], **Donggun Kim**[1], **Jeongbin You**[1], **Il-Youp Kwak**[2,3,*], **Gaeun Heo**[2], **Eujin Kim**[2], **Nahyun Lee**[3], **Sunmook Choi**[4], **Soyul Han**[5]
[1]Department of Mathematics, Korea University, South Korea
[2]Department of Statistics and Data Science, Chung-Ang University, South Korea
[3]Department of Smart Cities, Chung-Ang University, South Korea
[4]Center for Applied Mathematics, Cornell University, USA
[5]Department of Big Data Application, Hannam University, South Korea
{cshzton, seungsang, dorage1105, dbwjdqls03}@korea.ac.kr, {ikwak2, hge403, rladbtls3, naa012}@cau.ac.kr, sc3377@cornell.edu, soyul5458@hnu.kr

Abstract

Recent advances in text-to-audio (TTA) and audio-to-audio (ATA) generation models have enabled the creation of highly realistic environmental sounds, raising growing concerns about malicious audio manipulation in real-world scenarios. To address this emerging threat, the ESDD 2026 Challenge was introduced as the first large-scale benchmark for Environmental Sound Deepfake Detection (ESDD), featuring two tracks that evaluate generalization to unseen generators and robustness under black-box, low-resource conditions. In this paper, we present BEAT2AASIST, an enhanced deepfake detection framework built upon the BEATs-AASIST baseline. Motivated by the observation that token-based audio representations may weaken explicit preservation of structured acoustic cues, the proposed method introduces a dual-branch AASIST architecture that explicitly splits BEATs-derived representations along frequency or channel dimensions. This design enables specialized modeling of complementary spoofing artifacts that may be attenuated in unified representations. To further enrich acoustic features, we incorporate multi-layer fusion strategies that aggregate information from multiple transformer layers using concatenation, CNN-gated, and SE-gated mechanisms. In addition, vocoder-based data augmentation with multiple high-fidelity neural vocoders is employed to enhance robustness against unseen and black-box spoofing attacks. Experimental results on the EnvSDD dataset demonstrate that BEAT2AASIST achieves strong and consistent performance across both challenge tracks. In particular, the proposed approach attains 3rd place in Track 2 and 4th place in Track 1 in the ESDD 2026 Challenge, despite using fewer ensemble components than top-ranked systems. These results suggest that explicit modeling of heterogeneous acoustic subspaces, combined with targeted representation fusion and data augmentation, provides an effective and efficient design strategy for real-world environmental sound deepfake detection. The code is available at https://github.com/ikwak2/BEAT2AASIST.

1 Introduction

Recent progress in audio generation technologies has enabled the synthesis of highly realistic soundscapes extending beyond speech to a wide range of environmental sounds. Modern text-to-audio (TTA) and audio-to-audio (ATA) generative models are now capable of producing immersive and convincing audio content, which has accelerated applications in film production, gaming, and virtual or augmented reality. At the same time, these advances raise serious security concerns. Fabricated environmental sounds can be maliciously injected into recordings or disseminated as misleading evidence, undermining the credibility of audio-based information. These emerging threats have intensified the demand for reliable Environmental Sound Deepfake Detection (ESDD) systems.

While deepfake detection has been extensively studied in the speech domain, ESDD remains relatively underexplored. In speech-based spoofing detection, substantial progress has been achieved through the adoption of large-scale self-supervised learning (SSL) foundation models, particularly when combined with multi-layer fusion of transformer features and vocoder-based data augmentation [Han *et al.*, 2023; Martín-Doñas *et al.*, 2024; Wu *et al.*, 2024; Shin *et al.*, 2024; Wang and Yamagishi, 2023; Wang and Yamagishi, 2024]. Models such as wav2vec2 [Baevski *et al.*, 2020], HuBERT [Hsu *et al.*, 2021], and WavLM [Chen *et al.*, 2022] have become dominant front-end representations. These approaches have consistently delivered state-of-the-art results in major challenges such as ADD 2023 [Yi *et al.*, 2023] and ASVspoof5 [Wang *et al.*, 2026], demonstrating the effectiveness of SSL pre-training for robust spoofing detection.

In contrast, ESDD poses additional challenges due to the

*Corresponding authors.

significantly higher variability of acoustic conditions and the diversity of environmental sound categories. Environmental audio encompasses a broad range of non-linguistic events with less structured temporal patterns than speech, making generalization to unseen conditions substantially more difficult. To address these challenges, recent studies have explored SSL foundation models tailored for environmental audio, most notably BEATs [Chen *et al.*, 2023] and EAT [Chen *et al.*, 2024]. Among them, BEATs pre-trained on the large-scale AudioSet-2M corpus has shown strong generalization across diverse sound events and has emerged as a particularly effective representation for environmental sound analysis.

This research direction is formalized by the Environmental Sound Deepfake Detection Challenge (ESDD 2026) [Yin *et al.*, 2026], which introduces the first large-scale benchmark dedicated to ESDD based on the EnvSDD dataset. EnvSDD consists of 45.25 hours of real audio and 316.7 hours of fake audio, covering a wide range of environmental sound types. The challenge is structured into two tracks reflecting real-life scenarios. Track 1 evaluates generalization to unseen TTA / ATA generators, while Track 2 simulates a more challenging black-box setting with extreme uncertainty and limited data, where only 1% of black-box training data is available. The official baseline, BEATs-AASIST, highlights the importance of combining environmental-sound-centric SSL representations with robust anti-spoofing back-end architectures.

In this paper, we propose BEAT2AASIST, an enhanced framework that extends the BEATs-AASIST baseline to better capture the heterogeneous characteristics of environmental sound deepfakes. Unlike the baseline, which treats the BEATs output as a flattened sequence of tokens, inherently obscuring the explicit spectrotemporal structure, our approach introduces a dual branch AASIST architecture. By explicitly splitting the representations along either the frequency or channel dimension, we recover and independently model these structural components. This design enables specialized modeling of distinct acoustic artifacts, such as frequency dependent or channel specific spoofing patterns, before aggregating the branch outputs for final prediction.

To further enhance representation quality, we introduce three multi-layer fusion strategies that aggregate information from the top-k transformer layers of BEATs. By leveraging intermediate representations through concatenation, CNN-gated, and SE-gated fusion mechanisms, the model captures multi-scale acoustic cues beyond the final-layer embedding. In addition, we apply vocoder-based data augmentation using several high-quality, publicly available GAN-based vocoders to expose the model to diverse synthesis artifacts, improving robustness to both unseen-generator and black-box scenarios.

We evaluate our approach on the EnvSDD dataset under both challenge tracks. Our system achieves competitive performance, ranking 3rd in Track 2 with EER 0.35% and 4th in Track 1 with EER 1.60% in the ESDD 2026 Challenge, demonstrating strong generalization to unseen generators as well as robustness under black-box low-resource conditions. These results validate the effectiveness of our architectural and training strategies for ESDD.

2 Backgrounds

2.1 Foundation Models for Spoofing Detection

With increasingly realistic spoofing speech from advanced text-to-speech (TTS) and voice conversion (VC) systems, spoofing speech detection (SSD) has become a key task in speech processing. Recent SSD approaches have shifted from handcrafted features to end-to-end models operating on raw waveforms (e.g., RawNet2 [Tak *et al.*, 2021a]) and, more prominently, to SSL foundation models that provide strong transferable representations. Because spoofing artifacts can be subtle and localized, attention mechanisms and graph-based methods (e.g., RawGAT-ST [Tak *et al.*, 2021b], AASIST) are often employed to capture complex spectro-temporal dependencies. Thus, hybrid systems that combine SSL front-ends with attentive/graph back-ends have become an effective design for robust spoofing detection.

2.2 BEATs-AASIST

BEATs (Bidirectional Encoder representation from Audio Transformers) [Chen *et al.*, 2023] is a general audio foundation model designed to learn semantic representations from large-scale unlabeled audio. It introduces an iterative pre-training framework that alternates between an acoustic tokenizer and a Transformer-based audio SSL model, formulating masked prediction over discrete acoustic tokens rather than continuous reconstruction. Through mutual refinement via knowledge distillation, the tokenizer and SSL model progressively capture high-level semantics and discard the redundant details for audio. BEATs employs a Transformer encoder to model temporal and spectral context and is compatible with various masked audio modeling backbones. Trained on large-scale audio corpora such as AudioSet-2M, it achieves strong performance across diverse audio understanding tasks. Owing to its robust transfer learning capability, BEATs serves as an effective front-end for ESDD, where modeling heterogeneous acoustic patterns and background conditions is critical.

AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks) [Jung *et al.*, 2022] is an end-to-end deep learning framework specifically developed for voice spoofing detection. AASIST models audio representations as spectro-temporal graphs and employs graph attention mechanisms to capture discriminative features across frequency and time dimensions. By jointly learning temporal and spectral relationships, the model effectively identifies subtle artifacts introduced by spoofing and synthesis processes. AASIST is typically initialized randomly and optimized through supervised training on spoofing detection tasks, enabling it to adapt its graph-based reasoning to task-specific spoofing characteristics.

BEATs-AASIST integration combines self-supervised acoustic representations with graph-based anti-spoofing reasoning. In this framework, the BEATs encoder serves as a robust feature extractor that provides high-level, environment-aware embeddings, while the AASIST back-end operates as a discriminative classifier that models spectro-temporal relationships within these embeddings. This hybrid architecture has been adopted as the official baseline in the ESDD 2026

Challenge and has demonstrated strong generalization to diverse and unseen environmental sound manipulations.

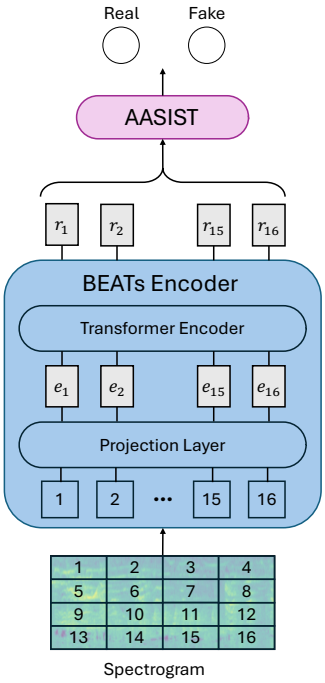

Figure 1: BEATs-AASIST architecture. Note that the explicit spectrotemporal structure is lost when the patch embeddings are flattened into a 1D token sequence $(r_1, \ldots, r_{16})$.

2.3 Multi-layer Fusion of Transformer Layers

Recent studies have shown that using only the final layer of a large pre-trained Transformer is often suboptimal for downstream tasks. Because Transformer encoders learn hierarchical representations from low-level acoustic patterns in lower layers to higher-level semantic information in higher layers, aggregating features from multiple layers generally yields richer and more robust representations than relying on a single layer.

This insight is supported by the SUPERB benchmark for speech, which employs the learnable weighted-sum of the features from all transformer layers for downstream evaluation, where the layer-weights are task-specific [Yang *et al.*, 2024]. Similarly, WavLM adopts layer-wise aggregation in speaker recognition, demonstrating improved robustness to acoustic variability [Chen *et al.*, 2022]. Multi-layer fusion following the SUPERB setup has also proven effective in audio deepfake detection [Martín-Doñas *et al.*, 2024].

Motivated by these findings, our work adopts a top-k layer fusion strategy to exploit complementary information across different transformer layers, offering stronger generalization and robustness than single-layer representations.

2.4 Neural Vocoders

Neural vocoders have emerged as an efficient and scalable source of synthetic audio for data augmentation in spoofing and deepfake detection. Since constructing comprehensive spoofing datasets is challenging due to diverse real-world attack pipelines, prior studies show that effective countermeasures can be trained using vocoder-generated audio alone, without relying on full TTS or voice conversion systems [Wang and Yamagishi, 2023; Wang and Yamagishi, 2024].

HiFi-GAN [Kong *et al.*, 2020] is a GAN-based neural vocoder that achieves computationally efficient and high-fidelity speech synthesis by effectively modeling periodic patterns in audio signals. Its strong generalization to unseen speakers and noisy conditions makes it particularly suitable for large-scale spoofed audio generation. Recent GAN-based vocoders such as UnivNet [Jang *et al.*, 2021] and BigVGAN [gil Lee *et al.*, 2023] further advance high-fidelity audio synthesis and generalization. UnivNet employs a multi-resolution spectrogram discriminator leveraging spectral representations at different scales, enabling robust waveform generation for both seen and unseen speakers. BigVGAN focuses on universal generalization by introducing a periodic activation function and anti-aliased representation, allowing stable synthesis across varied environments without fine-tuning.

Overall, neural vocoders provide a practical means of generating diverse spoofing examples, supporting the development of robust and generalizable deepfake detection systems when real attack data is limited.

3 Methods

Our proposed framework BEAT2AASIST extends the official BEATs-AASIST baseline by introducing targeted architectural and training-level modifications that improve the modeling of diverse acoustic cues inherent in environmental sound deepfake detection.

In the baseline pipeline, the self-supervised BEATs encoder takes a mel-spectrogram as input and partitions it into frequency-time patches of size 16×16. Each patch is embedded into a 768-dimensional token, producing a flattened feature sequence of shape $(H \times W, 768)$, where H and W correspond to the number of patches along the frequency and time axes, respectively. However, this flattening process inherently obscures the explicit spectrotemporal structure of the input. Consequently, the resulting representations are directly forwarded to the AASIST back-end without explicit spatial or channel distinction, potentially limiting the modeling of local frequency dependencies.

Building upon this baseline, our framework introduces three complementary improvements:
(1) a dual branch AASIST architecture that reorganizes the flattened BEATs representations to recover frequency or channel wise structural information and processes them with parallel AASIST branches;
(2) multi-layer fusion strategies that aggregate information from the top-k transformer layers instead of relying solely on the final layer in the BEATs; and
(3) vocoder-based data augmentation to improve robustness against diverse and previously unseen spoofing attacks during training.

3.1 BEAT2AASIST Architecture

The proposed BEAT2AASIST extends BEATs-AASIST by splitting BEATs-derived representations along frequency or channel dimension and processing them with a dual-branch AASIST back-end as illustrated in Fig. 2. Given the BEATs output feature sequence of shape $(H \times W, 768)$, the proposed architecture divides the representation into two disjoint subsets, each of which is processed by an independent AASIST network. The outputs from the two AASIST branches are then concatenated to produce the final prediction. Unlike the baseline BEATs-AASIST pipeline that feeds a single unified representation into one classifier, our design explicitly separates the feature space and enables specialized modeling of distinct spoofing characteristics. We investigate two complementary splitting strategies that target different dimensions of the BEATs representation.

Frequency-based Feature Split

For frequency-aware modeling, the BEATs output tensor is first reshaped into $(H, W, 768)$ and evenly divided along the frequency axis, yielding low- and high-frequency feature subsets. Each subset is then flattened back to $(H/2 \times W, 768)$ and processed by a separate AASIST branch. By explicitly decoupling low- and high-frequency components, this strategy allows each branch to focus on distinct spectral characteristics, improving sensitivity to frequency-dependent spoofing artifacts that often arise from synthesis operations.

Channel-based Feature Split

Alternatively, the BEATs representation is partitioned along the channel dimension into two tensors of size $(H \times W, 384)$, which are independently fed into parallel AASIST networks. This channel-wise decomposition encourages each branch to specialize in different subspaces of the learned BEATs embeddings, enabling the model to capture complementary channel-specific cues embedded in the high-dimensional representation. Such specialization is particularly beneficial for modeling subtle acoustic inconsistencies that may not be uniformly distributed across embedding channels.

3.2 Multi-layer Fusion

Transformer-based encoders capture complementary acoustic information across different layers, ranging from low-level spectral patterns to higher-level semantic cues. Relying only on the final layer may therefore limit the expressive capacity of the learned representation. To address this limitation, we adopt a multi-layer fusion strategy that aggregates the representations, say $\boldsymbol{h}_t^{13-k}, \ldots, \boldsymbol{h}_t^{12} \in \mathbb{R}^{768}$ for each token t with $1 \leq t \leq H \times W$, from the top-k layers among the 12 transformer layers of BEATs. Three different fusion strategies are illustrated in Fig. 3.

Concatenation Fusion

In the simplest fusion scheme, the outputs of the selected top-k layers, each with shape $(H \times W, 768)$, are stacked along a newly introduced layer dimension to form a tensor of shape $(H \times W, k, 768)$. This fused representation is directly forwarded to the downstream deepfake detection network without additional layer-weighting. Despite its simplicity, this approach preserves all layer-wise information from the selected layers and allows the classifier to implicitly learn inter-layer relationships.

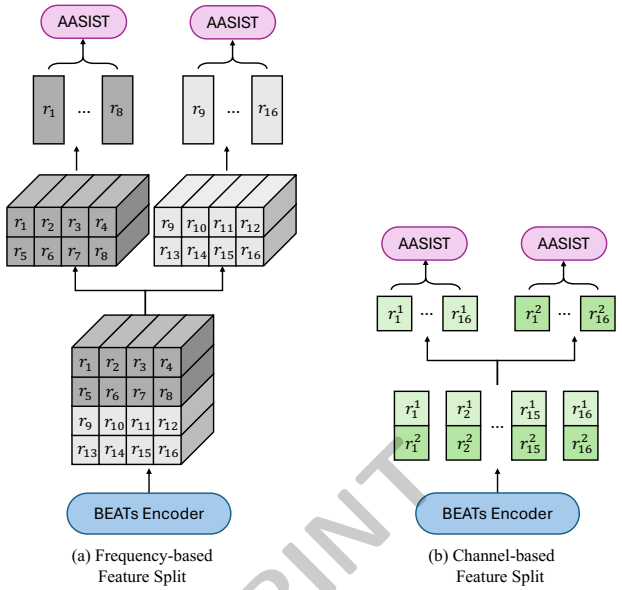

Figure 2: The proposed BEAT2AASIST architecture. Unlike the flattened baseline, our dual-branch design explicitly splits the representation along frequency or channel dimensions to recover and independently model the distinct structural information, enabling the detection of subtle, heterogeneous spoofing artifacts.

CNN-gated Fusion

To adaptively weight transformer layers based on input characteristics, we design a convolutional gating network conditioned on the input mel-spectrogram. The spectrogram is processed by a lightweight CNN composed of three 2D convolution layers, each followed by batch normalization, non-linear activation, and max pooling. The resulting feature map is then passed through global average pooling and a linear layer with softmax activation to produce a normalized k layer-weights, say $\alpha^{13-k}, \ldots, \alpha^{12}$, which are applied to the corresponding transformer layer outputs to compute a weighted-sum,

$$\tilde{\boldsymbol{h}}_t = \textstyle\sum_{l=13-k}^{12} \alpha^l \cdot \boldsymbol{h}_t^l,$$

thereby enabling the model to emphasize transformer layers that are more informative for a given acoustic input. The unified feature representation $\tilde{\boldsymbol{h}}_t$ is subsequently split along the frequency or channel dimension and processed by the dual-branch AASIST back-end.

SE-gated Fusion

Unlike the CNN-gated approach conditioned on the input spectrogram, the SE-gated fusion strategy derives layer-weights directly from the transformer layer outputs themselves. Specifically, the top-k layer representations are first aggregated and then passed through a Squeeze-and-Excitation (SE) module [Hu *et al.*, 2018], which captures global dependencies across the selected layers. The resulting layer-wise attention weights $\alpha^{13-k}, \ldots, \alpha^{12}$ are then used to combine the layer representations into a unified embedding,

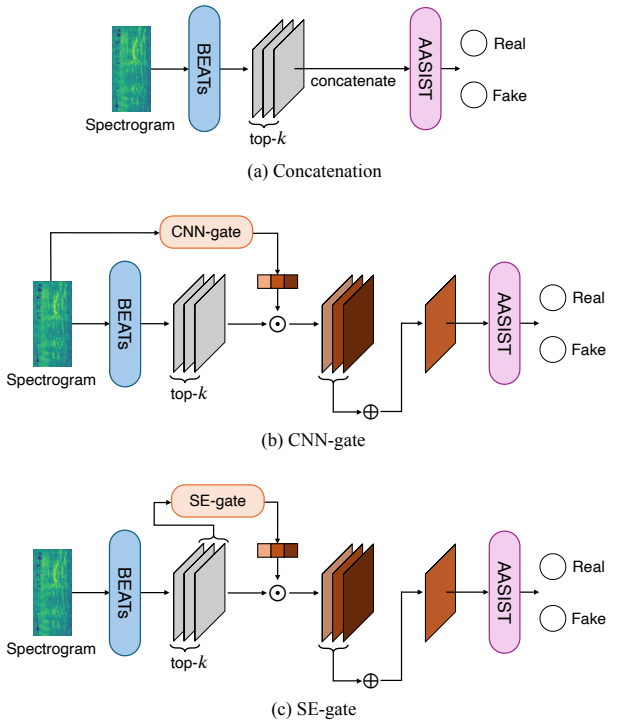

Figure 3: Multi-layer fusion strategies. By aggregating the top-k transformer layers using (a) Concatenation, (b) CNN-gate, and (c) SE-gate, we capture diverse hierarchical acoustic cues beyond the final-layer embedding.

enabling coherent acoustic patterns while selectively emphasizing the most relevant layers for deepfake detection.

3.3 Vocoder-based Data Augmentation

To enhance robustness against diverse spoofing conditions defined in the ESDD 2026 Challenge, we adopt vocoder-based data augmentation during training. Real-world environmental sound deepfakes can be generated by a wide range of unseen TTA / ATA generators, making it difficult to construct sufficiently diverse training data from limited generators alone. Neural vocoders provide an efficient way to simulate such variability by introducing distinct synthesis artifacts while preserving the semantic content of the original audio.

In accordance with the challenge design, we employ different augmentation strategies for the two tracks. For Track 1 (ESDD in Unseen Generators), which focuses on generalization to unseen TTA / ATA models, we use HiFi-GAN to generate spoofed audio via copy-synthesis. HiFi-GAN produces high-quality waveforms with realistic periodic structures, allowing the model to learn general spoofing cues without overfitting to specific generators.

For the more challenging Track 2 (Black-Box Low-Resource ESDD), which simulates extreme uncertainty and limited training data, we further expand augmentation by incorporating UnivNet and BigVGAN in addition to HiFi-GAN. These vocoders exhibit complementary generalization behaviors across acoustic conditions and recording environments, thereby introducing a broader spectrum of synthesis artifacts. This diversity is particularly important under black-box settings, where the test-time generation process is entirely unknown.

By leveraging multiple vocoders with different inductive biases, our augmentation strategy encourages the model to focus on intrinsic spoofing characteristics rather than vocoder-specific artifacts, leading to improved generalization across both unseen-generator and black-box scenarios.

4 Experiments

4.1 Dataset

The EnvSDD dataset [Yin *et al.*, 2026] is the first large-scale dataset designed for ESDD, consisting of 45.25 hours of real audio and 316.7 hours of fake audio. The training set for Track 1 contains 27,811 real audio samples and 111,244 fake samples, where the fake audio is synthesized using a fixed set of known TTA/ATA generators. The Track 1 test set includes 1,000 real and 3,000 fake samples, with fake audio generated by previously unseen models, thereby explicitly evaluating generalization to unseen generators [Yin *et al.*, 2026]. The training set for Track 2 includes all samples from the Track 1 training set, along with an additional 270 real audio samples and 1,083 fake samples. The generation process of these additional fake samples is undisclosed, and thus they are referred to as black-box data. The Track 2 test set consists of 1,994 real and 7,980 fake audio samples, all of which are generated under black-box conditions, making Track 2 representative of real-world deployment scenarios. Table 1 summarizes the detailed statistics of the EnvSDD dataset across different usage splits and tracks, highlighting the significant imbalance between real and fake samples as well as the increased scale and uncertainty introduced in Track 2. This design allows systematic evaluation of both unseen-generator generalization (Track 1) and robustness to unknown synthesis pipelines (Track 2). All audio samples are resampled to 16 kHz and have a duration of 4 seconds. For model training, we used only the training datasets of Track 1 and Track 2, respectively.

Usage	Track 1		Track 2	
	Real	Fake	Real	Fake
Training	27,811	111,244	270	1,083
Validation	7,942	31,768	90	361
Evaluation	478	1,522	973	4,014
Test	1,000	3,000	1,994	7,980

Table 1: Statistics of the EnvSDD dataset.

4.2 Experimental Setup

For training the BEAT2AASIST model, we employed the BEATs-iter3 version, which was pre-trained on the AudioSet dataset, and fine-tuned all parameters. Input audio clips were either trimmed or padded through repetition to a fixed length of 4 seconds. Each audio sample was converted into a mel-spectrogram with a frame length of 25 ms, a frame shift of

Model	Model description for Track 1	Vocoder	Eval EER	Test EER
#1	BEATs-AASIST	Hifi-GAN	1.69%	1.88%
#2	BEAT2AASIST, Frequency split, SE-gate ($k=4$)	Hifi-GAN	1.69%	1.92%
#3	BEAT2AASIST, Channel split, SE-gate ($k=10$)	Hifi-GAN	1.66%	1.70%
#4	**Ensemble (0.4 × #1 + 0.6 × #3)**		N/A	**1.60%**
#5	Ensemble (0.3 × #2 + 0.7 × #3)		1.66%	1.62%

Model	Model description for Track 2	Vocoder	Eval EER	Test EER
#1	BEATs-AASIST, Concatenation ($k=4$)	N/A	0.61%	0.75%
#2	BEAT2AASIST, Frequency split, CNN-gate ($k=4$)	Hifi-GAN, UnivNet, BigVGAN	0.42%	0.46%
#3	BEAT2AASIST, Frequency split, CNN-gate ($k=10$)	Hifi-GAN, UnivNet, BigVGAN	N/A	0.60%
#4	BEAT2AASIST, Frequency split, SE-gate ($k=4$)	Hifi-GAN, UnivNet, BigVGAN	0.82%	0.76%
#5	Ensemble (0.4 × #1 + 0.6 × #2)		0.30%	0.40%
#6	**Ensemble (0.2 × #1 + 0.5 × #2 + 0.3 × #3)**		N/A	**0.35%**
#7	Ensemble (0.5 × #2 + 0.4 × #3 + 0.1 × #4)		N/A	0.41%

Table 2: BEAT2AASIST performance on Tracks 1 and 2 in the ESDD 2026 Challenge

10 ms, and 128 mel bins. For data augmentation, we applied SpecAug [Park *et al.*, 2019] and additionally generated fake audio using vocoders. As vocoders, we used HiFi-GAN, UnivNet, and BigVGAN. All vocoder-generated fake audio was resampled to 16 kHz before being used. We trained the model using the Adam optimizer with a learning rate of 10^{-6} and weight decay of 10^{-4}. The batch size was set to 32, and training was conducted for 20 epochs. Because the dataset contains significantly more fake audio than real audio, we applied class weighting to the cross-entropy loss, setting the weights to (fake, real) = (0.1, 0.9).

4.3 Experimental Results

We evaluated our models on the official test sets of both tracks using equal error rate (EER) as the evaluation metric. The Eval EER corresponds to performance on the progress-phase data, whereas the Test EER reflects the final ranking-phase results. Table 2 summarizes the Track 1 performance, where Systems 1-3 are individual models and Systems 4-5 are ensemble variants constructed from them. Table 2 also presents the Track 2 results, with Systems 1-4 denoting individual models and Systems 5-7 representing their corresponding ensembles. The specific weights for the ensemble models were empirically tuned based on the individual model accuracy observed during the challenge participation. Detailed model configurations are provided in the table descriptions, and the vocoders used for generating synthetic spoofed audio are noted where applicable.

The evaluation results demonstrate the competitiveness of our proposed approach. In particular, the model exhibits strong performance on Track 2, which focuses on detecting black-box fake audio.

4.4 Ablation Study

To investigate how much each component of the proposed BEAT2AASIST contributes to performance improvement, we conducted additional ablation experiments. Using the single BEAT2AASIST models employed in Track 1 and Track 2, we evaluated the performance degradation caused by individually removing the feature split architecture, multi-layer fusion, and vocoder augmentation components. All additional experiments were conducted under the same experimental settings, except for the use of a smaller learning rate 7×10^{-7} for more fine-grained training. This adjustment of the learning rate led to slight performance improvements in several models. The results of the ablation experiments are summarized in Table 3.

Across most single BEAT2AASIST models, the largest performance degradation was observed when vocoder augmentation was removed. This effect is particularly pronounced in Track 2, indicating that vocoder augmentation plays a critical role in enhancing robustness under black-box conditions, where attack generators and synthesis processes are unknown at training time. By exposing the model to diverse vocoder-induced artifacts, this component substantially improves generalization to unseen spoofing strategies.

In contrast, removing the feature split architecture or the multi-layer fusion module consistently degrades performance across both Track 1 and Track 2, suggesting that these components provide track-agnostic structural benefits. The feature split architecture enables more effective discrimination of heterogeneous spoofing characteristics by introducing subspace specialization, while multi-layer fusion facilitates richer representation learning from the transformer-based encoder. Together, these results indicate that vocoder augmentation primarily addresses distributional uncertainty in black-box scenarios, whereas feature splitting and multi-layer fusion contribute complementary and robust gains across different evaluation conditions.

Model	Track 1				Track 2					
	(#2)		(#3)		(#2)		(#3)		(#4)	
	Eval	Test	Eval	Test	Eval	Test	Eval	Test	Eval	Test
Baseline	1.45	1.60	1.66	1.70	0.43	0.46	0.82	0.60	0.72	0.61
w/o Feature Split	2.10	2.22	2.47	2.24	1.04	0.95	1.13	1.05	1.34	1.15
w/o Multi-layer Fusion	1.69	1.90	2.30	2.30	1.02	0.80	1.02	0.80	1.02	0.80
w/o Vocoder Augmentation	2.54	2.60	2.30	2.10	1.83	1.75	1.87	1.91	2.06	1.96

Table 3: Ablation results on the impact of each component on generalization performance (EER, %).

4.5 Comparison with other Top-Performing Systems

As summarized in Table 4, top-performing systems in the ESDD 2026 Challenge are predominantly built upon strong environmental audio foundation models such as EAT, BEATs and SSLAM, often combined with advanced back-end architectures and ensemble inference. The highest-ranked approaches in both tracks employ large ensembles (up to five models), indicating that model diversity plays a key role in achieving peak performance. The SSLAM+FFN system [Guo *et al.*, 2025], which ranked 3rd in Track 1 and 4th in Track 2, demonstrates that competitive results can be obtained using a frozen SSL encoder with lightweight back-end modeling and intermediate-layer fusion. In contrast, BEAT2AASIST adopts a task-specific architectural design that explicitly models heterogeneous spoofing cues via feature splitting and multi-layer fusion. Despite using fewer ensemble components, our method achieves 4th place in Track 1 and 3rd place in Track 2, suggesting that architectural inductive biases can partially compensate for reduced ensemble size and offer an efficient alternative to large-scale ensemble-based systems.

Team Rank	System Description	Ensemble	Test EER
Track 1			
1st	EAT+AASIST	5	0.30
2nd	EAT_Large+BiCrossMamba	4	0.80
3rd	SSLAM+FFN	1	1.20
4th (**Ours**)	BEAT2AASIST	2	1.60
5th	EAT+ArcFace	1	2.40
Track 2			
1st	EAT_Large_SSLAM+BiCrossMamba	5	0.25
2nd	EAT+AASIST	5	0.25
3rd (**Ours**)	BEAT2AASIST	3	0.35
4th	SSLAM+FFN	1	1.05
5th	BEATs+KNN	1	2.96

Table 4: Leaderboard results on the ESDD 2026 Challenge test set (Test EER, %).

5 Conclusion

This paper introduced BEAT2AASIST, an enhanced framework for environmental sound deepfake detection that addresses the growing risks posed by increasingly realistic audio generation technologies. Beyond the specific architecture, our approach highlights the importance of explicitly modeling heterogeneous acoustic subspaces when applying token-based audio foundation models such as BEATs to spoofing detection tasks. By adopting a dual-branch AASIST design that splits BEATs representations along frequency or channel dimensions, the proposed method captures complementary spoofing cues that may be attenuated in unified representations. Multi-layer transformer fusion further enriches acoustic representations, while vocoder-based data augmentation improves robustness to unseen and black-box attacks.

Experiments on the EnvSDD dataset demonstrate the effectiveness of BEAT2AASIST, achieving competitive performance in the ESDD 2026 Challenge across both tracks. These results suggest that architectural inductive biases and targeted data augmentation can partially compensate for reliance on large-scale ensemble inference, offering a practical and efficient alternative for real-world ESDD scenarios.

Beyond the challenge performance, BEAT2AASIST is particularly relevant for real-world deployment, where environmental sounds exhibit high variability in source, context, and recording conditions, and where attack models are often unknown in advance. The modular design of our framework allows it to flexibly adapt to different operating conditions, making it suitable for applications such as forensic audio analysis, multimedia content verification, and security monitoring systems.

Several promising directions remain for future work. First, incorporating temporal or event-level localization could enable not only detection but also attribution of spoofed segments within long audio streams. Second, extending the framework to jointly leverage multimodal cues, such as synchronized visual or textual information, may further improve robustness against sophisticated attacks. Finally, exploring more adaptive or continual learning strategies could help the model remain effective as audio generation technologies continue to evolve.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00341075, RS-2026-25477127)

References

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc., 2020.

[Chen *et al.*, 2022] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[Chen *et al.*, 2023] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 5178–5193. PMLR, 23–29 Jul 2023.

[Chen *et al.*, 2024] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen. Eat: self-supervised pre-training with efficient audio transformer. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[gil Lee *et al.*, 2023] Sang gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. BigVGAN: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*, 2023.

[Guo *et al.*, 2025] Xiaoxuan Guo, Hengyan Huang, Jiayi Zhou, Renhe Sun, Jian Liu, Haonan Cheng, Long Ye, and Qin Zhang. Envsslam-ffn: Lightweight layer-fused system for esdd 2026 challenge, 2025.

[Han *et al.*, 2023] Soyul Han, Taein Kang, Sunmook Choi, Jaejin Seo, Sanghyeok Chung, Sumi Lee, Seungsang Oh, and Il-Youp Kwak. Cau ku deepfake detection system for add 2023 challenge. In *Proceedings of the IJCAI Workshop on Deepfake Audio Detection and Analysis (DADA)*, pages 23–30, 2023.

[Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[Jang *et al.*, 2021] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. In *Interspeech 2021*, pages 2207–2211, 2021.

[Jung *et al.*, 2022] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6367–6371, 2022.

[Kong *et al.*, 2020] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc., 2020.

[Martín-Doñas *et al.*, 2024] Juan M. Martín-Doñas, Aitor Álvarez, Eros Rosello, Angel M. Gomez, and Antonio M. Peinado. Exploring Self-supervised Embeddings and Synthetic Data Augmentation for Robust Audio Deepfake Detection. In *Interspeech 2024*, pages 2085–2089, 2024.

[Park *et al.*, 2019] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, page 2613–2617, September 2019.

[Shin *et al.*, 2024] Hyun-seo Shin, Jungwoo Heo, Ju-ho Kim, Chan-yeong Lim, Wonbin Kim, and Ha-Jin Yu. Hm-conformer: A conformer-based audio deepfake detection system with hierarchical pooling and multi-level classification token aggregation methods. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10581–10585, 2024.

[Tak *et al.*, 2021a] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6369–6373, 2021.

[Tak *et al.*, 2021b] Hemlata Tak, Jee weon Jung, Jose Patino, Madhu Kamble, Massimiliano Todisco, and Nicholas Evans. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 1–8, 2021.

[Wang and Yamagishi, 2023] Xin Wang and Junichi Yamagishi. Spoofed training data for speech spoofing countermeasure can be efficiently created using neural vocoders.

In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Wang and Yamagishi, 2024] Xin Wang and Junichi Yamagishi. Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end? In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10311–10315, 2024.

[Wang *et al.*, 2026] Xin Wang, Héctor Delgado, Hemlata Tak, Jee weon Jung, Hye jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, Nicholas Evans, Kong Aik Lee, Junichi Yamagishi, Myeonghun Jeong, Ge Zhu, Yongyi Zang, You Zhang, Soumi Maiti, Florian Lux, Nicolas Müller, Wangyou Zhang, Chengzhe Sun, Shuwei Hou, Siwei Lyu, Sébastien Le Maguer, Cheng Gong, Hanjie Guo, Liping Chen, and Vishwanath Singh. Asvspoof 5: Design, collection and validation of resources for spoofing, deepfake, and adversarial attack detection using crowdsourced speech. *Computer Speech & Language*, 95:101825, 2026.

[Wu *et al.*, 2024] Haochen Wu, Wu Guo, Zhentao Zhang, Wenting Zhao, Shengyu Peng, Jie Zhang, and China Merchants Bank. Spoofing Speech Detection by Modeling Local Spectro-Temporal and Long-term Dependency. In *Interspeech 2024*, pages 507–511, 2024.

[Yang *et al.*, 2024] Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T. Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, Tzu-hsun Feng, Po-Han Chi, Yist Y. Lin, Yung-Sung Chuang, Tzu-Hsien Huang, Wei-Cheng Tseng, Kushal Lakhotia, Shang-Wen Li, Abdelrahman Mohamed, Shinji Watanabe, and Hung-yi Lee. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2884–2899, 2024.

[Yi *et al.*, 2023] Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, Le Xu, Junzuo Zhou, Hao Gu, Zhengqi Wen, Shan Liang, Zheng Lian, Shuai Nie, and Haizhou Li. Add 2023: the second audio deepfake detection challenge. In *Proceedings of the IJCAI Workshop on Deepfake Audio Detection and Analysis (DADA)*, pages 125–130, 2023.

[Yin *et al.*, 2026] Han Yin, Yang Xiao, Rohan Kumar Das, Jisheng Bai, and Ting Dang. Environmental sound deepfake detection challenge: An overview. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 21772–21774, 2026.