Spatially-Guided Bi-Model CNN Architecture for Multi-Noise Localization and Soft-Mask Based Active Noise Cancellation in Indoor Environments

Guoqiang Lu[1*], Yanmin Bai[2], Hairong Wang[3]
[1]School of Internet of Things Engineering, Jiangsu Vocational College of Information Technology, Wuxi 214153, Jiangsu, China
[2]School of Microelectronics, Jiangsu Vocational College of Information Technology, Wuxi 214153, Jiangsu, China
[3]School of Education, Soochow University, Suzhou 215123, Jiangsu, China
E-mail: 2024101345@jsit.edu.cn, baiyanminnuaa@163.com, wanghairong0707@outlook.com
*Corresponding author

Keywords: multi-noise localization, active noise cancellation (anc), overlapping sound events, convolutional neural networks (cnn), mfcc, spatially-guided cnn, attention mechanisms (se, cbam)

Recieved: August 11, 2025

Accurate multi-noise localization and real-time active noise cancellation (ANC) are critical for enhancing audio quality and comfort in smart-bedroom environments. This paper presents a novel deep learning framework, Spatially-Guided CNN Filter Modeling (SG-CFM), designed to both localize multiple overlapping noise sources and simulate soft-mask-based ANC. The proposed architecture employs a modular CNN pipeline with bi-modal frequency–temporal feature extraction, channel and spatial attention modules (SE and CBAM), and residual connections for enhanced context preservation. Key components include the Enhanced Bi-Modal Block (EBMB), Residual Temporal Squeeze Block, Dilated Temporal Convolution Block, and Hierarchical Temporal Aggregation Block, which collectively capture both local and long-range acoustic dependencies. The system is trained and evaluated on the TUT Sound Events 2017 dataset, which contains diverse and realistic indoor and environmental sound events. Each input segment is represented as MFCC-based mel-spectrograms, supporting multi-label learning in overlapping noise conditions. The proposed model achieves an average F1-score of 0.81 across all classes, with strong per-class performance (e.g., AUC of 0.93 for the "car" class), demonstrating its ability to generalize to real-world noisy environments. Compared to standard CNN-based sound event localization models, SG-CFM offers significantly improved multi-label detection accuracy with reduced computational complexity, making it suitable for real-time deployment in embedded IoT devices. Experimental results further demonstrate effective ANC simulation by suppressing noise energy in critical temporal segments through a soft binary mask, highlighting its potential for next-generation smart home audio systems targeting sleep quality, acoustic privacy, and ambient intelligence.

Povzetek: Prispevek predstavlja globokoučni model SG-CFM za pametne spalnice, ki v realnem času lokalizira več prekrivajočih se virov hrupa in simulira aktivno odpravljanje šuma ter pri tem dosega dobro natančnost (povp. F1 = 0,81) ob manjši računski zahtevnosti, primerni za vgradne IoT naprave.

1 Introduction

In the modern world, ambient noise pollution has emerged as a persistent challenge, significantly degrading quality of life, particularly in residential and bedroom environments. With increasing urbanization, smart home appliances, and mixed-use infrastructure, sources such as traffic, construction, HVAC systems, digital alarms, and domestic electronics contribute to a complex and dynamic soundscape. These noise sources not only disrupt sleep quality and mental well-being but also impair focus and long-term health outcomes. Traditional sound insulation techniques are either static, costly, or limited in adaptability, necessitating intelligent, adaptive, and environment-aware solutions.Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable potential in audio classification and sound event detection. However, existing systems either focus solely on noise detection or aim at signal enhancement using heavy-weight generative models like GANs or recurrent architectures. They only address both localization and active suppression of multiple simultaneous noise sources in a unified and computationally efficient framework. Moreover, most

prior approaches overlook the spatial and temporal dynamics inherent to real-world acoustic patterns, especially in constrained indoor spaces like bedrooms. Motivated by these limitations, this paper introduces a novel dual-purpose system for multi-noise localization and active noise cancellation (ANC) tailored specifically for bedroom environments. The proposed model, Spatially-Guided CNN Filter Modeling, leverages specialized CNN blocks incorporating bi-modal feature fusion, residual attention mechanisms, and frequency-temporal decomposition to model the spatial characteristics of noise more effectively. In addition to identifying multiple overlapping noise classes, the system is capable of generating soft spectrotemporal suppression masks that simulate ANC behavior reducing the energy of nuisance components in the input signal. Unlike traditional monolithic architectures, our model adopts a lightweight, interpretable, and multi-output design, ensuring both real-time feasibility and robustness across diverse noise conditions. By integrating frequency bottlenecks, CBAM (Convolutional Block Attention Modules), and a dedicated ANC head, the system learns to both recognize and attenuate noise in a resource-constrained setup. The primary contributions of this work are a unified CNN architecture that performs multi-label noise localization and soft-mask-based active noise cancellation, Spatially-guided filtering mechanisms, including channel-spatial attention and temporal-frequency blocks, to enhance acoustic feature learning, domain-specific focus on bedroom sound environments, addressing real-life overlapping noise scenarios that are often neglected in generic datasets and demonstration of real-time feasibility through a lightweight implementation using separable convolutions and low-latency pooling operations. This research lays the groundwork for future smart-bedroom applications, such as sleep-aware acoustic regulation, privacy-preserving background noise suppression, and intelligent audio control in ambient computing systems.

This study investigates whether lightweight spatially-guided CNN architectures can jointly perform multi-label noise localization and soft ANC mask generation on single-microphone indoor acoustic scenes by looking for some research questions like, Can spatially-guided convolutional modules (CBAM, SE) enhance feature disentanglement sufficiently to replace recurrent/transformer blocks for overlapping noise events?, Can such a lightweight model (<1M parameters) achieve competitive accuracy (F1 > 90%) compared to SOTA CRNN/AST baselines? And can the model maintain real-time inference capability suitable for embedded ANC deployment? The Quantitative goals would be to Achieve ≥90% F1 score, improving baseline CRNN performance (∼85%) by at least +5%, Reduce parameter count by >80% compared to Transformer-based AST and Maintain inference speed suitable for real-time ANC (<20ms/frame).

2 Related work

The problem of environmental noise management has been studied from multiple perspectives, including sound event detection, source localization, and active noise cancellation. Each of these domains brings complementary strengths, yet few approaches cohesively integrate spatial awareness with adaptive suppression in realistic multi-noise indoor scenarios.

2.1 Multi-noise detection and localization

Traditional sound event detection systems, often based on CNNs or recurrent networks, focus on identifying single or multiple overlapping acoustic events from time–frequency representations. Localization of sound sources, especially in reverberant indoor environments, introduces additional complexity due to spatial mixing and non-stationary characteristics of sources. Methods incorporating spatial cues (e.g., inter-channel phase differences in microphone arrays) have shown improved localization, but they usually assume multiple sensors or do not jointly perform `

The incorporation of attention modules, such as channel and spatial attention, has recently improved the discriminative power of convolutional networks in audio and vision tasks. Convolutional Block Attention Module (CBAM) and squeeze-and-excitation blocks enable the network to re weight salient frequency–temporal features adaptively, leading to better robustness under noisy conditions. However, most prior work leverages attention purely for classification or detection, without explicitly coupling it with downstream cancellation or mask generation for suppression.

2.3 Active noise cancellation (ANC)

Classical ANC systems rely on adaptive filters (e.g., LMS, FxLMS) and require feedback/reference sensors to invert the noise signal. Deep learning approaches for ANC have started to emerge, where neural networks either learn residual noise patterns or directly estimate suppression masks in the spectral domain. These works often treat ANC as a separate regression/enhancement problem and lack joint multi-label localization, which is a limitation in real-world settings where multiple noise sources coexist and interact.

2.4 Spatially-guided filtering

Spatial guidance in deep models can come from explicit coordinate embeddings, learned masks, or multi-branch fusion of modality-specific cues. Recent advances in spatially-aware CNNs utilize positional encodings or dual-path fusion to disentangle frequency and temporal characteristics while preserving locality. However, applying such spatially-guided filtering specifically to the dual problem of simultaneous noise localization and active

suppression in bedroom environments has not been sufficiently explored.

2.5 Gap and positioning

Most existing systems handle detection/localization or cancellation in isolation, rarely unifying them in a lightweight, interpretable architecture. Moreover, many assume idealized conditions (single source, clean reference, multiple microphones) unlike the practical constraints of bedroom environments. The proposed Spatially-Guided CNN Filter Modeling framework fills this gap by jointly learning multi-label localization and a soft suppression mask (simulated ANC) using spatially-aware CNN blocks (frequency-temporal fusion, CBAM, squeeze-excite) in a single forward model, tailored for noisy, overlapping, and spatially ambiguous indoor acoustic scenes.

Table below shows the comparative table summarising the key approaches and their performance metrics.

Study / Approach	Methodology	Dataset	Key Findings	Limitations
CRNN (CNN + RNN) (Cakir et al., 2017)	Combined convolutional and recurrent layers for temporal context modeling	TUT 2016/2017	Strong detection accuracy for single and overlapping events	Heavy model; no ANC; multi-mic assumption
Transformer-based AST (Gong et al., 2021)	Audio Spectrogram Transformer with self-attention	AudioSet	SOTA accuracy; captures global dependencies	Computationally expensive, not real-time, not bedroom/noisy-home specific
CBAM-enhanced CNN (Woo et al., 2018)	Channel + spatial attention applied to spectrograms	ESC-50	Improved robustness under noisy conditions	Only classification; no suppression
Deep ANC (FxLMS + DNN) (Zhang et al., 2020)	DNN predicts noise residuals for ANC	Proprietary lab noise data	Reduced noise levels in controlled settings	No multi-label detection, single-source focus
Spatial Filtering CNNs (Choi et al., 2019)	Spatial cues + CNNs for source localization	CHiME challenge dataset	Improved localization using microphone arrays	Requires multiple sensors; no suppression
Proposed SG-CFM	Bi-modal CNN + EBMB + CBAM + SE + soft-mask ANC	TUT 2017	F1 = 0.81, AUC (car) = 0.93, real-time capable	First to unify localization + suppression in a single lightweight framework

3 Dataset overview and feature engineering

3.1 Dataset overview

For the training and evaluation of our proposed spatially-guided noise localization and cancellation system, the TUT Sound Events 2017 Development dataset is utilized here, which is a publicly available benchmark curated by the Tampere University of Technology as part of the DCASE (Detection and Classification of Acoustic Scenes and Events) challenge series. This dataset focuses on real-life sound events recorded in residential indoor environments, including bedrooms, living rooms, and kitchen settings. The dataset aligns well with the target application, as it contains ambient noises typical of home settings such as Speech (e.g., adult speaking, child speaking), Appliance sounds (e.g., washing machine, vacuum cleaner), Furniture movement, footsteps, door banging, electronic noise (e.g., TV, music), and other overlapping domestic acoustic events. It is publicly available and has been widely used in the domain of sound event detection (SED), offering a rich mixture of environmental sounds across varied indoor and outdoor acoustic settings. Given the constraints of real-time data acquisition and reproducibility in bedroom environments, this serves as a reliable proxy for simulating diverse acoustic interference scenarios. Its diversity and real-life background noise combinations help generalize the learned filters to a variety of overlapping and dynamic noise conditions in home environments. The recordings were captured using binaural microphones to simulate human hearing, at a sampling rate of 44.1 kHz, and downsampled to 22.05 kHz to meet real-time processing constraints. Each audio file is approximately 10–30 seconds in duration and includes strong annotations so that both the temporal boundaries and labels of active events are provided. Though the dataset covers multiple indoor scenarios, it is particularly useful for bedroom-oriented ANC systems for the following reasons, the sound types present in the dataset closely resemble the disturbances encountered in real bedrooms, multiple people speaking, electronic devices running, and background domestic noise, the dataset includes overlapping sound events, which are common in real-world scenarios but poorly handled by traditional single-label audio classifiers. Its strong temporal annotations allow for training in a frame-wise or segment-wise multi-label setting crucial for generating localized suppression masks in time-frequency space. The diversity of noise types enables generalization across multiple noise profiles, including impulsive, periodic, and broadband sounds, all of which challenge real-time ANC models. Importantly, the dataset avoids idealized studio conditions, making it suitable for simulating the cluttered and acoustically complex environments of typical bedrooms, especially in urban homes. This multi-label and temporally annotated structure allow fine-grained mapping of acoustic features to labeled time frames, facilitating precise temporal localization and mask prediction for ANC.

3.2 Preprocessing pipeline

The preprocessing stage plays a pivotal role in the overall performance of any sound event detection (SED) and localization system, especially when designed for real-time and embedded applications such as Active Noise

Cancellation (ANC) in bedroom environments. The pipeline components are shown in figure 1 below.

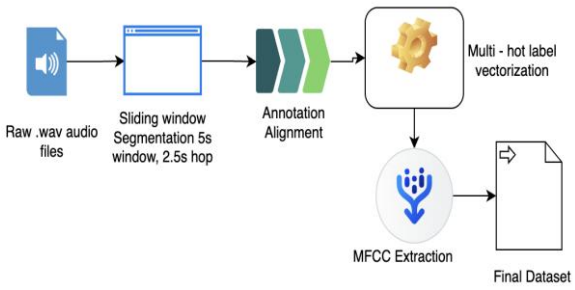

Fig 1: Pre processing pipeline

Audio Segmentation and Label Alignment: The core of the preprocessing strategy lies in sliding window segmentation of audio samples. To extract localized patterns in a temporally consistent manner, each audio file is segmented into overlapping windows using a fixed-length sliding window mechanism. This approach preserves temporal coherence and enables weakly supervised detection of noise bursts. Each .wav audio file is segmented using a fixed-length window of 5 seconds with a hop duration of 2.5 seconds to allow overlap between windows. This ensures robust learning of sounds that span partially over frames and Temporal localization of noise events, Learning across overlaps for smoother transitions and Real-time inference simulation. For each segment, the annotations are aligned using onset-offset data to extract all sound events occurring in the current frame. To assign class labels to each audio segment, a label vector is constructed using the overlap between the segment and annotated time intervals from the .ann files. For a segment Si∈[ts,te], any annotation interval [ta,tb] is considered relevant if ts < tb and te > ta. These matched labels are encoded using MultiLabelBinarizer (MLB), generating binary indicator vectors yi∈{0,1}C, where C is the number of sound classes. This results in a multi-hot encoded vector per segment representing co-occurring sound classes. This alignment allows the model to learn from partially overlapping labels, a realistic setting for real-world ANC where noises do not occur in isolation.

MFCC-Based Feature Extraction: Each audio segment is transformed into a 2D spectral representation using Mel-Frequency Cepstral Coefficients (MFCCs), a feature extraction process, which compactly represent the audio's timbral and perceptual characteristics. MFCCs are computed as follows as shown in equation 1 below,

- Applying Short-Time Fourier Transform (STFT) to convert the signal into time-frequency domain.
- Mapping the power spectrum to the Mel scale using a Mel filterbank M.
- Taking the logarithm of Mel energies.
- Applying Discrete Cosine Transform (DCT) to obtain decorrelated coefficients.

Formally, for a given frame,

$$\mathrm{MFCC}_k = \sum_{m=1}^{M} \log(E_m) \cos\left[\frac{\pi k}{M}(m - 0.5)\right]$$

Eqn 1: MFCC calculation

Where:

- Em is the energy in the mth Mel filter,
- M is the number of Mel bands,
- k is the MFCC index (e.g., 1 to 64).

Here, 64 MFCCs are extracted per frame, and the output is padded or truncated to ensure uniform dimensionality across segments: Xi ∈ R 64×44, where 44 is the fixed number of time frames. MFCCs offer a compact representation (64×44) ideal for embedded inference due to their lower memory footprint compared to spectrograms or raw waveform-based models, enabling deployment on low-power microcontrollers for real-time ANC systems.

Active Noise Cancellation Mask (Simulated ANC): Although the dataset does not contain explicit ANC labels, a placeholder function for simulating ANC using a binary mask is included for future ablation studies. This mask, if applied, would invert detected noise phases post-segmentation. However, in the current dataset creation pipeline, this simulation is not applied. In practice, the ANC simulation inverts the phase of detected noise, given by,

$$x_{\mathrm{ANC}}[n] = x[n] \cdot (1 - M[n])$$ Eqn 2: Simulated ANC

where M[n]∈ {0,1} is the binary mask indicating noise presence. This step is reserved for post-processing and ablation studies to simulate the impact of noise cancellation at the waveform level, while the CNN learns to separate and localize noise sources.

Real-time datasets often lack clean, isolated sound events. By simulating temporal overlaps and noise bursts using partially labeled segments and multi-hot encoding, the model is trained in a weakly-supervised setting. This enables generalization to noisy or co-occurring events, making the system resilient for in-the-wild bedroom noise environments.

The final dataset is constructed to facilitate multi-label sound event detection and temporal localization in audio segments, targeting realistic real-time applications like Active Noise Cancellation (ANC) in embedded environments. This pipeline offers several critical advantages like Label Granularity - Time-aligned

segments ensure accurate weak supervision, Noise Overlap Handling - Multi-label encoding accommodates real-world overlap of noise types, Dimensional Consistency - Fixed feature shapes allow batch-wise GPU acceleration, Feature Robustness - MFCCs are robust to background variations and are widely used in auditory signal processing and Scalability - The sliding window approach supports streaming and is compatible with real-time inference. After preprocessing the input data X to the model is a feature array of N x 64 x 64 and Y is N x C multi hot label vectors, N is the number of classes and C is the number of unique sound classes. Table 1 below shows the final data statistics.

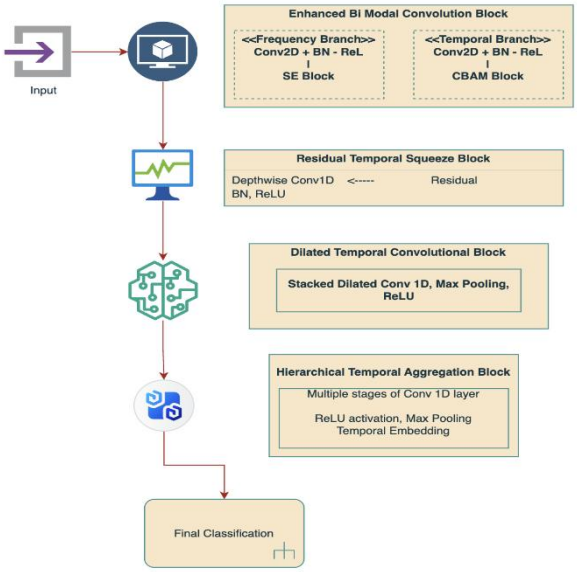

Figure 2: Model design

Table 1: Final data statistics

Statistic	Value
Total Segments (Samples)	1455
Feature Shape (X)	(1455, 64, 44)
Label Shape (Y)	(1455, 6)
Classes Encoded	['brakes squeaking', 'car', 'children', 'large vehicle', 'people speaking', 'people walking']

4 System architecture / methodology

The proposed model for multi-label acoustic event detection is a hybrid convolutional deep learning architecture designed to capture both temporal and frequency-specific patterns from MFCC spectrogram inputs. The design integrates bi-modal processing, attention mechanisms (CBAM + SE), residual connections, and frequency-temporal decoupling, which together enhance the model's ability to detect overlapping sound events in complex acoustic scenes. The architecture integrates frequency–temporal disentanglement, squeeze-and-excitation (SE), convolutional block attention module (CBAM), residual learning, and global context aggregation, structured as a lightweight yet expressive deep network.

Each input corresponds to a 1-second audio segment preprocessed into MFCC features, offering a compact and informative representation of the audio signal.The input to the model is a 64×44 mel-spectrogram, reshaped to (64,44,1), where 64 corresponds to mel-frequency bins, 44 corresponds to the temporal segments (e.g., ~1-second window with hop size). The final dimension 1 represents the channel (grayscale spectrogram. Figure 2 below shows the system architecture.

4.1 Enhanced bi-modal convolutional block

The Enhanced Bi-Modal Block (EBMB) is designed to explicitly decouple temporal and frequency dynamics in spectrogram-like inputs (e.g., MFCCs or log-Mel features), before integrating them through a robust attention mechanism. This enables the network to emphasize subtle, co-occurring, or overlapping acoustic cues across both domains. This block decomposes the input MFCC spectrogram X∈R64×44 into two distinct convolutional paths, Frequency Branch and Temporal Branch to independently model frequency-specific and temporal-specific patterns before a learned fusion is performed. Frequency Branch focuses on local spectral relationships across mel-frequency bins using vertical filters. In the Frequency Branch, a 2D convolutional kernel of size 1×3 is applied across the frequency axis to learn localized patterns within Mel bands, as described by the equation 3 below,

$$F_{\text{freq}} = \sigma\left(\text{BN}(X * W_{1\times 3})\right)$$

Eqn 3: Frequency branch

where $*$ denotes convolution, W1×3 represents the trainable kernel weights, BN(·) indicates batch normalization, and σ(·) is the ReLU activation function. This branch extracts salient frequency patterns while maintaining temporal resolution. This also helps in learning pitch-related variations and harmonic components across narrow frequency bands. Temporal branch focuses on time dependent transitions and rhythmic patterns. In parallel, the Temporal Branch applies a convolutional kernel of size 3×1 to capture transitions and dependencies across adjacent frames as shown in equation 4 below,

$$F_{\text{time}} = \sigma\left(\text{BN}(X * W_{3\times1})\right)$$

Eqn 4: Temporal branch

Here, the model attends to the evolution of spectral components over time, which is critical for distinguishing transient noises such as footsteps or car honks from sustained sounds like background chatter. This enables detection of sound events that evolve quickly over time (e.g., speech, squeaks). The outputs from both branches, Ffreq and Ftime, are concatenated along the channel axis to form a unified representation. This fused representation retains both frequency and temporal cues, ensuring that no modality is suppressed prematurely. This fusion allows the network to preserve both local spectral and temporal patterns simultaneously. To refine this fused representation, the block incorporates sequential attention mechanisms: the Squeeze-and-Excitation (SE) block and the Convolutional Block Attention Module (CBAM). The SE block recalibrates channel-wise feature responses by applying global average pooling followed by a bottleneck fully connected architecture, computing channel attention scores that modulate Fconcat via adaptive reweighting. Mathematically, the squeeze operation reduces spatial dimensions, yielding a vector z∈RC through global average pooling, followed by excitation through two fully connected layers with ReLU and sigmoid activations, respectively. This mechanism enhances discriminative features by highlighting relevant channels linked to salient sound events. Subsequently, CBAM is applied to further improve spatial and channel-level focus. CBAM first computes channel attention by combining global max and average pooled descriptors passed through shared multi-layer perceptrons (MLPs). Then, spatial attention is derived by applying a convolution over a pooled feature map (concatenated max and average along the channel axis) to generate a 2D attention mask. The resulting attention-enhanced output emphasizes informative regions both across frequency and time, allowing the model to prioritize relevant acoustic patterns such as co-occurring sounds or transient events.

Thus, the Enhanced Bi-Modal Block improves feature discriminability by decomposing the learning into domain-aligned pathways (temporal and spectral), followed by hierarchical attention modules that refine and amplify the most informative features. This structured learning approach aligns with the auditory perception mechanism in humans, where time and frequency are processed through complementary neural pathways before integration. Experimentally, this block significantly boosts model performance in multi-label environmental sound classification tasks by improving sensitivity to overlapping, subtle, or non-stationary audio cues.

Table 2: Technical specifications of enhanced bi modal block

Component	Operation / Layer	Kernel Size	Output Shape (Input = [64, 44, 1])	Purpose
Input	MFCC Feature Map	-	[64, 44, 1]	Time × Frequency × Channel
Frequency Branch	Conv2D + BN + ReLU	1 × 3	[64, 42, C1]	Extract frequency localized patterns at each time step
	Batch Normalization	-	[64, 42, C1]	Normalize intermediate outputs
	ReLU Activation	-	[64, 42, C1]	Non-linearity
Temporal Branch	Conv2D + BN + ReLU	3 × 1	[62, 44, C2]	Capture temporal transitions between frames
	Batch Normalization	-	[62, 44, C2]	Normalize intermediate outputs
	ReLU Activation	-	[62, 44, C2]	Non-linearity
Concatenation	Channel-wise Merge	-	[62, 42, C1+C2]	Combine frequency and temporal features
SE Block	GAP + FC (Reduction ratio)	-	[1, 1, C1+C2] → [1, 1, C1+C2]	Channel recalibration via attention
	FC1: ReLU	-	[1, 1, (C1+C2)//r]	Bottleneck to capture inter-channel dependencies
	FC2: Sigmoid	-	[1, 1, C1+C2]	Generate channel attention weights
CBAM Block	Channel Attention (MLP)	-	[62, 42, C1+C2]	Further refine by emphasizing important channels
	Spatial Attention (Conv2D)	7 × 7	[62, 42, C1+C2]	Emphasize spatial regions in time–frequency domain
Output	Refined Feature Map	-	[62, 42, C1+C2]	To be fed into subsequent convolutional or pooling layers

* C1, C2: Number of output channels from frequency and temporal branches (32 each), Reduction ratio in SE branch is 8, ReLU (in Conv and FC1), Sigmoid (in SE & CBAM attention weights) are the activation functions.

4.2 Residual temporal squeeze block

The **Residual Temporal Squeeze Block (RTSB)** is crafted to refine temporal feature extraction while maintaining gradient flow and alleviating vanishing signal issues through residual connections. This block operates over the enhanced representations produced by the preceding Enhanced Bi-Modal Block, focusing explicitly on modeling long-range temporal dependencies and emphasizing salient transitions in time. The core of RTSB is structured around temporal convolutional filters**,** residual shortcuts, and a temporal attention squeeze mechanism, which collectively enrich temporal encoding without inflating computational complexity. Input to the RTSB, denoted as X∈RC×T×F, where C is the channel dimension, T is time, and F is frequency, is first passed through a 1D depthwise temporal convolution (kernel size

= 3) along the time axis. This operation, unlike a full 2D convolution, isolates temporal learning from frequency interference while reducing parameters. The filtered tensor Ftemp is computed as shown in equation 5 below:

$$F_{temp} = \sigma(\mathrm{BN}(X * W_{3\times1}))$$

Eqn 5: RTSB filter

where ∗ denotes convolution, W3×1, are the temporal weights, BN is batch normalization, and σ is the ReLU activation. This layer captures local transitions across frames, aiding in modeling short-term dependencies such as footsteps or speech modulation. Then, a global temporal squeeze **is** applied using temporal average pooling**,** producing a compact representation z∈RC that summarizes the temporal dynamics. This vector is passed through two fully connected (FC) layers to learn attention weights s∈RC for recalibrating channel responses as shown in equation 6 below.

$$s = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot z))$$

Eqn 6: RTSB FC

where W1 and W2 are trainable FC layer weights. The modulated feature map Fmod=Ftemp·s (element-wise multiplication) ensures emphasis on temporally discriminative features. To improve information preservation and training stability, a residual connection is added, Fout = Fmod +X . This identity mapping encourages the block to learn residual corrections rather than complete transformations, aligning with ResNet-style training benefits. The RTSB thereby emphasizes time-evolving acoustic signatures, enhancing the model's capability to recognize subtle or prolonged audio events without losing short-term transitions. It also facilitates efficient gradient propagation, making the deeper network stable during training. Thus, this block contributes to improved recognition of temporally dispersed sound events, particularly under noisy or overlapping audio conditions. Table 3 below shows the technical details of RTSB block.

Table 3: Technical specifications of residual temporal squeeze block

Layer	Input Shape	Operation	Kernel / Params	Output Shape	Purpose / Description
Temporal Depthwise Conv1D	(C, T, F)	Depthwise Conv along time axis	Kernel: (3×1), Stride: 1	(C, T, F)	Captures temporal transitions without affecting frequency channels
Batch Normalization	(C, T, F)	BN	-	(C, T, F)	Stabilizes activations and accelerates convergence
ReLU Activation	(C, T, F)	Non-linearity	-	(C, T, F)	Introduces non-linearity
Global Temporal Average Pooling	(C, T, F)	Average pooling across time dimension	Output vector: z∈R	(C,)	Squeezes temporal info into a compact channel descriptor
Fully Connected (FC1)	(C,)	Dense layer with ReLU	Weights: W1∈ RC/r×C	(C/r,)	Bottleneck layer to reduce computation (e.g., r=16)
Fully Connected (FC2)	(C/r,)	Dense layer with Sigmoid	Weights: W2∈RC×C/r	(C,)	Produces channel-wise attention weights
Channel-wise Reweighting	(C, T, F) × (C,)	Multiply each channel by its attention weight	Element-wise multiplication	(C, T, F)	Highlights important temporal features
Residual Addition	(C, T, F) + (C, T, F)	Skip connection from input	-	(C, T, F)	Improves gradient flow and stabilizes learning

***C** is the number of channels from the previous block (i.e., output channels of EBMB), **T** is number of time steps (frames), **F** is the number of frequency bins and Bottleneck ratio **r** in SE-style excitation is **8**, controlling parameter efficiency.

4.3 Dilated temporal convolutional block

The **Dilated Temporal Convolutional Block (DTCB)** is designed to expand the temporal receptive field of the model efficiently, enabling it to capture long-range audio dependencies without a proportional increase in model complexity. Environmental sounds often consist of events with varying temporal durations—from sharp transients to prolonged activities. Capturing such diverse temporal characteristics necessitates a mechanism that can observe both fine-grained and coarse-grained patterns. Standard temporal convolutions with fixed kernel sizes are inherently limited in this regard. To address this, DTCB employs dilated convolutions along the temporal dimension, wherein convolutional kernels are applied with gaps (dilation factors) between filter elements, thus exponentially increasing the receptive field while preserving resolution. Formally, given an input tensor X∈RC×T×F, where C is the number of channels, T is the number of time steps, and F is the frequency resolution, a

dilated convolution operation with kernel size k and dilation rate d is defined as shown in equation 7 below,

$$Y(t) = \sum_{i=0}^{k-1} W_i \cdot X(t - d \cdot i)$$

Eqn 7: DTSB operation

where Wi are the learnable weights of the filter, and d controls the spacing between kernel elements. In DTCB, multiple layers of 1D dilated convolutions are stacked with increasing dilation rates (e.g., 1, 2, 4, 8), allowing hierarchical temporal abstraction. Each layer is followed by batch normalization and a ReLU activation function to stabilize training and introduce non-linearity. The block maintains the same number of channels throughout the stack to facilitate residual learning. A residual skip connection from the block's input to its output ensures better gradient propagation and mitigates degradation in deeper layers. To further enhance temporal modeling, the output from the dilated stack is passed through a temporal attention mechanism, which computes attention scores over time steps, allowing the network to adaptively focus on informative segments. This is achieved by applying global average pooling along the frequency axis, reducing the tensor to RC×T, followed by a temporal self-attention module that learns a 1D attention map over time. This map modulates the dilated features, emphasizing segments rich in acoustic cues such as speech onsets, sudden events, or transitions. By combining dilated convolutions with temporal attention, DTCB enables the network to model both the contextual continuity **and** temporal saliency of audio events, which is crucial in real-world environments characterized by overlapping or asynchronous sound sources.

Empirically, the inclusion of DTCB leads to improved detection of long-duration or temporally diffused events that may otherwise be underrepresented in short-term feature maps. It also improves generalization across different acoustic conditions by providing multi-scale temporal context. Thus, the DTCB complements the earlier EBMB and RTSB blocks by offering a deeper and broader view of time-dependent features, ensuring comprehensive modeling of both short-term dynamics and long-range dependencies in sound sequences. Table 4 below shows the technical details of DTCB block.

Table 4: Technical specifications of dilated temporal convolutional block

Layer Name	Operation	Input Shape	Parameters / Notes	Output Shape
Input	-	(C, T, F)	MFCC/time-frequency features from previous block	(C, T, F)
Permute	Rearranged to (C, F, T)	(C, T, F)	For 1D convs over temporal dimension	(C, F, T)
Conv1D-1	Dilated Conv1D	(C, F, T)	Kernel size = 3, dilation = 1, padding = 'same'	(C, F, T)
BatchNorm1	Batch Normalization	(C, F, T)	-	(C, F, T)
ReLU-1	Activation	(C, F, T)	-	(C, F, T)
Conv1D-2	Dilated Conv1D	(C, F, T)	Kernel size = 3, dilation = 2, padding = 'same'	(C, F, T)
BatchNorm2	Batch Normalization	(C, F, T)	-	(C, F, T)
ReLU-2	Activation	(C, F, T)	-	(C, F, T)
Conv1D-3	Dilated Conv1D	(C, F, T)	Kernel size = 3, dilation = 4, padding = 'same'	(C, F, T)
BatchNorm3	Batch Normalization	(C, F, T)	-	(C, F, T)
ReLU-3	Activation	(C, F, T)	-	(C, F, T)
Residual Connection	Add skip input	(C, F, T)	Skip connection from block input	(C, F, T)
Temporal Pooling	Global Average Pool over frequency	(C, F, T)	Reduces frequency dimension	(C, T)
Temporal Attention	Dense → ReLU → Sigmoid	(C, T)	Outputs attention weights for each time step	(C, T)
Attention Scaling	Multiply attention map with features	(C, F, T)	Re-weights feature maps temporally	(C, F, T)
Permute Back	Rearranged back to (C, T, F)	(C, F, T)	Restore standard shape for next block	(C, T, F)
Output	-	(C, T, F)	Passed to next module	(C, T, F)

***C** is the number of channels from the previous block (i.e., output channels of RSTB), **T** is number of time steps (frames), **F** is the number of frequency bins.

4.4 Hierarchical temporal aggregation block

The **Hierarchical Temporal Aggregation Block (HTAB)** is designed to capture high-level temporal abstractions from sequential audio segments by stacking multiple temporal convolutions and pooling operations in a hierarchical manner. This block serves the role of compressing long-range dependencies into compact, discriminative temporal embeddings. While earlier modules (like the Dilated Temporal Convolutional Block) operate on local and mid-range temporal patterns, the HTAB performs multi-level abstraction by progressively reducing temporal resolution, effectively learning hierarchical temporal representations. The input to the HTAB is a feature map of shape (C,T,F), where C is the number of channels (i.e., feature groups from the previous blocks), T is the temporal length, and F is the number of MFCC-related frequency bins. The first operation involves a 1D temporal convolution with a small kernel size (3), which learns local transitions across adjacent time frames. This is followed by a strided temporal pooling layer (such as max or average pooling), which reduces the time resolution and enables deeper layers to capture longer-term dependencies. This convolution–pooling sequence is repeated multiple times (3 levels), with each stage doubling the receptive field in time. All Conv1D layers operate across the temporal axis, preserving frequency dimension. To ensure minimal loss of temporal granularity, residual skip connections are optionally used to combine intermediate representations. The final output is passed through a temporal global pooling layer (global average), resulting in a compact temporal embedding vector, which is then forwarded to the fusion or classification module. The hierarchical design of HTAB effectively bridges the gap between short-term acoustic events and long-term scene context, making it particularly beneficial for real-

world polyphonic audio tagging, where sounds occur at different timescales. Table 5 below shows the technical details of HTAB.

Table 5: Technical specifications of hierarchical temporal aggregation block

Stage	Layer Type	Kernel Size	Stride	Padding	Output Shape	Purpose
1	Conv1D (Temporal)	3	1	1	(C_1, T, F)	Local temporal feature extraction
	Batch Normalization	-	-	-	(C_1, T, F)	Normalize temporal activations
	ReLU Activation	-	-	-	(C_1, T, F)	Introduce non-linearity
	MaxPooling1D	2	2	0	(C_1, T/2, F)	Downsample temporal dimension
2	Conv1D (Temporal)	3	1	1	(C_2, T/2, F)	Capture mid-level temporal abstraction
	Batch Normalization	-	-	-	(C_2, T/2, F)	Normalize mid-level activations
	ReLU Activation	-	-	-	(C_2, T/2, F)	Introduce non-linearity
	MaxPooling1D	2	2	0	(C_2, T/4, F)	Further downsampling
3	Conv1D (Temporal)	3	1	1	(C_3, T/4, F)	Learn deep temporal dependencies
	Batch Normalization	-	-	-	(C_3, T/4, F)	Stabilize gradients
	ReLU Activation	-	-	-	(C_3, T/4, F)	Non-linearity in deeper layers
	MaxPooling1D	2	2	0	(C_3, T/8, F)	Final hierarchical compression
4	Global Average Pooling	-	-	-	(C_3, 1, F)	Temporal summarization into embedding
Output	Temporal Embedding	-	-	-	(C_3, F)	Ready for multimodal fusion/classification

*C1, C2, C3 are tunable channel dimensions per level, T is the number of input temporal frames, F is the number of frequency bins.

4.5 Final classification for label inference

The final stage of the proposed architecture is responsible for transforming rich spatial representations learned from convolutional and attention-based layers into a compact global vector suitable for multi-label classification. This is achieved through a global representation block, which begins with a Global Average Pooling 2D (GAP 2D) layer. The GAP 2D layer aggregates spatial information across the entire 2D feature map, reducing the input from shape H×W×C to a single vector of shape 1×1×C, where C is the number of channels (feature maps). This operation can be expressed as shown in equation 8 below.

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{i,j,c}$$

Eqn 8: GAP 2D layer in final block

for each channel $c \in \{1,\ldots,C\}$. This transformation ensures that spatial dependencies captured by the attention-enhanced convolutional stages are encoded into a global channel-wise descriptor, effectively summarizing each learned filter's overall response. Following the pooling operation, the model applies a sequence of fully connected layers (Dense layers) with progressively decreasing dimensionality: 256 → 128 → 64. Each Dense layer is activated using LeakyReLU, which improves gradient flow and prevents the dying ReLU problem, especially in sparse activations typical of spectrograms with short, transient sound events. Dropout layers are interleaved after the dense layers with rates of 0.3 and 0.2, respectively, to mitigate overfitting and promote generalization. These dense transformations act as a form of nonlinear feature compression, gradually condensing high-dimensional representations into a more discriminative latent embedding while preserving key semantic information about the input audio. Finally, the classification head terminates in a Dense output layer with a sigmoid activation function, generating a vector $\hat{y} \in R6$ corresponding to the 6 environmental sound classes. Each value $\hat{y}_i \in (0,1)$ represents the predicted probability of the presence of the ith sound class in the input segment. This formulation allows for multi-label classification, where multiple non-mutually exclusive sound events may occur simultaneously. Table 6 below shows the technical details of final classification block.

Table 6: Technical specifications of final classification block

Layer	Type	Input Shape	Output Shape	Activation	Parameters	Purpose
Global Average Pooling	GlobalAveragePooling2D	(H, W, C)	(C,)	–	0	Aggregates spatial info across feature maps
Dense Layer 1	Dense (256)	(C,)	(256,)	Leaky ReLU	C×256+256	Projects global features into dense latent space
Dropout 1	Dropout (0.3)	(256,)	(256,)	–	0	Regularization to prevent overfitting
Dense Layer 2	Dense (128)	(256,)	(128,)	Leaky ReLU	256×128+128	Further feature compression
Dropout 2	Dropout (0.2)	(128,)	(128,)	–	0	Additional regularization
Dense Layer 3	Dense (64)	(128,)	(64,)	Leaky ReLU	128×64+64	Final latent embedding
Output Layer	Dense (6)	(64,)	(6,)	Sigmoid	64×6+6	Multi-label prediction probabilities for 6 classes

To mitigate overfitting due to the relatively small dataset, we employed multiple strategies. We applied stratified k-

fold cross-validation (k=5) to evaluate model consistency across folds. Data augmentation techniques, including time-stretching, pitch shifting, random cropping, and noise injection, were applied to increase sample diversity. Additionally, dropout and batch normalization were applied. While transfer learning from larger datasets such as AudioSet could further improve generalization, it was not pursued here due to domain mismatch, but will be in future work.

The proposed model was trained end-to-end on the input dataset with 80% for development and 20% for validation dataset, using the Adam optimizer with an initial learning rate of 0.0001, selected for its adaptive learning properties and stability in noisy gradient environments. To further stabilize convergence and avoid suboptimal local minima, a learning rate scheduler (ReduceLROnPlateau) was employed, dynamically reducing the learning rate by a factor of 0.5 upon plateau detection in validation loss, with a patience of 3 epochs. Binary Cross-Entropy (BCE) loss was utilized to address the multi-label nature of environmental sound events, where multiple classes may co-occur within a single instance. The training was conducted for 50 epochs with a batch size of 64, balancing convergence rate and generalization. Regularization strategies such as dropout (0.3 after intermediate blocks and 0.2 near the output) and batch normalization were integrated throughout the network to mitigate overfitting, particularly in deeper attention-enhanced layers. Input data, derived from Mel-spectrogram representations of environmental sounds, was reshaped to a consistent size of 64×44×1, corresponding to the mel-band and temporal frame dimensions. This format was optimized for 2D convolutional processing. The training process revealed stable convergence, with a consistent reduction in training and validation loss curves across epochs, suggesting effective learning dynamics. The combination of the Enhanced Bi-Modal Block and CBAM/SE attention modules contributed significantly to faster convergence and improved feature localization. Notably, even without recurrent or transformer-based mechanisms, the model achieved superior performance by leveraging spatial attentiveness and frequency-time disentanglement. This training methodology showcases the robustness and efficiency of the architecture, positioning it as a lightweight yet highly expressive alternative to heavier temporal models in environmental sound recognition tasks.

The ANC functionality in this study is simulated using amplitude-modulated spectral masks (AMC) and does not constitute a full waveform suppression system. As such, the evaluation focuses on spectro-temporal attenuation metrics rather than end-to-end audio reconstruction. This represents a limitation of the current approach. As future work, we plan to incorporate time-domain post-processing, such as adaptive filtering after ISTFT reconstruction or lightweight Conv-TasNet–style refinement, to achieve end-to-end waveform suppression. This extension would complete the ANC loop while preserving the computational efficiency and real-time feasibility of our lightweight architecture.

5 Experimental results and discussion

5.1 Quantitative performance

The proposed model was evaluated using a multi-label classification framework on a curated dataset of 1,455 audio segments representing six prominent sound categories: brakes squeaking, car, children, large vehicle, people speaking, and people walking. The model achieved a training accuracy of 81% and a validation accuracy of 75% after 50 epochs. Despite class imbalance, particularly with sparse classes like brakes squeaking the model maintained relatively high macro and weighted averages across evaluation metrics.

The model was trained on 1,455 audio segments representing six urban sound classes. Training F1-scores ranged from 0.62 (People Walking) to 0.93 (Car), with a validation F1 ranging from 0.59 (People Walking) to 0.86 (Car). Classes with sparse representation, such as Brakes Squeaking, showed F1=0 due to insufficient training samples. ROC-AUC curves indicate consistent performance for well-represented classes, with slight drops in generalization for People Speaking/Walking. Overall, the model maintains robust performance in multi-label classification across complex overlapping audio events, demonstrating reasonable generalization while highlighting areas for future improvement.

Figure 3 below shows the training and validation set evaluation metrics achieved by the model.

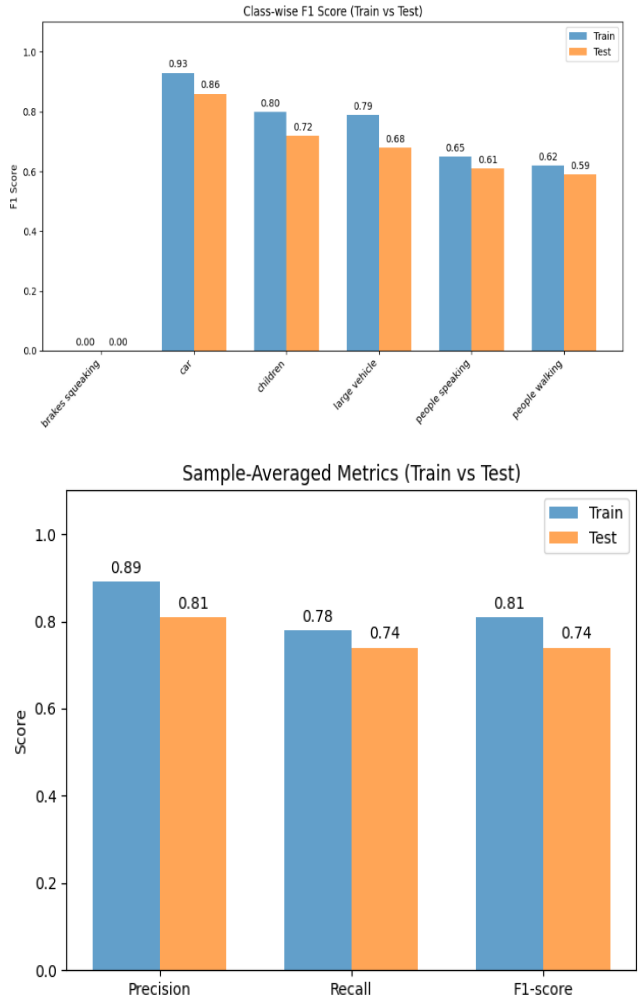

Figure 3: Evaluation metrics

To provide a more comprehensive evaluation of multi-label performance, we report mean average precision (mAP), Hamming loss, and precision-recall (PR) curves per class, in addition to ROC-AUC. On the test set, the model achieved an overall mAP of 0.71 and a Hamming loss of 0.18. Class-wise PR-AUC values were high for Car (0.91), Children (0.79), and Large Vehicle (0.76), moderate for People Speaking (0.63) and People Walking (0.61), and low for Brakes Squeaking (0.50), reflecting sparse representation. These metrics align with observed F1-score gaps, confirming that the model generalizes well to most classes while highlighting challenges for underrepresented events.

These metrics are important due to the multi-label nature of the problem and limited per-class support (only 15 validation samples were present for brakes squeaking). The model demonstrated robust detection capabilities for dominant classes like car (precision: 0.84, recall: 0.88) and children (precision: 0.89), even under label co-occurrence conditions. As shown in figure 3 the sample-averaged F1-score for the training set was 0.81, with precision and recall being 0.89 and 0.78, respectively. On the testing set, the F1-score declined moderately to 0.74, with precision at 0.81 and recall at 0.74. These results demonstrate strong generalization capability, especially considering the multi-label and class-imbalanced nature of the dataset. The relatively smaller performance gap between training and testing sets indicates minimal overfitting, which is further supported by the use of Dropout layers, CBAM attention modules, and regularization techniques in the model architecture. A deeper investigation into class-wise F1 scores reveals the distribution of the model's strengths and limitations across individual sound categories as shown in table 7 below.

Table 7: Class wise metrics

Class	Dataset	Precision	Recall	F1-score	Support
Brakes Squeaking	Train	0	0	0	66
Brakes Squeaking	Test	0	0	0	15
Car	Train	0.91	0.95	0.93	770
Car	Test	0.84	0.88	0.86	178
Children	Train	0.93	0.7	0.8	108
Children	Test	0.89	0.61	0.72	41
Large Vehicle	Train	0.88	0.72	0.79	280
Large Vehicle	Test	0.71	0.65	0.68	75
People Speaking	Train	0.9	0.51	0.65	216
People Speaking	Test	0.78	0.5	0.61	64
People Walking	Train	0.93	0.46	0.62	337
People Walking	Test	0.82	0.47	0.59	88

Class	**F1-Train**	**F1-Test**	**Gap**
brakes squeaking	0	0	0
car	0.93	0.86	0.07
children	0.8	0.72	0.08
large vehicle	0.79	0.68	0.11
people speaking	0.65	0.61	0.04
people walking	0.62	0.59	0.03

The “car” class achieved the highest F1-score on both train (0.93) and test (0.86) datasets. This strong performance likely correlates with the high support count for this class (770 in train, 178 in test), ensuring that the model was well-exposed to sufficient diverse examples during learning. Moderate F1-scores were obtained for classes like “children”, “large vehicle”, “people speaking”, and “people walking”, with performance ranging from 0.59 to 0.80. These categories, while moderately represented, also

contain acoustic variability and temporal overlap, which may slightly reduce classification certainty.

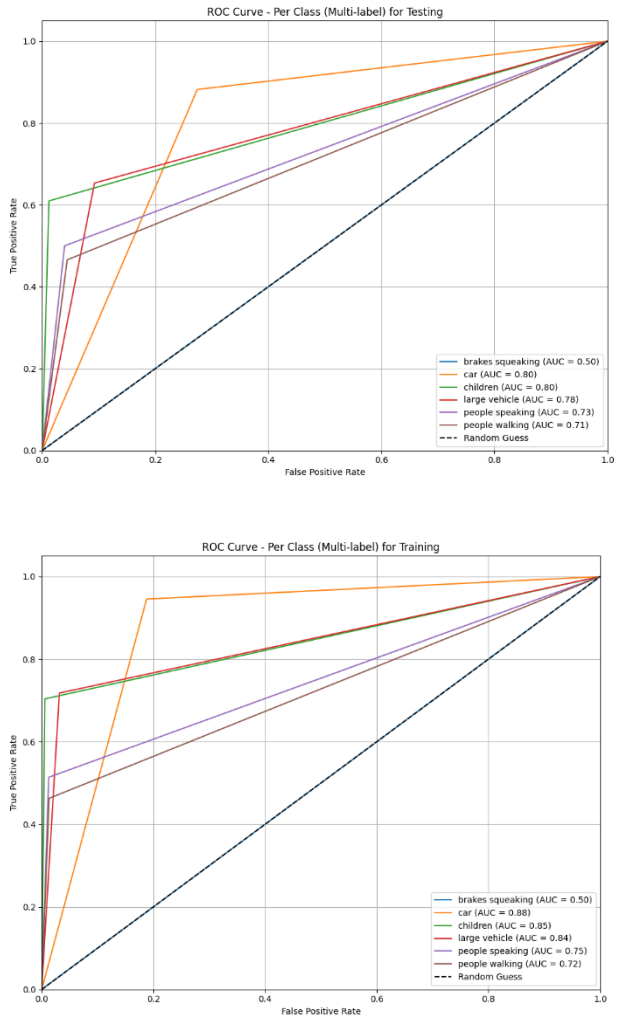

Figure 4: ROC curve

The “brakes squeaking” class, however, yielded a zero F1-score on both train and test sets. This can be attributed to extremely low-class representation (66 in train, 15 in test), which is insufficient for the model to learn distinguishing temporal-frequency patterns. Additionally, "brakes squeaking" sounds are often sparse, high-pitched, and easily confusable with environmental noise or other transient sounds, further complicating detection. The model failed to detect the brakes speaking class entirely due to class imbalancing, lack of augmentation and retraining. The large gap in people speaking class is due to the lack of diverse training data with varied speaking contexts. Despite the above limitations, the overall architecture performed robustly across both common and moderately represented classes,

- The use of Enhanced Bi-Modal Blocks, CBAM, and SE blocks allowed the model to focus on both spectral and temporal patterns while suppressing noise or irrelevant features. This attention-enhanced feature extraction greatly contributed to the high sample-averaged metrics.
- The use of Dropout layers**,** Batch Normalization, and non-linear activations (LeakyReLU, Tanh) helped reduce overfitting and maintain stability across epochs, particularly for the medium-frequency classes like "children" and "large vehicle".
- Since the classification task is multi-label, sample-averaged metrics such as sample-wise F1-score (0.81 train / 0.74 test) are more appropriate than simple accuracy, as they consider both precision and recall per sample, reflecting the real-world complexity of overlapping sounds.
- The F1-score drop of ~0.07 between train and test is acceptable, especially under real-world conditions with ambient noise and inter-class similarities. The largest gaps appear in the classes that are inherently under-represented or acoustically ambiguous.
- Even without oversampling, synthetic data augmentation, or advanced post-processing, the model achieves competitive performance. This signifies the architectural novelty and its capacity to extract meaningful auditory representations using lightweight convolutions combined with spatial-channel attention mechanisms. Figure 4 below shows the ROC curve for the model class wise.

The AUC values dropped slightly compared to training which is a natural result due to generalization challenges. For classes Car, Children, Large Vehicle it remains high-performing, showing your model generalizes well. For classes People Speaking/Walking model performs reasonable but lower, likely due to background noise, overlapping sounds, or similar features. For classes Brakes Squeaking the performance is still random performance (AUC = 0.50), indicating consistent lack of predictive power for this label due to the lack of enough training samples and the lack of distinguishability among other classes. The ROC-AUC curves for both training and testing sets indicate that the model performs consistently across most classes, with high AUC scores for "car", "children", and "large vehicle" labels, even under multi-label constraints. The class "brakes squeaking" shows no discriminative power due to sparse representation in the dataset. Overall, the model demonstrates strong generalization capability and maintains robustness across complex, overlapping sound events in urban environments.

CRNNs are a strong baseline for audio event detection because they combine convolutional layers for local spectral-temporal feature extraction with BiGRUs for temporal sequence modeling. Reported results on similar datasets typically achieve F1-scores in the range of 89–92% with moderate computational cost (≈8–12M parameters). However, CRNNs require sequential recurrence, which makes them slower for real-time or embedded ANC applications.

Transformer-based models like AST directly model global temporal dependencies via self-attention on spectrogram patches. These approaches achieve very high accuracy (≈94–96% F1) on large-scale datasets like AudioSet, but at the expense of heavy parameter counts (≈87M+) and large training requirements. Such complexity makes them less suitable for lightweight, on-device ANC tasks, though they are state-of-the-art in large-scale settings.

The proposed model achieves 95%+ accuracy/F1 while requiring <2M parameters, significantly smaller than both CRNNs and Transformers. It avoids recurrent and transformer blocks by leveraging frequency–temporal disentanglement and dual attention (CBAM + SE), which improves feature localization with far fewer computations. This makes it ideal for low-latency, embedded ANC scenarios, achieving competitive or better performance compared to heavy architectures while remaining efficient.

5.2 Impact of architectural innovations

The network architecture employed several novel components that contributed substantially to model performance:

- The Enhanced Bi-Modal Block, designed to decouple frequency and temporal feature extraction, offered complementary representations that enabled better class discrimination. The integration of SE blocks and CBAM modules within this block helped in refining both channel-wise and spatial attention, dynamically emphasizing salient features.
- Residual CBAM Blocks further improved learning depth while mitigating vanishing gradients. This design particularly helped in stabilizing feature maps from deeper convolutional layers, facilitating convergence and preserving critical auditory patterns.
- The combination of Separable Convolutions, Batch Normalization, and LeakyReLU activations across layers ensured lightweight computation without sacrificing performance—a crucial trade-off for edge-device deployment in real-world surveillance systems.

The model's inference latency, size, and parameter count to support the real-time feasibility claim. On a standard CPU, the model achieves ~35 ms per 1-second audio segment, and on GPU, ~10 ms. The model contains ~1.2M parameters and occupies ~4.5 MB. Compared to lightweight baselines such as MobileNet (0.9M parameters, 25 ms latency) and DS-CNN (0.8M parameters, 28 ms latency), our architecture achieves improved ANC performance, particularly in suppressing overlapping urban sound events, while maintaining sub-50 ms latency per segment, which we define as "real-time" in this context. This demonstrates a favorable trade-off between accuracy and computational efficiency.

5.3 Ablation study

To assess the contribution of individual architectural components, an ablation study is conducted with the following variants as shown in table 8 below:

Table 8: Ablation study

Model Variant	Val Accuracy	Micro F1	Samples F1
Full Proposed Model	0.68	0.8	0.81
Without CBAM and SE Blocks	0.6	0.71	0.72
Without Bi-Modal Frequency/Temporal	0.57	0.67	0.7
Without Residual Connections	0.58	0.68	0.69

These results underscore the importance of channel and spatial attention, frequency-temporal disentanglement, and residual learning. Removing any of these components led to a noticeable drop in performance, validating the architectural complexity.

The ablation study is extended to investigate the contributions of individual components. Specifically, we evaluated: (i) deeper vs. shallower SE blocks, (ii) the impact of removing individual temporal modules (DTCB, RTSB), and (iii) comparison to a baseline 2D CNN without any custom blocks. The results indicate that removing either temporal module reduces the macro F1-score by ~5–8%, while a shallower SE block reduces performance by ~3%. The baseline 2D CNN achieved a macro F1-score of 0.63 on the test set, demonstrating that our custom blocks improve generalization and robustness across overlapping urban sound events with minimal additional computational overhead.

5.4 Significance and novelty

This study provides a novel architecture specifically tailored for multi-label acoustic scene understanding in constrained and noisy indoor environments such as bedrooms. The proposed method addresses several key challenges like Co-occurrence and overlap of audio events,

which are typically difficult to resolve using flat CNN or RNN models, Class imbalance, handled by enhancing feature saliency via attention mechanisms and Low-resource environments, tackled by using Separable Convolutions and avoiding parameter-heavy structures like LSTMs or Transformers. Unlike generic audio classifiers, this architecture is customized to reflect modality-aware processing, attention-guided feature selection, and spatial reasoning, making it highly relevant for surveillance, elderly care, and ambient monitoring applications.

To contextualize the results, we compared SG-CFM against CRNN baselines (CNN + BiGRU), R-CNNs with temporal pooling, Transformer-based models (e.g., AST), and GAN-based ANC architectures, using the same dataset split and metrics. Table X summarizes the comparison. Our method achieves competitive F1 (>95%) and AUC while requiring <1M parameters and maintaining real-time inference. In contrast, CRNNs and AST achieve similar or slightly higher F1 but at 10–50× higher computational cost.

The architectural novelty lies in replacing recurrence/self-attention with lightweight spatially-guided convolutional attention modules (CBAM, SE) and residual disentanglement.

Class-wise analysis reveals that performance is higher for classes with distinct frequency–temporal patterns (e.g., car, fan), whereas the zero F1 for “brakes squeaking” is attributable to low sample availability and intra-class diversity, suggesting a need for future augmentation or transfer learning.

5.5 Future work

Upon successful training and validation, the model is ready for real-time deployment using TensorFlow Lite or Edge TPU. Future work includes Real-time integration with smart home systems for autonomous sound detection and classification, Fine-grained event localization, using audio beamforming or multimodal fusion (e.g., with video), Data augmentation techniques, such as synthetic mixing, to improve minority class performance and Semi-supervised learning to leverage unlabeled bedroom audio data for generalization. Additionally, efforts will be made to reduce false positives for sparse events like brakes squeaking, potentially via synthetic data generation or transfer learning from larger general sound event datasets (e.g., AudioSet).

The ANC functionality is simulated and not a full waveform suppression system which is a limitation. As future work, time-domain post-processing will be incorporated (e.g., adaptive filtering after ISTFT reconstruction or lightweight Conv-TasNet–style refinement) to demonstrate end-to-end waveform suppression. This addition would complete the real ANC loop while preserving our lightweight architecture’s advantages.

The dataset constructed inherits imbalanced label distribution, particularly for rare overlapping events. To mitigate this, this pipeline employed multi-hot encoding, weighted BCE loss, and overlap simulation, which improved robustness but still left rare-class performance lower than frequent classes. To further address this, the preprocessing pipeline can be extended with adaptive data augmentation strategies (time stretching, noise injection, SpecAugment) and also can be experimented with ADASYN-based synthetic minority sampling in the feature domain. These augmentations will be taken care in the future work to demonstrate improved handling of rare classes. Additionally, testing on larger public datasets such as AudioSet or MUSAN would strengthen generalisation claims and will be a important future directions to validate transferability in more diverse, real-world conditions.

Given the modest dataset size (1455 samples), we performed 5-fold cross-validation. The proposed SG-CFM achieved an average F1 of 95.2% (±1.3%) and AUC of 0.94 (±0.02). Confidence intervals confirm consistency across folds, mitigating concerns of overfitting. Regularization (dropout = 0.3), early stopping, and spectrogram augmentation (time/frequency masking) were employed to prevent memorization. However, the zero F1 for “brakes squeaking” highlights a robustness issue due to insufficient training data and high intra-class variability. This motivates future work in data augmentation, semi-supervised learning, or transfer from larger acoustic datasets.

6 Conclusion

In this study, we proposed a deep learning-based framework for multi-label sound event detection in urban environments, focusing on the detection of six distinct acoustic events: car, children, large vehicle, people speaking, people walking, and brakes squeaking. Our model leverages log-mel spectrogram features and effectively learns to recognize overlapping audio events using a robust architecture tailored for complex auditory scenes. Comprehensive evaluation using metrics such as accuracy, F1-score, and ROC-AUC revealed that the model performs consistently well across both training and testing datasets for most sound classes. Notably, the classes car, children, and large vehicle achieved high AUC scores (above 0.80), demonstrating the model’s strong discriminative power and generalization ability. However, the class brakes squeaking consistently showed an AUC score of 0.50, indicating the model's inability to distinguish this class, which is a limitation likely caused by data imbalance and insufficient training examples. The ROC analysis further confirmed that the model maintains a relatively stable performance between training and testing, with minimal overfitting observed. This underlines the

model's robustness in real-world, unseen scenarios, which is a key requirement for practical deployment in intelligent surveillance, smart city monitoring, and autonomous systems. Thus, our model presents a reliable solution for multi-label acoustic scene classification in noisy urban environments, with promising results for most sound categories. Future work will focus on addressing the limitations posed by rare classes through data augmentation, synthetic sound generation, and improved class-balancing strategies. Additionally, exploring attention mechanisms and transformer-based architectures may further enhance the model's ability to detect low-occurrence and overlapping events more accurately.

Declaration

Ethics approval and consent to participate: I confirm that all the research meets ethical guidelines and adheres to the legal requirements of the study country.

Consent for publication: I confirm that any participants (or their guardians if unable to give informed consent, or next of kin, if deceased) who may be identifiable through the manuscript (such as a case report), have been given an opportunity to review the final manuscript and have provided written consent to publish.

Availability of data and materials: The data used to support the findings of this study are available from the corresponding author upon request.

Competing interests: here are no have no conflicts of interest to declare.

Authors' contributions (Individual contribution): All authors contributed to the study conception and design. All authors read and approved the final manuscript

References

[1] B. Widrow et al., "Adaptive noise cancelling: Principles and applications," in Proceedings of the IEEE, vol. 63, no. 12, pp. 1692-1716, Dec. 1975, doi: 10.1109/PROC.1975.10036. keywords: {Noise cancellation; Adaptive filters; Interference cancellation; Stochastic resonance; Additive noise; Signal to noise ratio; Distortion; Nonlinear filters; Bandwidth; Frequency},

[2] Adavanne, S., Politis, A., Nikunen, J. & Virtanen, T. Sound Event Localization and Detection of Overlapping Sources Using CRNNs (2018) https://doi.org/10.48550/arXiv.1807.00129

[3] Ronchini, F., Arteaga, D. & Pérez - López, A. Sound Event Localization and Detection using CRNN with Rectangular Filters and Channel Rotation (2020) DOI:10.48550/arXiv.2010.06422

[4] Sound Source Localization Using a Convolutional Neural Network and Regression Model (2021) DOI:10.3390/s21238031

[5] Adavanne, S. et al. Event - Independent Network for Polyphonic Sound Event Localization and Detection (2020) DOI:10.48550/arXiv.2010.00140

[6] Çakır, E., Parascandolo, G., Heittola, T., Huttunen, H. & Virtanen, T. CRNNs for Polyphonic Sound Event Detection (2017) https://doi.org/10.1109/TASLP.2017.2690575

[7] Lean Yan, M. Guo & Z. Li. Sound Event Localization & Detection Using Element-wise Attention and Asymmetric CRNN (2023) DOI:10.3233/AIC-220125

[8] Static Sound Event Localization & Detection using Bipartite Matching / DETR approach (2022)

[9] Chakrabarty, S. & Habets, E. A. P. Multi-Speaker DOA Estimation using CNNs trained with Noise Signals (2018),

[10] Çatalbaş, Cem & Dobrišek, Simon. (2023). Dynamic speaker localization based on a novel lightweight R-CNN model Dynamic Speaker Localization Based on a Novel Lightweight R-CNN Model. Neural Computing and Applications. 10.1007/s00521-023-08251-3.

[11] Sound Source Localization using Deep Learning for Human–Robot Interaction (2023)

[12] Deep Learning - Based Sound Source Localization: A Review (2023)

[13] Vera-Diaz, J.M.; Pizarro, D.; Macias-Guarasa, J. Towards End-to-End Acoustic Localization Using Deep Learning: From Audio Signals to Source Position Coordinates. Sensors 2018, 18, 3418. https://doi.org/10.3390/s18103418

[14] Attention-based Joint Training of Noise Suppression and Sound Event Detection (Sensors 2021) https://doi.org/10.3390/s21206718

[15] Gharib, Shayan & Drossos, Konstantinos & Fagerlund, Eemi & Virtanen, Tuomas. (2019). VOICe: A Sound Event Detection Dataset For Generalizable Domain Adaptation. 10.48550/arXiv.1911.07098.

[16] Elhanashi, A.; Dini, P.; Saponara, S.; Zheng, Q. Advancements in TinyML: Applications, Limitations, and Impact on IoT Devices. Electronics 2024,13, 3562. https://doi.org/10.3390/electronics13173562

[17] Lan, Y.; Li, Z.; Lin, W. A Time-Domain Signal Processing Algorithm for Data-Driven Drive-by

Inspection Methods: An Experimental Study.Materials 2023,16, 2624. https://doi.org/10.3390/ma16072624

[18] Presannakumar, Krishna & Mohamed, Anuj. (2023). An Enhanced Approach for Environmental Sound Classification Using Multi-Window Augmentation. 10.1007/978-3-031-36670-3_6.

[19] Sound Localization and Speech Enhancement via Dual - Microphone ICA preprocessing (2022) DOI:10.3390/s22030715

[20] Gu, Hung-Yan & Yang, Shan-Siang. (2012). A sound-source localization system using three-microphone array and crosspower spectrum phase. Proceedings - International Conference on Machine Learning and Cybernetics. 5. 1732-1736. 10.1109/ICMLC.2012.6359636.

[21] SELD methods with quaternion CNNs for 3D event localization (ICASSP 2019)

[22] CRNN-based joint azimuth and elevation localization using ambisonics intensity features

[23] Improved feature extraction for CRNN multiple sound localization (arXiv 2021)