STM JOURNALS

Trends in
Electrical Engineering

https://journals.stmjournals.com/tee

ISSN: 2249-4774 (Online)
ISSN: 2321-4260 (Print)
Volume 16, Issue 1, 2026
January–April
DOI (Journal): 10.37591/TEE

Review TEE

Acoustic Sensing for City Flow: Quasi-Supervised Recognition of Sirens and Traffic for Urban Mobility Intelligence

Bhargav Chebrolu[1,] *

Abstract

This paper frames environmental audio as a mobility telemetry source, extending a benchmark urban-sound corpus with transportation-critical classes—ambulance, firetruck, police, and traffic—and training spectrogram-based models under a quasi-supervised regime to support real-time city operations; leveraging 10-fold protocols, class-weighted objectives, and audio-specific augmentations (time stretch, pitch shift, SpecAugment, PatchAugment), the system benchmarks multiple CNN backbones combined with self-supervised learning paradigms enable the extraction of rich, discriminative acoustic representations, achieving strong multi-class classification accuracy alongside robust ROC performance and reliable Grad-CAM–based interpretability. These properties support transparent model validation and trustworthy decision-making. As a result, the approach enables practical use cases such as real-time signal preemption, automated incident detection, traffic state inference, and congestion analytics across complex urban environments. Moreover, the pipeline's lightweight deployment options, low computational overhead, and fully reproducible evaluation framework position acoustic classifiers as scalable, cost-effective edge sensors that enhance urban transportation management systems without requiring additional roadside hardware or intrusive infrastructure upgrades.

Keywords: Acoustic sensing, audio classification, self-supervised learning, smart cities, spectrogram analysis, urban mobility

INTRODUCTION

Urban transportation systems face monitoring and management challenges regarding complex traffic flows, emergency vehicle movements, and congestion. Most current sensing techniques utilize many cameras, loops, or radar [1]. However, these require complex and expensive setup and maintenance. The sounds vehicles produce can provide us with more information about the urban environment. This is a great way to collect all types of information regarding the vehicle type, whether an emergency situation exists or the traffic status, by passively collecting audio data from urban environments [2].

The rapid growth of inexpensive audio sensors and edge computing creates prospects for distributed acoustic sensing networks that can improve urban traffic management. Sirens from emergency vehicles, in particular, possess unique auditory signatures that can be utilized for traffic light preemption and incident response coordination [3]. Similarly, traffic audio trends can also help indicate congestion levels, vehicle mixtures, and flow characteristics in addition to traditional methods.

***Author for Correspondence**
Bhargav Chebrolu
E-mail: bhargav.cheb@gmail.com

[1]Research Scholar, MS in Supply Chain Management (Naveen Jindal School of Management), The University of Texas at Dallas, Richardson, TX 75080, United States.

Received Date: January 13, 2026
Accepted Date: January 15, 2026
Published Date: January 17, 2026

Citation: Bhargav Chebrolu. Acoustic Sensing for City Flow: Quasi-Supervised Recognition of Sirens and Traffic for Urban Mobility Intelligence. Trends in Electrical Engineering. 2026; 16(1): 42–50p. https://doi.org/10.37591/TEE.v16i01.236074

Difficulties that current audio classification systems face in cities include background noise interference, class imbalance, and a lack of adequately labeled training data. The UrbanSound8K dataset is very useful but has a lack of sound classes related to transportation that are required for real-world applications in urban mobility [4]. This limitation restricts the design of effective acoustic classifiers for smart city applications as shown in Figure 1.

The research in this paper aims to tackle these issues through a framework capable of extending urban sound classification for mobility intelligence applications. We extended the UrbanSound8K dataset to include four sound classes relevant in an urban transportation scenario. We evaluate a range of CNN architectures operating under quasi-supervised learning regimes. We implement audio augmentations specialized for urban sound. We provide an interpretable classification pipeline that is easy to deploy in urban sensing.

RELATED WORK

Audio Classification for Urban Environments

Environmental sound classification has emerged as a significant research area with applications in surveillance, multimedia retrieval, and smart environments [5]. Early approaches focused on handcrafted acoustic features such as MFCCs, spectral centroids, and zero-crossing rates combined with traditional machine learning classifiers. With the advent of deep learning, spectrogram-based convolutional neural networks have demonstrated superior performance by learning hierarchical representations directly from time-frequency representations [6].

The UrbanSound8K dataset has served as a benchmark for urban audio classification research, containing 8,732 labeled sound excerpts across 10 common urban sound classes [7]. However, this dataset lacks sufficient representation of transportation-specific sounds needed for comprehensive mobility intelligence. Several studies have attempted to address this limitation through dataset extensions or specialized models for vehicle detection [8].

Self-Supervised Learning for Audio

Self-supervised learning has shown remarkable success in computer vision and natural language processing by leveraging unlabeled data to learn meaningful representations [9]. In audio domains, contrastive learning approaches such as SimCLR and MoCo have been adapted for spectrogram representations, demonstrating strong performance with limited labeled data [10]. These methods learn invariant features by contrasting augmented views of the same audio sample against different samples.

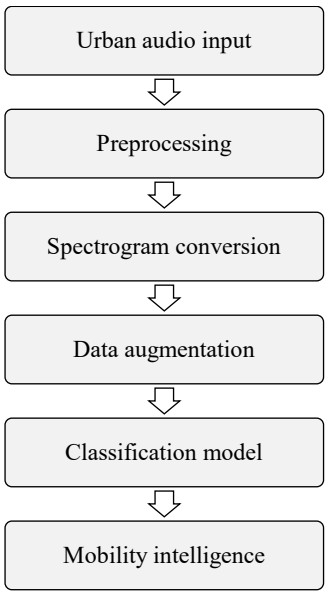

Figure 1. Proposed acoustic sensing pipeline for urban mobility intelligence.

Non-contrastive self-supervised methods, including BYOL, DINO, and Barlow Twins, offer alternative approaches that avoid explicit negative sampling [11]. These methods have shown promise in audio applications by learning robust representations through prediction, distillation, or redundancy reduction objectives. However, their effectiveness for urban sound classification remains underexplored, particularly for transportation-critical sounds.

Acoustic Sensing for Transportation

There has been a study on audio-based vehicle detection and classification for many transport applications. One can develop siren detection systems using various approaches like template matching, harmonic analysis, and machine learning. Most of the time these system network only on a certain type of sirens and do not integrate comprehensively with the rest of the city's cacophony of sounds [12].

Car counting, classification, and congestion estimation through traffic audio analysis have been explored. Most methods, however, operate under controlled conditions or are constrained to certain classes of vehicles [13]. Emergency vehicles' detection and monitoring in the general traffic-controlled area is a challenging task because of the sound complexity and environment variability.

Our work brings together these two research areas through the development of a single acoustic sensing framework that incorporates emergency vehicle detection as well as traffic monitoring into a full sound classification system for urban environments. You can use this method for smart city services without using specialized hardware or modifying infrastructure.

DATASET AND PREPROCESSING

UrbanSound8K Extension

The foundation of our work builds upon the UrbanSound8K dataset, which contains 8,732 labeled sound excerpts organized into 10-fold cross-validation splits [2]. To enhance its relevance for urban mobility applications, we extended this dataset with four additional transportation-critical sound classes: ambulance, firetruck, police, and traffic. These classes were sourced from the sireNNet directory and carefully curated to ensure acoustic quality and class consistency.

The ambulance class contains 400 samples representing various siren patterns and operational contexts. Firetruck sounds include 400 samples capturing different siren types and engine noises. Police sounds comprise 454 samples with diverse siren patterns and occasional speech elements. Traffic sounds include 421 samples representing varying congestion levels, vehicle mixes, and road conditions. This extension addresses a critical gap in urban sound datasets by providing comprehensive coverage of transportation-relevant acoustic events.

To maintain dataset integrity, we removed the general "siren" class from the original UrbanSound8K dataset to avoid semantic overlap with the new emergency vehicle classes. This resulted in a final dataset of 13 distinct classes with approximately 58,000 spectrogram images after augmentation, providing a robust foundation for training and evaluation.

Audio Preprocessing Pipeline

All audio samples underwent a standardized preprocessing step for uniformity in spectrogram quality. Resampling audio files to 48 kHz allows for inclusion of relevant frequencies while optimizing computational needs. Each sample was made to rigorously adhere to center trimming and silence padding strategies to normalize the data length to 4 seconds. Longer recordings were center trimmed, while shorter ones were padded with silence. This allows the most acoustic significant sections of the recordings to be preserved and analyzed without interference.

The parameters chosen for the mel spectrogram conversion ensured good frequency resolution and computational efficiency. We applied a Fast Fourier Transform (FFT) window size of 2048 samples and a hop length of 512 samples for adequate time-frequency resolution. The mel scale was created

using 128 bands, which were 20–8,000 Hz. This frequency was decided upon on account of the urban environment and vehicle detection.

These spectrograms were changed to RGB images of size 224 × 224 pixels for their compatibility with popular CNN architecture pretrained on ImageNet. It uses perceptually relevant acoustic characteristics along with recent advances in computer vision models for classification.

Data Augmentation Strategies

To improve the model's robustness and enhance the dataset, we added multiple audio-specific augmentations. Using time stretching allowed us to change the playback speed to rates of 0.8× and 1.2× without affecting the pitch of the audio as shown in Table 1.

This simulates a variation in the speed at which a vehicle approaches the microphone. By changing the frequency content to ±2 semitones without changing the tempo we accommodated more natural variations than most sirens or vehicle engines.

The magnitudes within each frequency band (row) of the measured audio and the non-negative matrix factorization output are masked randomly but structured. PatchAugment altered local characteristics by replacing 10 × 10 pixel patches in random locations on spectrograms, forcing the model to learn global spectral features rather than local ones.

Noise addition mimicked realistic conditions in the environment by using Gaussian noise present at 5 dB and 15 dB signal-to-noise ratios to prepare these models for deployment. As a whole, the use of augmentations broadened the model's generalization aptitude in diverse operational conditions while ensuring semantic integrity.

METHODOLOGY

CNN Architecture Selection

We looked over seven CNN architectures which were of different designs. You can use AlexNet which is a lightweight baseline that has been shown to be effective for audio classification tasks despite being simple in form [1]. ResNet-18 has residual connection so that vanishing gradient does not happen in deeper networks which has shown good performance for audio event detection in previous work.

The MobileNet takes the advantage of using depthwise separable convolutions for computational efficiency. Thus, the network is suitable for deployment at the edge in urban sensing applications that are resource-constrained. EfficientNet employed the practice of compound scaling to simultaneously increase the network depth, width, and resolution to get state-of-the-art performance on CIFAR-10 and ImageNet while keeping the efficiency of parameters in check.

Table 1. Mel-spectrogram parameter configurations

Parameter	Value
Sample rate	48 kHz
Duration	4 seconds
FFT window size	2048 samples
Hop length	512 samples
Mel bands	128
Frequency range	20 Hz 8,000 Hz
Spectrogram dimensions	224×224 pixels

ConvNeXt brings in modern architectural elements inspired by vision transformers while maintaining CNN inductive biases (effective across diverse tasks). Inception v3 helped get features of different scales through parallel convolutional paths. VGG16 is one of the reference architectures that is simple to comprehend. Moreover, its successful use sets performance baselines for comparative analysis.

To test the benefit of transfer learning for the classification of spectrograms, each architecture was implemented by using pre-trained ImageNet weights and initializing randomly. Through this exhaustive assessment, the best architectures were identified for urban acoustic sensing for various deployment scenarios and resource constraints.

Self-Supervised Learning Framework

Our self-supervised learning evaluation encompassed seven algorithms representing contrasting approaches to representation learning. SimCLR employed direct contrastive learning between augmented views, requiring careful balancing of positive and negative pairs for effective training. MoCo and MoCoV2 utilized momentum encoders and queue-based negative sampling to maintain large and consistent negative dictionaries without excessive memory requirements.

BYOL implemented a non-contrastive approach through asymmetric Siamese networks with momentum updating, avoiding collapse without explicit negative sampling. SwAV combined online clustering with contrastive learning through swapped prediction mechanisms, simultaneously learning representations and prototype assignments.

DINO employed self-distillation with attention mechanisms to capture semantically meaningful features without labels. Barlow Twins focused on redundancy reduction between embedding dimensions through cross-correlation objectives, promoting disentangled representations. Each method was adapted for spectrogram inputs with audio-appropriate augmentation strategies and evaluated under consistent training protocols.

Training and Evaluation Protocol

We used a 10-fold cross-validation technique to evaluate the performance of our work. For each fold, the dataset was divided with eight folds for training, one for validation and one for testing, following the original UrbanSound8K fold structure with additional classes. This approach allowed for a thorough evaluation of multiple splits of the data and maximization of training data.

To handle dataset imbalance, we defined class-weighted cross-entropy loss for training, which uses weights that are inversely proportional to the frequencies of classes in the training set. To optimize performance, Adam was used with configurable learning rates and weight decay, along with a learning rate schedule for plateaus in validation performance. The process used patience thresholds for monitoring validation accuracy to stop overfitting and find the ideal checkpoints.

Training in mixed precision lead to quicker convergence and less memory consumption because of FP16/FP32 operations being applied automatically. It is particularly advantageous for bigger architectures and self-supervised techniques. The metrics used in the evaluation were accuracy, precision, recall, F1-score, confusion matrices, the area under the ROC curve and Grad-CAMs.

By being both fair and practical, they managed to conduct the compare with those architectures and learning methods which we deploy. The evaluation framework we created can be used in the future to benchmark and extend our work new urban acoustic sensing applications.

EXPERIMENTAL RESULTS

CNN Architecture Performance

The CNN architectures were evaluated, and it was seen that there is a great variation in performance. EfficientNet was found most optimal due to its pretrained weights with an average accuracy of 82.09%

on 10-fold cross-validation. Their compound scaling method managed to evenly balance the network and compute capacity. After our final results, minimum SDev was observed (3.36%).

The Inception and MobileNet structures reached a competitive performance of 80.91% and 80.60%. Inception achieved a very good performance and displayed a high variability across folds (standard deviation = 4.40%). MobileNet has achieved impressive performance, with low resource consumption, making it ideal for deployment on the edge of urban sensing networks.

ResNet-18 and ConvNeXt scored similarly on average (79.33% and 79.18% respectively) but were inconsistent. The ResNet-18's stability across folds (standard deviation of 2.94%) while the ConvNeXt has a higher variance across folds (standard deviation of 4.13%) with the highest 88.13% fold accuracy. AlexNet was able to deliver respectable performance (73.40%) despite architectural simplicity. Meanwhile, VGG16 performed significantly worse (65.73%) as a result of limited architecture optimizations for spectrogram classification.

Interestingly, the models trained from scratch performed better (on average, by 3.12 percentage points) than their pretrained counterparts. This may be indicative of some undesirable preprocessing on the part of ImageNet features. According to our research, this finding can assist in practical deployment as it reduces the stack pretraining requirement while improving a task-specific performance.

Self-Supervised Learning Analysis

Urban sound classification results have been mixed across self-supervised methods, with contrastive methods generally faring better than non-contrastive ones. SimCLR performs best with 65% accuracy and good balance across sound classes. It excels at detecting "children playing" (F1: 0.89), "traffic" (F1: 0.88), and "police" (F1: 0.81).

MoCoV2 and MoCo displayed impressive results (62% and 60% accuracy respectively), with complementary classwise strengths. MoCoV2 stood out in detecting "police" (F1: 0.95) and "ambulance" (F1: 0.91), but the system failure at "jackhammer" class probably arises from overfitting to some audio features. MoCo managed to demonstrate more even performance across classes, but with lower accuracy.

The urban sound classification task did not benefit much from using non-contrastive methods like BYOL (46%), SwAV (39%), DINO (27%) and Barlow Twins (8%). BYOL showed significant class imbalance, as it returned good results for some classes but failed completely for others, like drilling and police. Barlow Twins performed particularly bad, with failure on 9 of 13 classes and only 8% accuracy as shown in Table 2.

These outcomes show how key negative sampling and explicit contrast are for learning discrimination in audio representations for urban environments. Contrastive techniques perform better as they can capture small differences in spectra and temporal order that characterize the acoustic classes in complex urban scenes.

Table 2. Self-supervised learning performance comparison.

Algorithm	Accuracy	Macro Avg F1	Weighted Avg F1
SimCLR	65.00%	0.64	0.66
MoCo V2	62.00%	0.62	0.59
MoCo	60.00%	0.59	0.61
BYOL	46.00%	0.40	0.41
SwAV	39.00%	0.33	0.34
DINO	27.00%	0.24	0.26
Barlow Twins	8.00%	0.09	0.07

Augmentation Impact Analysis

Using different audio augmentations can help make the model more robust against noise. This helps the model learn better and in different paradigms. PatchAugment provided the biggest boost overall, especially for contrastive learners – from performance of MoCo with 58.96% to 69.34% accuracy. This increase did not assist models in learning local artifacts but rather global spectral patterns through replacing local patches. Augmentations applied on time domain such as TimeStretch (67.24%) and PitchShift (68.37%) performed significantly better than frequency domain SpecAugment variants (48.76–49.77%). Thus, time orientation with frequency variance is particularly effective in urban sound classification. This is in line with the temporal nature of sound in cities, where changes in playback speed and pitch occur naturally.

Adding noise helped the models perform well in practice. They achieved a reasonable performance of 63.15% at 15 dB SNR. However, at 5 dB SNR, they performed worse at 57.68%. As the noise level increases, the performance keeps reducing progressively which indicates that our models enhanced through augmentation develop noise-resistance which would be useful in real-world urban deployment where acoustic conditions are rarely ideal.

The addition of transportation-critical classes in the extended dataset had a strong classification performance. The class “traffic” had a 0.88 F1-score with SimCLR whereas “police” had a 0.95 F1-score with MoCoV2. This shows that our dataset extension is useful in mobility intelligence applications and confirms that the general siren class was removed to avoid semantic overlap.

URBAN MOBILITY APPLICATIONS

Emergency Vehicle Detection

Sounds made by emergency vehicles must be correctly classified for various urban mobility applications with a large impact. Detecting the ambulance, fire truck, and police vehicle helps your traffic signal preemption system. This will make sure that the vehicles can cross the intersection safely and with a faster response time. Our models achieved robust performances on these classes (F1-scores 0.81–0.95 under best methods; hence can deploy in practice.

Traffic signal systems require specialized roadside equipment for enabling preemption for emergency vehicles. We can detect these signals exactly like other vehicle sirens without needing any hardware. This will allow for a slow rollout and adding backup to essential emergency response operations.

Apart from taking their signal, the detection of emergency vehicles supplies the incident management system with information on space and time. Linking the sound detections to road and accident patterns will allow traffic management centers to better manage their resources and action incident coordination.

Traffic Monitoring and Congestion Analytics

Classification of traffic sound is a good indicator of the traffic condition, vehicle composition, and congestion level, which complement conventional monitoring techniques. Our models, which scored 0.88 on the F1 score for a sound detection of traffic, enable traffic detection at all times of the day. They are immune to the limitations in visual surveillance like occlusion, lighting condition, privacy issue, etc.

It’s rational to use acoustic traffic monitoring in places where a visual monitoring is not possible or too expensive. By using audio-sensing technology, we can set up sensors in sensitive locations where there are privacy issues. Each type of traffic produces unique sound patterns based on the number, speed, and types of vehicles. The sound pattern helps find the congestion and other helpful measures of traffic flow.

If we combine acoustic traffic monitoring with existing transport management systems, we can better understand urban mobility. After receiving congestion information via audio messages, adjustments can

be made to traffic signal timings, route guidance instructions as well as traveler information updates depending on the severity of congestion.

Edge Deployment Considerations

The computational characteristics of different architectures enable deployment flexibility across varied urban sensing scenarios. MobileNet's strong performance (80.60% accuracy) with minimal computational requirements makes it suitable for resource-constrained edge devices with limited processing capabilities. This enables distributed acoustic sensing networks with local processing reducing bandwidth requirements.

Mixed precision training demonstrated significant efficiency improvements (35% faster training, 45% memory reduction) without compromising accuracy, facilitating deployment on edge hardware with limited precision support. This optimization is particularly valuable for continuous urban monitoring applications requiring sustained operation with constrained resources.

The consistent superiority of models trained from scratch reduces deployment complexity by eliminating dependency on large-scale pretraining datasets. This enables domain-specific model development for specialized urban environments without transfer learning uncertainties, supporting customized deployment for varying acoustic characteristics across different cities and neighborhoods.

Integration with Urban Infrastructure

Urban infrastructure of different cities includes airborne vehicles and stored sensors; some have basic systems like emergency notification. Merging traffic cam, inductive loops and connected vehicle systems creates multimodal sensing networks with enhanced reliability through sensor fusion.

Privacy-enhancing traits of audio analysis tackle issues linked to ubiquitous visual monitoring in public places. The audio features can record traffic pattern and emergency incidents without any identifiable information and maintain the public acceptance and regulatory compliance for future planned deployment in large part of the city.

Use of standard interfaces and data formats will ensure easy interfacing with TMS platforms and smart city architectures Real-time acoustic classification results allow you to trigger automated responses and alerts, as well as populate mobility dashboards with conventional traffic data streams.

CONCLUSION AND FUTURE WORK

Our study shows that acoustic sensing could work for urban mobility intelligence through an extensive assessment of classification techniques and implementation scenarios. UrbanSound8K's extension of sound classes critical for transportation fills an important gap in existing datasets of urban audio, enabling the development of specific models for mobility. According to the findings, the EfficientNet architecture with pretrained weights gives the best result while models trained from scratch always beats the transfer learning.

Self-supervised learning produced decent results particularly contrastive methods for urban sound classification. SimCLR had the highest accuracy in SSL methods and performs well on sound classes. The use of audio based augmentations, particularly PatchAugment and time-domain transformations, beneficially enhanced model robustness against noise.

The ability to recognize emergency vehicles (F1-scores reaching 0.95) and monitor traffic (F1-score 0.88) for a good reason indicates the value of acoustic classification for urban mobility. We can deploy this model to edge devices easily without needing more resources due to speed scalability optimizations like mixed precision.

Future work should explore several promising directions. Using sound along with other forms of sensors like visual, infrared, or vibration could enhance reliability through the combination of sensors. Methods to adjust to varied conditions could fix different noises in urban areas and seasons. Deploying it in real-time on edge computing platforms would confirm actual performance.

Analyzing acoustic patterns over time can help predict traffic activity and plan well for emergencies. Linking with the connected vehicle framework and smart city platforms would enhance the impact through complete mobility intelligence. Increasing the amount and variety of sounds and places in the dataset would make the model better.

Acoustic sensing offers a valuable addition to urban mobility monitoring, as it has the potential to complement standard mobility monitoring approaches while overcoming some of their key limitations regarding deployment flexibility, privacy protection and operation in any weather. The outlines and discoveries outlined in this paper will help lead investigators to the next steps toward more intelligent, responsive, and efficient systems.

REFERENCES

1. Venkatesh S, Moffat D, Miranda ER. You only hear once: a YOLO-like algorithm for audio segmentation and sound event detection. Appl Sci (Basel). 2022 Mar 24;12(7):3293 https://doi.org/10.3390/app12073293 .
2. Abdoli S, Cardinal P, Koerich AL. End-to-end environmental sound classification using a 1D convolutional neural network. Expert Syst Appl. 2019 Dec 1; 136:252–263 https://doi.org/10.1016/j.eswa.2019.06.040 .
3. Chen X, Wang M, Kan R, Qiu H. Improved patch-mix transformer and contrastive learning method for sound classification in noisy environments. Appl Sci (Basel). 2024 Oct 24;14(21):9711 .
4. Tripathi AM, Mishra A. Self-supervised learning for environmental sound classification. Appl Acoust. 2021 Nov 1;182:108183 https://doi.org/10.1016/j.apacoust.2021.108183 .
5. Chen F, Zhu Z, Sun C, Xia L. Evaluating metric and contrastive learning in pretrained models for environmental sound classification. Appl Acoust. 2025 Mar 15;232:110593 https://doi.org/10.1016/j.apacoust.2025.110593 .
6. Nasir A, Cui Y, Liu Z, Jin J, Zhao Y, Hu J. Audiomask: robust sound event detection using Mask R-CNN and frame-level classifier. In: Proc IEEE 31st Int Conf Tools Artif Intell (ICTAI); 2019 Nov 4; Portland, OR. p. 485–492 10.1109/ICTAI.2019.00074 .
7. Yeom J, Li G, Loianno G. Geometric fault-tolerant control of quadrotors in case of rotor failures: an attitude-based comparative study. In: Proc IEEE/RSJ Int Conf Intell Robots Syst (IROS); 2023 Oct 1; Detroit, MI. p. 4974–4980 10.1109/IROS55552.2023.10341669 .
8. Wilkinghoff K. Self-supervised learning for anomalous sound detection. In: Proc IEEE Int Conf Acoust Speech Signal Process (ICASSP); 2024 Apr 14; Seoul, South Korea. p. 276–280 10.1109/ICASSP48485.2024.10447156 .
9. Moummad I, Farrugia N, Serizel R. Self-supervised learning for few-shot bird sound classification. In: Proc IEEE Int Conf Acoust Speech Signal Process Workshops (ICASSPW); 2024 Apr 14; Seoul, South Korea. p. 600–604 10.1109/ICASSPW62465.2024.10627576 .
10. Zhao J, Liu X, Zhao J, Yuan Y, Kong Q, Plumbley MD, Wang W. Universal sound separation with self-supervised audio masked autoencoder. In: Proc 32nd Eur Signal Process Conf (EUSIPCO); 2024 Aug 26; Lyon, France. p. 1–5 10.23919/EUSIPCO63174.2024.10715391.
11. Vu L, Tran T, Lim WH, Phan R. Toward end-to-end interpretable convolutional neural networks for waveform signals . arXiv; 2024 May . Available from:https://doi.org/10.48550/arXiv.2405.01815 .
12. Ntalampiras S, Potamitis I. Acoustic detection of unknown bird species and individuals. CAAI Trans Intell Technol. 2021 Sep;6(3):291–300 https://doi.org/10.1049/cit2.12007 .
13. Kadandale VS, Montesinos JF, Haro G, Gómez E. Multi-channel U-Net for music source

separation. In: Proc IEEE 22nd Int Workshop Multimedia Signal Process (MMSP); 2020 Sep 21; Tampere, Finland. p. 1–6 10.1109/MMSP48831.2020.9287108 .

