Acoustic Non-Line-of-Sight Vehicle Approaching and Leaving Detection

Mingyang Hao, Fangli Ning, Ke Wang, Shaodong Duan, Zhongshan Wang, Di Meng, and Penghao Xie

***Abstract*— Early detection of occluded moving vehicles at intersections can prevent collisions, but most of the existing sensors only detect objects in sight. Sound propagates around obstacles by means such as reflection and diffraction, allowing passive acoustic sensing of the occluded vehicle. We propose a deep learning-based acoustic non-line-of-sight (NLOS) vehicle detection method. With the direction-of-arrival feature and time-frequency feature calculated from the microphone array data as inputs in image form, we designed and trained the parallel neural network to perceive the direction of the occluded moving vehicle at intersections. Since intelligent vehicles react differently to the approaching and leaving occluded moving vehicle, we further distinguished the occluded moving vehicle's approaching/leaving status. To evaluate the proposed method, we collected data from different locations in the urban environment. The experimental results show that the classification for 6-class intersection traffic conditions reached 96.71%, and the occluded approaching vehicle was detected 1 second before it entered the line of sight, providing additional reaction time for intelligent vehicles. The direction of the occluded moving vehicle is accurately predicted, and the approaching/leaving status is further inferred, providing detailed traffic information for the intelligent vehicles' response decisions. Furthermore, experiments show that the predictions of our method outperform the state-of-the-art acoustic NLOS approach vehicle detection baseline on real-world traffic datasets. Our code and dataset: https://github.com/RST2detection/Acoustic-Occluded-Vehicle-Detection.**

***Index Terms*— Acoustic traffic perception, sound event detection, intelligent vehicle, deep learning, non-line-of-sight.**

I. Introduction

MODERN sensor systems widely used on vehicles require a direct line-of-sight (LOS) channel to capture dynamic obstacles/vehicles [1]. However, when the sight is obscured by trees, parked vehicles on the roadside, and buildings at intersections, the sensing systems that vehicles rely primarily on lack awareness of non-line-of-sight (NLOS) areas [2], which may pose a danger to undetected road users [3]. NLOS methods aim to perceive the occluded objects from the information they indirectly reflect on the visible surfaces. The perception of dynamic objects in NLOS scenes applies to various domains, including robotics and remote sensing. In particular, it can be employed in autonomous driving applications to detect hidden road users, prevent collisions, and enhance safety [4].

Although researchers have experimented on different sensors, the existing NLOS object detection methods only work under specific conditions and do not apply to complex automatic driving scenarios. Camera-based NLOS sensing [5] works under limited conditions with optical reflections or additional illumination [6], [7], [8]. The LiDAR-based specialized device is difficult to promote due to its expensive price [9]; Since at most frequencies, the through-wall radar [10] cannot pass through obstacles made of metal or concrete, obstacles on the road are opaque to the radar; Additionally, the use of Vehicle to Everything (V2X) or Vehicle to Vehicle (V2V) cooperative sensing [11] requires sufficient vehicle-to-road collaborative software development and infrastructure construction, which cannot satisfy in the near/medium term. Therefore, it is necessary to explore new methods for detecting NLOS vehicles.

Even if there are obstacles between the sound source and the microphone, the sound wave emitted from the source reaches the microphone after various wave interactions (reflection, interference, diffraction, etc.) [12], which means that the invisible vehicle is potentially detectable by acoustic sensors. The information in the cross-power spectrum phase coefficients enables robust detection of multiple approaching vehicles behind corners [13]. To localize the NLOS vehicle, the localization problem of uniformly moving sound sources was transformed into solving a set of equations [14]. Furthermore, a diffraction theory-based NLOS vehicle localization method was proposed in [15]. However, for modeling the effect of occlusion on acoustic sensor measurements, these methods [13], [14], [15] rely on strong assumptions without regard for the impact of other propagation modes. In contrast to these efforts, recent work based on machine learning indicates that the data-driven method can predict the approaching vehicle behind corners in advance [2]. However, there are still three issues that remain unresolved in [2]. Firstly, in the adopted steered-response power phase transform (SRP-PHAT) method,

Manuscript received 4 April 2023; revised 29 September 2023 and 7 December 2023; accepted 2 January 2024. Date of publication 26 January 2024; date of current version 1 August 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 52075441, in part by the Shaanxi Key Research Program Project 2020ZDLGY06-09 and Project 2023-YBGY-219, in part by the Aeronautical Science Foundation of China under Grant 20200015053001, and in part by the Xi'an Key Industrial Chain Technology Research Project 23ZDCYJSGG0006-2023. The Associate Editor for this article was M. Yang. *(Corresponding author: Fangli Ning.)*

The authors are with the School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: mingy_hao@163.com; ningfl@nwpu.edu.cn; kewang5@mail.nwpu.edu.cn; duansd0516@163.com; wzs040101@163.com; di_meng@mail.nwpu.edu.cn; xpenghao0902@163.com).

Digital Object Identifier 10.1109/TITS.2024.3353749

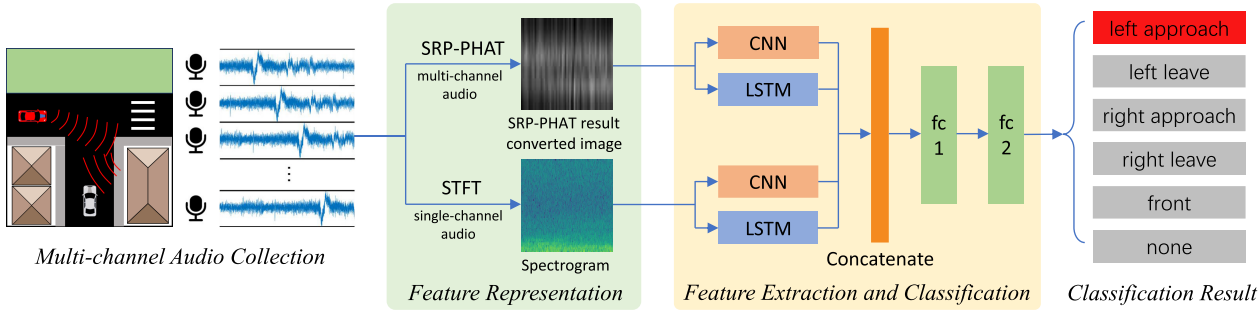


Fig. 1. Overview of the proposed NLOS vehicle approaching and leaving detection system.

only a frequency bandpass of [50, 1500] Hz is performed [2], which is unable to confirm whether the sound comes from the road user of interest. How can it be determined whether the sound is coming from the detected vehicle? Secondly, the direction of arrival (DOA) estimation results are computed only twice on each audio sample of 1 s duration [2], which breaks the temporal continuity of the DOA estimation results. How can the evolution of the DOA estimation results be coherently represented? Thirdly, acoustic NLOS vehicle detection and acoustic NLOS vehicle localization belong to two tasks. How to combine the two tasks to provide acoustic NLOS vehicle detection result with directional estimation? Furthermore, existing acoustic methods only detect the approach of the NLOS vehicle [2], [13]. In contrast, the NLOS moving vehicle has two relative states, approaching and leaving, and intelligent vehicles respond differently to these two types of traffic.

To solve the above problems, this work proposes a novel passive acoustic method for NLOS vehicle approaching and leaving detection. As shown in Figure 1, the microphone array captures ambient audio as the intelligent vehicle enters the intersection from the bottom. Unlike existing methods [2], [13], we consider the NLOS vehicle approaching and leaving as separate sound events [16], [17] at the feature representation stage. Firstly, the spectrogram is used as a feature representation for NLOS vehicle detection. Secondly, the continuous SRP-PHAT results are represented as spectrogram-like 2D images, thus transforming the NLOS vehicle localization task into an image classification problem. Therefore, at the feature extraction and classification stage, to combine the acoustic NLOS vehicle detection task and localization task, we construct a two-branch parallel CNN-LSTM (pCRNN) network with spectrogram and SRP-PHAT result converted image as input. The final output of pCRNN represents the six traffic types at the intersection. The contributions of this study are highlighted as follows.

1) We propose a novel two-branch neural network pCRNN. pCRNN combines acoustic NLOS vehicle detection task and localization task using spectrogram and DOA estimation results as inputs. Unlike existing studies [2], [13], [14], [15] based only on sound source localization information, the spectrogram is introduced to explore the application of diverse acoustic features.

2) We construct a new acoustic NLOS vehicle approaching and leaving detection dataset, filling a gap in existing publicly available dataset [2] that do not contain data on the vehicle leaving intersections.

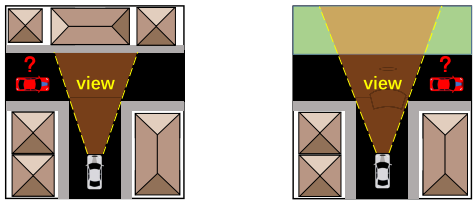


Fig. 2. Schematics of the two T-junction types. The intelligent vehicle approaches the junction from the vertical road, Type A has a significant reflective surface directly in front of the intelligent vehicle, but type B does not. Note that the rules for classifying T-junctions follow [2]. Another vehicle on the crosswise road may be approaching or leaving from the left or right blind corner.

3) We study the influence of time window length and data augmentation operations on the performance of the proposed method.

The following section (Sec. II) introduces the details of our method. Section III presents the evaluation of the predictive performance of our pCRNN by real-world datasets. Finally, conclusions are given in Section IV.

II. Approach

Using only passive acoustic sensors, the goal of our passive acoustic NLOS vehicle detection method is to detect the occluded moving vehicle in advance, including estimating its direction and predicting its approach or leave status.

After a short time of acoustic measurement of the intelligent vehicle near the intersection, the traffic around the corner should be predicted as one of the six categories. Specifically, our method focuses on the sound event detection for the following six classes of traffic conditions at two types of T-junctions (see Figure 2):

- *front:* a part or whole of the vehicle passing through the intersection appears in the view,
- *left approach:* an occluded vehicle approaches from the left corner and does not appear in the view,
- *left leave:* same, but the vehicle approaches from the right corner,
- *right approach:* an occluded vehicle leaves from the left corner and does not appear in the view,
- *right leave:* same, but the vehicle leaves from the right corner,

Fig. 3. Microphone array integrated device on our research vehicle.

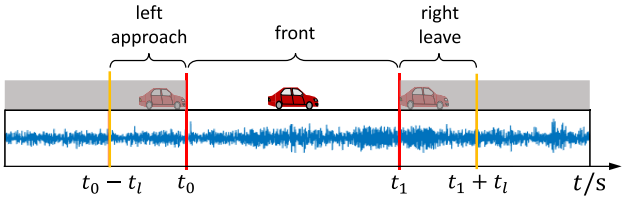


Fig. 4. Segmentation of a complete audio recording of a vehicle (the red vehicle in Figure 2a) passing through the intersection from the left side. t_0: the moment when the vehicle starts to appear, t_1: the moment when the vehicle leaves the line of sight, and t_l: a duration to avoid too weak sound signals from the NLOS vehicle in the audio segments of $left\ approach$ and $right\ leave$ classes.

- *none:* no car is approaching or leaving.

We split the process into three main stages: *1) Data collection and preprocessing*, *2) Feature representation*, and 3) *Neural network for feature extraction and classification*.

Since [2] is the first and only study on data-driven based acoustic NLOS vehicle approach detection to the best of our knowledge, we referred to the setup in [2] for the first and second stages to compare with it. It is worth noting that our method differs from [2] mainly in the core part, i.e., *2) Feature representation* and *3) Neural network for feature extraction and classification* stages.

A. Data Collection and Preprocessing

The first stage involves collecting multi-channel audio and processing the raw audio into short audio samples with labels.

1) Research Vehicle for Data Collection: To collect audio-visual data of the vehicle passing through intersections, a microphone array was installed on the roof of the research vehicle [18]. The microphone array measures 0.3 $m \times 0.3\ m$, and consists of 32 spirally arranged micro electro mechanical system (MEMS) microphones with a frequency response range of 100 to 80,000 Hz. It simultaneously collected audio data from 32 channels at a sampling rate of 48 kHz. A forward-facing camera located at the array's center is equipped to record the moments when the detected vehicle enters and exits the field of view, providing a basis for the division of various types of sound signals. The center of the roof-mounted microphone array was approximately 1.8 m above the ground, as shown in Figure 3. Notably, the microphone we use has a frequency response range of [100, 80,000] Hz. Therefore, the frequency for DOA estimation cannot be lower than 100 Hz, while it is 50 Hz in [2].

2) Non-Line-of-Sight Approaching and Leaving Vehicle Detection Dataset: To validate the performance of our method, we created a real-world NLOS vehicle approaching and leaving detection dataset.

We recorded multiple audio clips at two intersections in Xi'an, which belong to different T-junction types. For safety reasons, the research vehicle was stopped during the audio collection, and only the detected vehicle was driving at the intersection without other motor vehicles. In practice, the research vehicle was parked 10∼12 m away from the intersection, and the detected vehicle was driven at about 18 km/h. The roof-mounted microphone array recorded the complete process of the detected vehicle from approaching to leaving intersections, while the *none* class audio was the surrounding sound when no vehicle passed the intersection. At each intersection, we recorded 20 audio clips of the detected vehicle passes and 5 audio clips of *none* class with a duration of 15 s. Furthermore, environmental noise at various volume levels was unavoidable, such as external air-conditioner working, pedestrian passing, etc.

Following [2], we categorize T-junctions with the rules in [2] and collect audio at both types of T-junctions. However, the difference is that we record the audio of the detected vehicle from approaching to leaving intersections, while the dataset created in [2] does not include the audio of the vehicle leaving intersections.

3) Sample Extraction: Since the complete audio recording of the vehicle passing through an intersection contains three classes of audio: *left approach/right approach* (vehicle approaching), *front* (vehicle in view), and *right leave/left leave* (vehicle leaving), we first segmented the audio recordings and labeled the classes to which the audio segments belong.

Figure 4 shows how to segment the complete audio recording of the vehicle passing through an intersection from the $left$. For a complete audio recording, define the time t_0 as the moment when the detected vehicle entered into view, and the time t_1 represents the moment when it left view on the opposite side. The determination of t_0 and t_1 was based on the video with a limited frame rate (30 Hz) corresponding to the audio recording. From the recordings of the detected vehicle passing the intersection from the $left$, extracting short audio at $t_0 - t_l < t \leq t_0$ as $left\ approach$ audio segments, short audio at $t_0 < t < t_1$ as $front$ audio segments, and $left\ leave$ audio segments are extracted at $t_1 \leq t < t_1 + t_l$. The recordings of the vehicle passing through the intersection from the $right$ side were segmented and labeled in the same manner. Besides, *none* audio recordings were considered not to contain other types of sound.

After segmentation and class labeling of the audio recordings, the audio samples were extracted with a fixed duration δ_t. Specifically, the audio segments from each class were then split through the sliding time window of length δ_t and hop length t_h, where the audio from each time window split was used as a new audio sample for feature representation [19]. Since a sufficient amount of data is beneficial for feature learning, $t_h = 10$ ms was set. For example, a 2 s audio segment generates 101 audio samples when δ_t is set to 1 s. Table I counts the audio samples of various classes extracted from the audio segments.

However, Table I also shows a class imbalance. In this study, we balanced the samples of each class by shortening the hop

TABLE I
SAMPLES PER CATEGORY. IN THE T-JUNCTION TYPE A/B (SEE FIGURE 2)

T-junction Type	front	left approach	left leave	right approach	right leave	none	Sum
A	6581	1010	1010	1010	1010	7005	17626
B	9163	1010	1010	1010	1010	7005	20208
A&B (without shortening t_h)	15744	2020	2020	2020	2020	14010	37834
A&B (shortening t_h)	15744	16020	16020	16020	16020	14010	93834

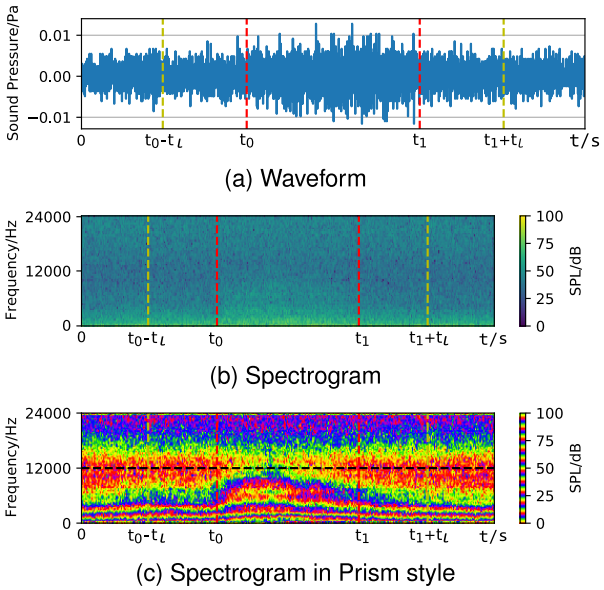


Fig. 5. Waveform and spectrograms of the audio clip containing three sound events: "Vehicle approaching", "Vehicle in view", and "Vehicle leaving", separated by red dashed lines.

length t_h of *approach* and *leave* classes. Specifically, the hop length of *approach* and *leave* classes was shortened to one-eight of that of the *front* and *none* classes, i.e., 1.25 ms. The number of samples in each class after shortening the hop length t_h is also shown in Table I, and it can be seen that the samples in each class are well balanced.

B. Feature Representation

The second stage uses audio samples to generate two feature representations, the spectrogram, and the DOA estimation converted image.

1) Spectrogram Generation: The recorded audio contains sound signals of the engine working and tire friction with the ground from the detected moving vehicle. Features for acoustic NLOS vehicle detection are included in the signals, and we represented these features by the spectrogram.

With the use of CNN, taking spectrogram as input provides better sound classification accuracy than taking Mel-Frequency Cepstral Coefficients [20], [21]. Furthermore, according to the Doppler effect, the frequency of the sound collected by the observer changes as the source approaches or moves away from the observer, which can be described by the spectrum distribution in the frequency domain, and the time-frequency feature has been proven effective in the LOS vehicle approaching warning [22]. Thus, transferring the single-channel audio samples to the frequency domain becomes necessary.

Road noise corresponds to traffic in real-time and is a non-smooth signal. The Short-Time Fourier Transform (STFT) can represent the frequency and phase components of a time-dependent signal [23]. STFT can be expressed mathematically as

$$STFT_x(t,\omega) = \int_{-\infty}^{+\infty} x(t)h(t-\omega)e^{-j\omega t}dt \tag{1}$$

where $x(t)$ is the time-domain signal to be transformed, $h(t-\omega)$ is the window function, and j is the complex unit.

To represent the power spectral density of the function visually, a modulo-square operation is performed on the results of $STFT_x(t,\omega)$ to obtain a spectrogram, which can be expressed as

$$SPEC_x(t,\omega) = |STFT_x(t,\omega)|^2 \tag{2}$$

Specifically, using the Hanning-window, STFT was created for each sample with the duration of δ_t, and the spectrogram as one of the feature representations was generated. Practically, with python, we generated spectrograms using *librosa* [24].

The spectrogram represents the sound signal with three variables: frequency, time, and intensity (color scale). Figure 5 shows a complete recording of the waveform and spectrograms of three sound events: the detected vehicle approaching the intersection ($t < t_0$), the detected vehicle traveling in view ($t_0 < t < t_1$), and the detected vehicle leaving the intersection from the opposite side ($t > t_1$). As shown in Figure 5b, the color of the spectrogram indicates the sound pressure level (SPL) of the sound signal.

The variation of the spectrogram can also provide a basis for our parameter settings. To better visualize the change in SPL of each frequency over time in the spectrogram, we show a Prism-style spectrogram in Figure 5c. It is worth noting that the spectrogram rather than the Prism style spectrogram is used as input to the model. As can be seen in Figure 5c, the SPL of some frequencies changes gradually as the detected vehicle approaches the intersection. When the detected vehicle comes into view, the SPL at certain frequencies, mainly below 12,000 Hz, significantly change and remain until the detected vehicle disappears. And after the detected vehicle disappears from view, the SPL changes of these frequencies gradually disappear. Therefore, we set up experiments with $f_{max} = \{6{,}000\ Hz,\ 9{,}000\ Hz,\ 12{,}000\ Hz,\ 15{,}000\ Hz\}$ to find the optimal frequency band for DOA estimation. And Figure 5b also shows that, the larger t_l is, the less variation there is in the spectrogram. Therefore, we set up comparison experiments with $t_l = \{2\ s,\ 2.5\ s,\ 3\ s\}$ to investigate the influence of t_l on classification performance and early detection time.

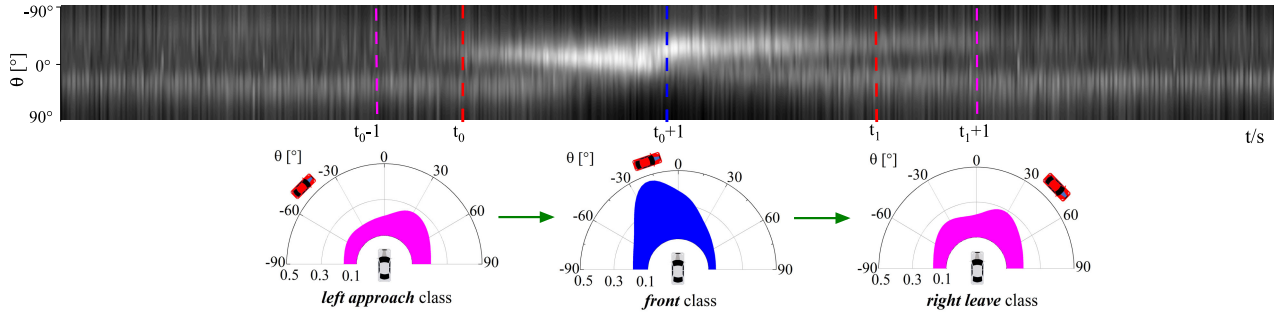


Fig. 6. Converted image of the continuous DOA estimation results of the detected vehicle passed through the intersection and the ambient sound power map of three of these moments. The red vehicle indicates the actual position of the detected vehicle. When the detected vehicle is fully in view ($t = t_0 + 1$: *front* class), the angle of the peak corresponds to the localization of the detected vehicle. When the detected vehicle is occluded ($t = t_0 - 1$: *left approach* class and $t = t_1 + 1$: *right leave* class), the NLOS vehicle cannot be localized by peak search.

2) DOA Estimation Converted Image Generation: Acoustic vehicle localization can be described as a problem of DOA estimation [25].

Steered Response Power (SRP) based method was chosen for DOA estimation. The DOA estimation methods [26] can be roughly classified into three groups: 1) Time Difference of Arrival (TDOA)-based algorithms [27], [28], [29]; 2) algorithms based on maximizing the SRP of the beamformer, such as SRP-PHAT [30]; 3) subspace algorithms, such as Multiple Signal Classification (MUSIC) [31], [32]. Compared to the other two groups of methods, the TDOA-based algorithms are sensitive to noise and reverberation [33]. Nevertheless, MUSIC needs to pre-estimate the number of active sound sources before DOA estimation, and incorrect estimation of source number leads to biased or even invalid DOA estimation result [34]. However, in real traffic scenarios, active sound sources changes almost constantly. Therefore, for non-smooth signals such as ambient sound, MUSIC is not applicable, we chose SRP-based method for DOA estimation.

Introducing PHAT $G_n(\omega) = 1/|X_n(\omega)|$ to compute the SRP is commonly known as the SRP-PHAT algorithm [30], the computation of the SRP is illustated in the Appendix. First, PHAT is the inverse of the amplitude of the GCC, allowing the mutual power spectrum of the signals to retain only the phase information associated with the time delay, which improves the noise and reverberation immunity of the SRP algorithm. However, PHAT also drops the amplitude information, which can represent the ambient sound intensity. Thus, one of the reasons we introduce the spectrogram as an auxiliary input is to supplement the amplitude information lost by SRP-PHAT. Second, the SRP-PHAT allows us to obtain an ambient sound power map with a maximum sound power angle corresponding to the source location [33]. As shown in Figure 6, based on the direct sound assumption, the SRP-PHAT computes the ambient sound power distribution. However, the sound of the NLOS vehicle is not direct to the microphone array, which means that the direction of the NLOS vehicle cannot be localized by peak search. Therefore, classifiers are required to model the mapping relationship between the SRP-PHAT result and the actual direction of the NLOS vehicle.

SRP-PHAT was used to calculate DOA estimates. Firstly, STFTs were created of 32 channels audio signals. Next, using the GCC of 32 STFTs, SRP-PHAT calculated the DOA energy $P(\theta)$ for any given azimuth angle θ ahead of the intelligent vehicle with a bandpass frequency range of $[f_{min}, f_{max}]$, the computation of the $P(\theta)$ is detailed in the Appendix. The band selection of filtering was analyzed in 1) *Spectrogram generation* of Section II-A. Specifically, to represent the temporal evolution in DOA estimation results, the STFT result of each channel was divided into 61 pieces along the temporal dimension. To form a square representation of the DOA estimation converted image, the azimuth range $[-90°, 90°]$ in front of the intelligent vehicle was equally divided into 61 nodes $\theta_1, \ldots, \theta_{61}$. DOA energy at the azimuth of 61 nodes was calculated (1×61) and 61 response vectors were calculated from an multichannel audio sample. These response vectors were then connected chronologically into a (61×61) dimensional feature S.

We adopted the min-max normalization for S, S_{max} and S_{min} were the maximum and minimum element values of S, then the pixel value of the corresponding image with the size of (61×61) was:

$$Pix(i, j) = 255 \times (S(i, j) - S_{min})/(S_{max} - S_{min}) \quad (3)$$

Finally, the feature matrix S was converted into an spectrogram-like image, thus transforming the task into an image classification problem.

C. Neural Network of Feature Extraction and Classification

In the third stage, we propose a novel neural network-based NLOS vehicle approaching and leaving detection model, pCRNN.

The approach and leave of the NLOS vehicle possess temporal characteristics that can benefit classification. Certain sound events can be easily distinguished due to their impulsive characteristics [35], (e.g., glass smash), while the approach and leave of the NLOS vehicle lasts for a while. Therefore, classification methods that can retain temporal context along the feature representations are well suited for detecting the approaching or leaving NLOS vehicle. To realize the extraction and classification of the input feature representations' spatial and temporal correlation features, the main building blocks of our system are two CRNNs that share the same parallel convolution and LSTM architecture.

The network consists of two branches. One branch takes DOA estimation converted image as input. In parallel, the

spectrogram is fed into another branch. The feature is extracted for each image pair input through the convolution and LSTM blocks. The two-branch structure of the pCRNN combines the acoustic NLOS vehicle detection task and localization task using the spectrogram and DOA estimation results converted image as input.

1) Convolutional Block: The input of the neural network is composed of three sequential CNN blocks. In particular, each sequential block consists of a convolutional layer with a kernel size of 5, a ReLU activation function, a batch normalization, and an average pooling of size 2×2 is followed to downsample the feature map. With the input image size of 61×61, the output of the convolutional layer is of dimension $32 \times 4 \times 4$. Finally, flatten the feature map into a vector of 512×1.

2) LSTM Block: The output of convolutional layers contains rich feature representations but omits hidden significant features with long-term dependency. However, the self-recurrent weights in LSTM hidden layers do not only depend on the current input, and the cell in the memory block retains previous information. Thus, the two-stacked LSTM layers are employed in the proposed neural network, and every layer contains 122 LSTM units. The output size of the LSTM block is 122×1. Next, merging the features extracted from the convolution and LSTM block into a vector of 634×1, we concatenate the merged feature vector as the input to fully connected (FC) layers, which enable the proposed model to learn complementary temporal context feature and high-level feature simultaneously [36]. These features are extracted and fused to form a feature vector with dimension 1268×1.

3) FC Layers: There are two FC layers, the first FC layer consists of 2,048 output nodes with linear activation, followed by a Dropout layer, a ReLU layer, and finally, a linear layer with 2,048 input nodes generates a vector of size 6 representing the probability distribution of the intersection traffic categories.

In contrast to [2], which uses SVM as the classifier, we propose a neural network-based model, pCRNN. Moreover, unlike existing studies [2], [13], [14], [15] where only the SSL information is used as input, by allowing the spectrogram and the DOA estimation result converted image to be used together as input, the two-branch structure of the pCRNN achieves the combination of acoustic NLOS vehicle detection and localization tasks.

III. Experiments

In this section, we validated the classification performance of pCRNN on two real-world datasets. First, our dataset and then the Occluded Vehicle Acoustic Detection dataset (OVAD dataset) [2], since the OVAD dataset is the first and only publicly available dataset related to acoustic NLOS vehicle approaching detection.

A. Evaluation Metrics

We used two metrics to evaluate the classification performance of the pCRNN model:

(1) Overall Accuracy ($Acc/\%$):

$$Acc = \sum_{c \in C} \frac{TP_c}{TP_c + TN_c + FP_c + FN_c} \times 100\% \quad (4)$$

(2) Jaccard Index ($J/\%$):

$$J_c = \frac{TP_c}{TP_c + FP_c + FN_c} \times 100\% \quad (5)$$

where c denotes a class belonging to $C = \{front, left\_approach, left\_leave, right\_approach, right\_leave, none\}$. Regarding the classification result as positive for target class c and negative for the other five categories, TP_c/TN_c are the True Positives/Negatives, FP_c/FN_c are the False Positives/Negatives.

Specifically, Acc was used to measure the overall classification accuracy. J_c was computed involving samples in each class c that were correctly classified, samples misclassified into other classes, and samples in other classes that were misclassified into that class. For both metrics, the larger the value, the better the classification performance is.

B. Data Augmentation

The data augmentation method was used to alleviate the problem of overfitting and the large amount of training data required for model training. We converted continuous DOA estimation results into a spectrogram-like image form, allowing us to utilize acoustic image data augmentation. Therefore, we used two data augmentation operations, Frequency Masking and Time Masking from SpecAugment [37], as follows.

- Frequency Masking: masks f consecutive frequency channels $[f_0, f_0 + f)$, where f_0 is randomly generated from $[0, F - f)$, where $F = 24{,}000$ Hz. Replace frequency parameters with angles in the DOA estimation conversion result image.
- Time masking: masks t consecutive time steps $[t_0, t_0 + t)$, where t_0 is randomly selected from $[0, \tau - t)$, $\tau = \delta_t$.

We validated the effectiveness of each data augmentation operation by comparing the classification performance with that trained on the raw data. The experimental results and conclusions can be found in Sec. III-C. 6) *Influence of data augmentation operation.*

C. Influence of Parameters and Features

Based on the description in Sec II, we performed experiments with different settings and compared classification performance with the baseline on our dataset.

1) Choosing Parameters: The parameters of our method mainly include the frequency band range $[f_{min}, f_{max}]$, the audio sample duration (δ_t), and the duration of the audio segment (t_l).

To study the influence of different parameter settings on the classification performance, we adjusted these parameters and conducted experiments. Firstly, as mentioned in Sec.II-A-1) and Sec.II-B-1), we fixed f_{min} at 100 Hz and adjusted $f_{max} = \{6{,}000$ Hz, $9{,}000$ Hz, $12{,}000$ Hz, $15{,}000$ Hz$\}$ to select the optimum frequency band. Then, fixing the optimal frequency band, we performed experiments on different data

TABLE II

BASELINE COMPARISON AND PARAMETER STUDY ON OUR DATASET. OUR REFERENCE CONFIGURATION: FREQUENCY BAND [100, 12,000] HZ, t_l = 2 S, δ_t = 1 S, WITH FREQUENCY MASKING AND TIME MASKING FOR DATA AUGMENTATION. * INDICATES OUR PIPELINE

Number	Method	Acc	J_{front}	$J_{left\_approach}$	J_{left_leave}	$J_{right\_approach}$	J_{right_leave}	J_{none}
1	*pCRNN	**96.71**	96.61	**92.94**	87.26	95.34	**95.30**	95.16
2	pCRNN [100, 6,000] Hz	94.74	98.13	85.38	77.70	94.22	93.49	94.18
3	pCRNN [100, 9,000] Hz	95.78	97.62	89.67	81.99	95.32	94.26	94.51
4	pCRNN [100, 12,000] Hz	96.27	98.46	91.89	85.61	94.48	90.37	**97.73**
5	pCRNN [100, 15,000] Hz	95.69	96.13	90.91	84.85	**96.39**	92.94	90.04
6	pCRNN (with Frequency Masking only)	96.30	**99.47**	89.54	84.59	95.50	94.09	95.29
7	pCRNN (with Time Masking only)	96.44	98.11	90.72	**87.53**	93.65	92.72	97.65
8	*pCRNN ($\delta_t = 0.5$ s)	90.26	95.68	75.72	67.65	81.05	82.20	93.06
9	*DOA-only (CRNN)	95.57	98.07	91.75	84.37	92.39	92.09	91.71
10	DOA+SVM [2]	91.56	97.56	81.18	75.82	89.26	84.69	82.33

augmentation operations, different audio sample durations δ_t = {0.5 s, 1 s} and different audio segment lengths t_l = {2 s, 2.5 s, 3 s}.

2) Competing Method: Because of the lack of acoustic NLOS vehicle approaching and leaving detection methods in existing studies, we only compared relevant acoustic NLOS vehicle approaching detection method [2].

In our baseline, each audio sample was first segmented into two temporally non-overlapping segments, and the DOA estimate for each segment was computed with an angular resolution of 6°. Then, the DoA estimates, represented as a 2×30 matrix, were taken as input and trained by the SVM classifier. For a fair comparison on our dataset, the output of the SVM [2] was adjusted from 4 to 6 classes.

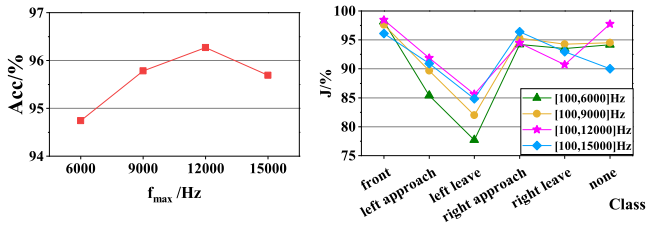
3) Implementation & Model Training: Processing is done in python, using pyroomacoustics [38] for acoustic feature extraction, and PyTorch [39] for classifier training.

As a multi-classification problem, the model optimizes cross-entropy loss by Adam. For training, we manually adjusted and set the learning rate to a constant 0.0001, the training epoch to 50, and the batch size to 32. To prevent overfitting, we applied a dropout with the probability of 0.5 and regularized the network with a weight decay of λ = 0.01.

To minimize the effect of different dataset divisions on accuracy, we employed 5-fold cross-validation for model performance evaluation. Notably, unlike randomly selecting samples from all samples as training set, we equally split 25 audio clips recorded at each intersection into five groups, which ensures that the samples in the training set and test set do not come from the same audio recording, avoiding data leakage. Thus, all the audio recordings were divided into five parts, and for each fold of the training process, always 4 parts of the dataset were taken as input for training, and the remaining part for testing.

4) Experimental Results: Table II shows the overall accuracy and Jaccard Index for each category with different parameters, data augmentation operations, feature representation inputs, and the baseline.

Our final choice and reference is frequency band [100, 12,000] Hz, Frequency Masking and Time Masking data augmentation operations, δ_t = 1 s, t_l = 2 s, using spectrogram and DOA estimation result converted image pair as input. It can be seen that our reference in Experiment 1 achieves the highest overall accuracy, and the Jaccard Index of $left\ leave$ with the lowest classification accuracy also reaches 87.26%, which indicates that the proposed passive acoustic perception method achieves early detection of the approaching and leaving NLOS vehicle. In the following, we specifically analyze the influence of different settings on classification performance.

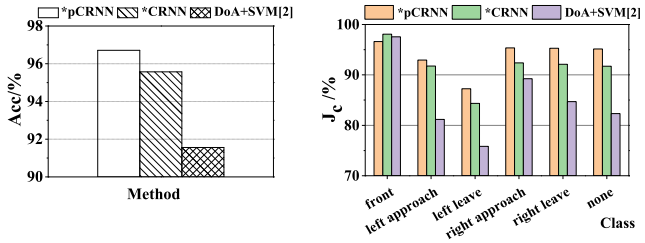


(a) Acc (b) Jaccard Index

Fig. 7. Comparison of results on different frequency bands on our dataset. (a) Overall accuracy at different frequency bands; (b) Jaccard Index of each class at different frequency bands.

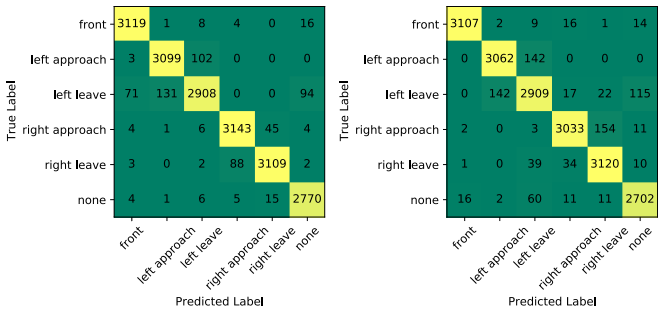
5) Influence of Frequency Band: Vehicles emit sounds in a specific frequency range; thus, choosing a suitable frequency band range enables the DOA estimation results to contain sufficient vehicle localization information and avoids non-vehicle frequency band sound interference. To select the appropriate frequency band, with the fixed f_{min} = 100 Hz, we conducted Experiments 2, 3, 4, and 5 in Table II with the upper band limit f_{max} = {6,000 Hz, 9,000 Hz, 12,000 Hz, 15,000 Hz}. Figure 7 shows a comparison of the results on different frequency bands.

In Figure 7a, the overall accuracy of our pCRNN first increases and then decreases as the frequency band expands. Specifically, we can see that the overall accuracy has been growing when the frequency band range is extended from [100, 6,000] Hz to [100, 12,000] Hz. However, the overall accuracy decreases when f_{max} exceeds 12,000 Hz, i.e., the frequency band continues to expand to [100, 15,000] Hz. The trend of the overall accuracy with f_{max} in Figure 7a is in good agreement with the spectral variation in Figure 5, where the spectral variation during the vehicle passing is concentrated below 12,000 Hz. Therefore, we took [100, 12,000] Hz as the frequency band for DOA estimation.



(a) Acc (b) Jaccard Index

Fig. 8. Comparison of overall accuracy and Jaccard Index of pCRNN, CRNN and the baseline on our dataset.



(a) pCRNN (b) CRNN

Fig. 9. Comparison of the confusion matrices of pCRNN, CRNN on our dataset.

Figure 7b compares the Jaccard Index for each class at different frequency bands. It can be seen that the narrower frequency band, i.e., [100, 6,000] Hz is unfavorable for the classification of *approach* and *leave* classes. As the frequency band increases, the classification accuracy of the *approach* and *leave* classes improves. In addition, we observe that the Jaccard Index of *left leave* is lower than that of *left_approach* in all frequency bands, as well as on the *right* side. This suggests that the leaving NLOS vehicle is always more challenging to detect than the approaching one on the same side, which may be related to the influence of the vehicle's sounding position and body structure on sound propagation.

6) Influence of Data Augmentation Operation: To verify the effectiveness of Frequency Masking and Time Masking on the acoustic NLOS vehicle detection task. We conducted Experiments 4, 6, and 7 in Table II, where Experiment 4 used the un-augmented raw feature representation for training, and Experiments 6 and 7 used Frequency Masking and Time Masking for data augmentation, respectively.

Table II shows that both data augmentation operations improve the classification performance, and the Time Masking is more obvious in comparison. Therefore, we combined these two data augmentation operations in Experiment 1 of Table II. Compared with the single operation, the combination of Frequency Masking and Time Masking further improves the classification performance with the highest overall accuracy of 96.71%.

7) Influence of Sample Duration: If the sample duration is shorter, it cuts off potential features and creates too much variability [40]. Therefore, choosing an appropriate sample duration δ_t is essential. We set δ_t to 0.5 s and 1 s for comparison in Experiment 8 and Experiment 1 in Table II, respectively.

Comparing Experiment 8 and Experiment 1 in Table II, it can be seen that the classification performance of pCRNN improves significantly as δ_t increases from 0.5 s to 1 s. Compared to $\delta_t = 0.5$ s, the overall accuracy of pCRNN at $\delta_t = 1$ s is improved by 6.45%, indicating that a sufficiently long audio sample duration is crucial for NLOS vehicle approaching and leaving detection. Therefore, we set $\delta_t = 1$ s.

8) Influence of the Introduction of Spectrogram: To verify the effectiveness of introducing the spectrogram as an input, in Experiment 1 of Table II, we used the DOA estimation result converted image together with the spectrogram as inputs to pCRNN. And in Experiment 9 of Table II, we used only the DOA estimation result converted image as the input without the spectrogram, correspondingly, we removed a branch of pCRNN, denoted as CRNN. A comparison of the experimental results of pCRNN and CRNN is also shown in Figure 8.

As shown in Figure 8, on our dataset, pCRNN outperforms CRNN in overall accuracy and Jaccard Index for all classes except the *front* class. These results indicate that the time-frequency feature in the spectrogram is effective for the NLOS vehicle approaching and leaving detection.

To further analyze the misclassification between the categories, we computed the confusion matrix, and Figure 9 shows the confusion matrices of pCRNN and CRNN. As shown in Figure 9, the most misclassified categories of pCRNN and CRNN are the same-side *approach* and *leave* categories, i.e., the *left approach* and *left leave* categories as well as the *right* side, which suggests that the acoustic patterns of the same-side NLOS vehicle approaching and leaving are similar. However, compared to CRNN, pCRNN reduces the number of misclassifications between *approach* and *leave* categories, which indicates that the spectrogram is beneficial for differentiating the approaching and leaving states of the NLOS vehicle. In addition, compared to CRNN, pCRNN has 34 fewer misclassifications of other categories to *none* category and 69 fewer misclassifications of *none* category to other categories, indicating that introducing the spectrogram can reduce both the miss rate and false alarm rate.

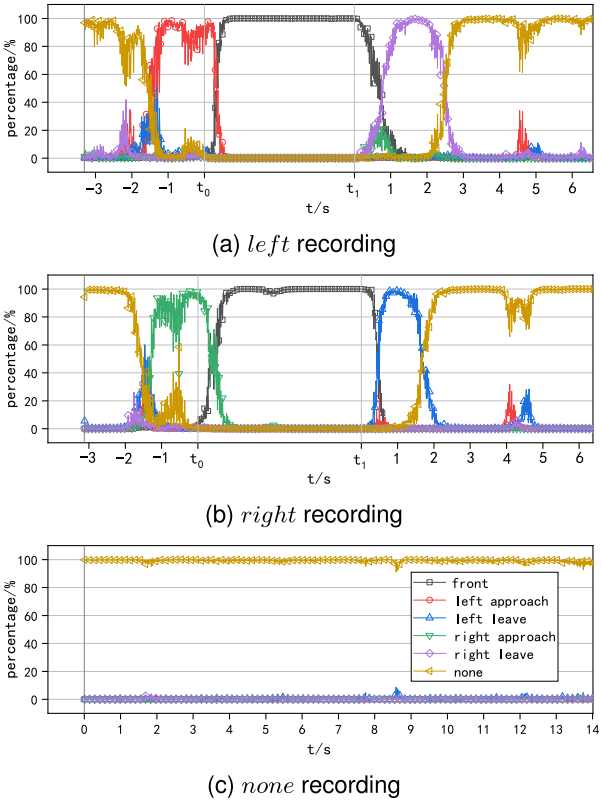
9) Compare With the Baseline: To compare our proposed pCRNN model with the baseline, we performed experiments with both methods on the dataset we created. We report the experimental results of pCRNN and the baseline in Experiment 1 and Experiment 10 of Table II, and Figure 8 compares the experimental results for these two methods.

Compared with the state-of-the-art method [2], our pCRNN achieved advanced classification performance on this dataset. Figure 8a shows that our pCRNN outperforms the baseline by 5.15% overall accuracy. Figure 8b indicates that pCRNN significantly improves the Jaccard Index on the NLOS categories, i.e., the other five categories except for the *front* category. These experimental results demonstrate the effectiveness of our pCRNN model, thus suggesting our method's advantages on our newly created dataset.

In addition, as can also be seen in Figure 8, using only DOA estimation information as input, the CRNN has a higher overall

TABLE III
CLASSIFICATION RESULTS FOR DIFFERENT t_l ON OUR DATASET

Number	Method	t_l	Acc	J_{front}	$J_{left\_approach}$	J_{left_leave}	$J_{right\_approach}$	J_{right_leave}	J_{none}
1	pCRNN	2 s	**96.71**	96.61	**92.94**	**87.26**	**95.34**	**95.30**	**95.16**
2	pCRNN	2.5 s	93.92	96.19	84.39	76.13	93.51	89.10	94.26
3	pCRNN	3 s	89.35	93.17	75.93	68.63	84.86	75.86	89.87
4	DOA+SVM [2]	2 s	91.56	**97.56**	81.18	75.82	89.26	84.69	82.33



(a) $left$ recording

(b) $right$ recording

(c) $none$ recording

Fig. 10. The classification result of the entire audio clips using pCRNN. The detected vehicle *approached* and passed the intersection from the $left$ and $right$ blind corners.

accuracy and higher Jaccard Index in each category than the baseline, which demonstrates that, compared to the baseline, our proposed representation of continuous DOA estimation results in the form of images and neural network-based image classification model can better utilize the DOA estimation information.

D. Advanced Detection of NLOS Vehicle Approaching and Leaving

1) Detection Results of pCRNN on Whole Audio Clips: To better understand the pCRNN model, we tested the entire audio clips of the vehicle from $approach$ to $leave$. First, a sliding time window with $\delta_t = 1$ s and a hop length of 10 ms was used to slide the entire audio. Then, the image pairs generated from the audio in the time window were input to pCRNN in chronological order. Finally, the traffic category probability corresponding to each pair of images were output, and the category with the highest probability was considered as the traffic situation predicted by the model.

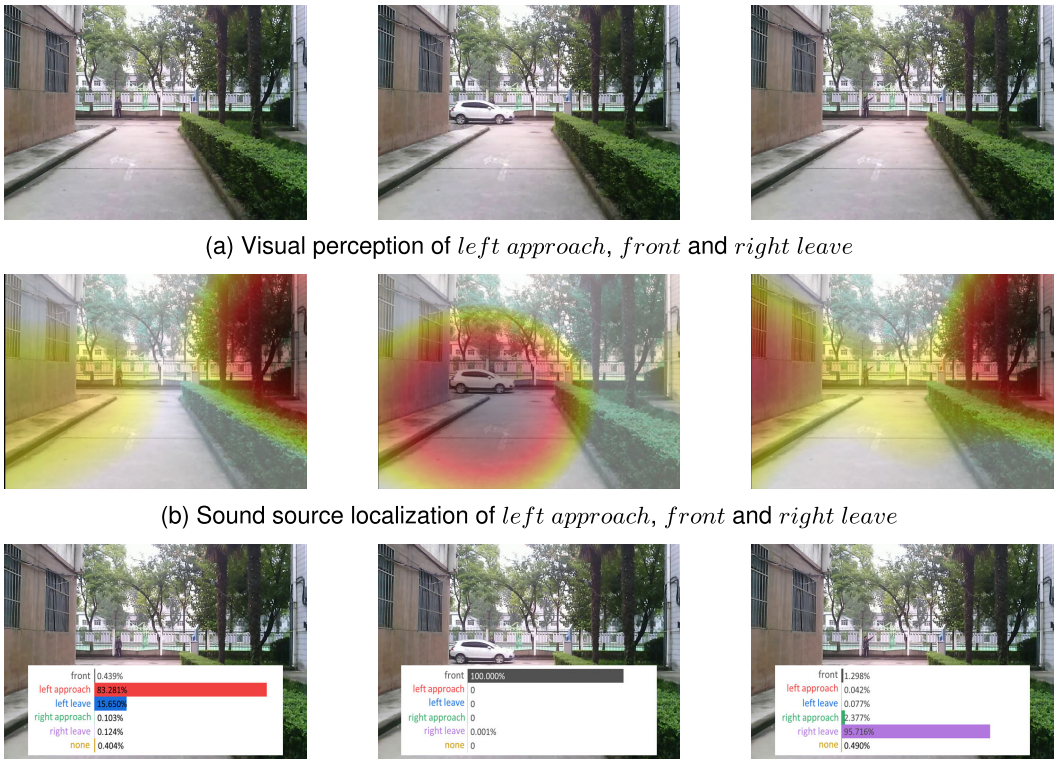
Figure 10 shows the functional relationship between the probability of each category and extraction time t of test audio clips, which was distinguished by the direction of the detected vehicle entering the intersection. As shown in the probability curve of $left$ recording, the probability of $none$ is highest at the beginning, which means that the model predicts that no vehicle is approaching or leaving the intersection. As time goes on, the probability of $none$ class drops sharply, and the classification result starts to shift to $left\ approach$, remained until t_0. For a short time after t_0, the detected vehicle gradually appears in LOS, the probability of $left\ approach$ decreases rapidly, while the probability of $front$ increases to the maximum probability. At t_1, the detected vehicle completely leaves the visible area of the intersection from the $right$ side. At this point, the maximum probability starts to switch from $front$ to $right\ leave$ rapidly. During the occluded detected vehicle leaves the intersection from the $right$ side, the prediction result maintains the classification result of $right\ leave$ for a period of time, and finally converts to $none$ when the detected vehicle leaves far away. The mirrored prediction result was obtained from the $left$ recording. And the maximum probability category of $none$ recording is kept at $none$.

In general, the category output of the model in the entire audio clip corresponds to the process of the detected vehicle passing through the intersection. The passive acoustic perception system realizes the detection of the NLOS vehicle more than 1 s in advance through audio cues. It also distinguishes the approaching and leaving status of the detected vehicle.

2) Influence of t_l on Advance Detection Time: To explore earlier detection of the NLOS vehicle, we expanded the t_l to 2.5 s and 3 s for training and testing.

As shown in Figure 4, the larger t_l is, the further away the NLOS vehicle is from the intersection, and the weaker the sound signal received by the microphone array is. Thus, we used t_l to ensure the intensity of the NLOS vehicle sound signal in the training samples. According to our settings of $t_l = 2$ s and $\delta_t = 1$ s, the theoretical advance detection time is $t_{adv} = t_l - \delta_t = 1$ s. However, Figures 10a and 10b show that pCRNN detected the NLOS vehicle more than 1 s in advance, which indicates that pCRNN can also detect weaker NLOS vehicle sound signals. Therefore, we used larger $t_l = \{2.5\ s,\ 3\ s\}$ to explore earlier detection of the NLOS vehicle by pCRNN. We report the classification results on our dataset for $t_l = \{2\ s,\ 2.5\ s,\ 3\ s\}$ in Table III. In addition, we also report the classification results of the method in [2] on our dataset for $t_l = 2$ s in Table III.

As can be seen from Table III, the classification accuracy of the pCRNN is reduced as t_l increases, to ensure the classification accuracy, $t_l = 2$ s is chosen in our method. However,



(a) Visual perception of $left\ approach$, $front$ and $right\ leave$

(b) Sound source localization of $left\ approach$, $front$ and $right\ leave$

(c) Our passive acoustic perception result of $left\ approach$, $front$ and $right\ leave$

Fig. 11. Perceived results of three modes for type A intersection traffic at three identical moments. The left-to-right moments of the three images in the same line are $left\ approach : t = t_0 - 1s$, $front : t = t_0 + 1s$, $right\ leave : t = t_1 + 1s$.

as shown in Experiment 2 of Table III, the overall accuracy of the pCRNN still reaches 93.92% when $t_l = 2.5$ s, i.e., the pCRNN detects the NLOS vehicle 1.5 s in advance with an accuracy of 93.92%. Therefore, by appropriately increasing t_l for training, the pCRNN can detect the NLOS vehicle more than 1 s in advance. It is worth noting that, as shown in Experiment 4 in Table III, the accuracy of the method in [2] detects the NLOS vehicle 1 s in advance with an accuracy of 91.56%, which is lower than the accuracy of our method in detecting the NLOS vehicle 1.5 s in advance. It indicates that our method detects the NLOS vehicle earlier than the method in [2] on our dataset.

3) Visual Representation of Conventional Sound Source Localization Results and pCRNN Detection Results: Our passive acoustic perception method aims to detect the occluded approaching and leaving vehicle in advance, and therefore we compared the traffic perception result of the vision-based, conventional beamforming (CBF) algorithm and our method at three identical moments, which are $t = t_0 - 1s$, $t = t_0 + 1s$ and $t = t_1 + 1s$. Figure 11 shows the comparison of the detected vehicle entering the intersection from the left.

First, since the detected vehicle at the moments of $t = t_0 - 1s$ and $t = t_1 + 1s$ is completely occluded by the intersection building, the camera couldn't capture the moving vehicle under the corner. Next, the 2D DOA estimation superimposed on the camera images shows the SSL algorithm calculating multiple sound sources generated by acoustic reflection and diffraction, etc., under both LOS and NLOS environments. As shown in Figure 11b, the moving detected vehicle is recorded as salient sound source when the vehicle is fully in view. However, the most salient SSL result is not indicative of correct localisation when the detected vehicle is completely occluded. Notably, as shown in Figure 11c, our method predicts the direction and relative motion state of the detected vehicle at these two moments ($t = t_0 - 1s$ and $t = t_1 + 1s$), which indicates that our passive acoustic perception method provides additional reaction time for intelligent vehicles.

E. Experimental Validation on the Publicly Available Dataset

We evaluated the classification performance of the pCRNN model on another real-world dataset: the OVAD dataset. Since the OVAD dataset is the first and only publicly available dataset related to acoustic NLOS vehicle approach detection.

1) Data Description: Schulz Y et al. [2] created the OVAD dataset collected in urban outdoor environments. The OVAD dataset was recorded at five T-junctions containing types A and B around Delft, The Netherlands. There are 623 static recordings and 441 dynamic recordings. The difference between static and dynamic data is whether or not the research vehicle is moving. In the more challenging dynamic data, the research vehicle reached the intersection at ∼15 km/h. Unlike our collection of complete audio of the detected vehicle passing through the intersection, the OVAD dataset does not contain audio of the vehicle leaving the intersection, which means that the number of classes changes from 6 to 4, i.e., $C = \{front, left\_approach, right\_approach, none\}$.

2) Implementation & Model Training: The classification performance was validated using 5-fold cross-validation on

TABLE IV
CLASSIFICATION RESULTS ON THE STATIC DATA OF OVAD DATASET

Number	Method	Acc	$J_{left\_approach}$	J_{front}	$J_{right\_approach}$	J_{none}
1	pCRNN	**94.45**	**82.55**	**94.33**	**87.50**	**89.33**
2	DOA-only (CRNN)	92.58	81.86	91.93	84.29	83.90
3	DOA+SVM [2]	89.63	71.25	90.82	78.08	79.66

TABLE V
CLASSIFICATION RESULTS ON THE DYNAMIC DATA OF OVAD DATASET

Number	Method	Acc	$J_{left\_approach}$	J_{front}	$J_{right\_approach}$	J_{none}
1	pCRNN	**81.69**	**46.86**	**92.94**	**57.81**	**66.43**
2	DOA-only (CRNN)	80.45	42.40	90.96	57.74	66.02
3	DOA+SVM [2]	77.72	38.83	88.52	46.00	64.83

static and dynamic data. Due to the limitation of the audio segment duration in the OVAD dataset, we shortened the audio duration δ_t from 1 s to 0.5 s with a hop length t_h of 10 ms, and adjusted the model output from 6 to 4 classes.

3) Competing Methods: To the best of our knowledge, only Schulz Y et al. [2] studied the data-driven based acoustic NLOS vehicle approach detection. Therefore, we used the state-of-the-art method in [2] as the baseline. Then, to verify the effectiveness of introducing the spectrogram as an input, we removed the spectrogram from the input and downgraded pCRNN to CRNN by removing one branch.

4) Comparisons on the OVAD Dataset: We compared pCRNN, and CRNN with the baseline on the static data and dynamic data of the OVAD dataset. The results are shown in Table IV and Table V.

As shown in Table IV and Table V, compared with the state-of-the-art method [2], our pCRNN achieved advanced classification performance on both static and dynamic data, with the highest overall accuracy and the Jaccard Index of every class. Notably, our pCRNN outperforms the baseline by 3.97% overall accuracy, which is a significant improvement on the more challenging dynamic data. These experimental results further demonstrate the effectiveness of our method, thus showing the advantages of the proposed method on the OVAD dataset.

Table IV and Table V also show that the classification performance of CRNN is lower than that of pCRNN and higher than that of the baseline. Compared with CRNN, our pCRNN achieved superior classification performance, which indicates that introducing the spectrogram as an input is effective for acoustic NLOS vehicle approaching and leaving detection. Taking only the DOA estimation results as input, CRNN outperformed the baseline on all metrics, which suggests that our proposed continuous DOA estimation results representation in the form of spectrogram-like image and CRNN can utilize the DOA estimation information more fully.

IV. CONCLUSION

A passive acoustic NLOS vehicle approaching and leaving detection method was proposed to assist existing vehicle sensors in perceiving NLOS traffic. Taking the spectrogram and DOA estimation result converted image as input, pCRNN combines the acoustic NLOS vehicle detection task and the localization task. In our experimental setup, the proposed scheme achieved an overall accuracy of 96.71% on the task of 6-class intersection traffic conditions classification. The vehicle in the blind corner can be detected and inferred whether it is approaching or leaving. Detecting the approaching NLOS vehicle under corners 1s in advance, our method provides additional reaction time for intelligent vehicles. Moreover, compared to the state-of-the-art acoustic NLOS approach vehicle detection baseline [2], pCRNN has advantages in both datasets.

Although these preliminary findings provide the possibility for applying image classification methods based on deep learning in passive acoustic NLOS vehicle detection, our pCRNN has no advantage in computational overhead compared with the method in [2] that uses SVM as the classifier. And the false detection caused by the noise emitted by other sound sources in the environment is non-negligible. Therefore, in future work, we will optimize the network structure to reduce the computational overhead. And we will introduce noise reduction and sound separation methods to focus on detecting target sound sources. In addition, casting the problem of NLOS vehicle detection to an image classification task, the data enhancement and neural network structure research will help to improve the classification performance.

APPENDIX

SRP CALCULATION OF THE BEAMFORMER

For multi-channel audio signals, the signal received by the n^{th} sensor in the microphone array can be modeled as

$$x_n(t) = a_s(t) * h_n(\theta_s, t) + v_n(t) \tag{6}$$

where $a_s(t)$ is the signal received from the direction θ_s, under the assumption of acoustic directivity, θ_s is the source angle or position, h_n is the impulse response from θ_s to the n^{th} sensor, and $v_n(t)$ is the uncorrelated noise of this sensor, which is typically assumed to be white, Gaussian noise.

In the frequency domain, the output of the filter-and-sum beamformer steered at the direction or position θ can be defined as

$$Y(\omega, \theta) = \sum_{n=0}^{N-1} G_n(\omega) X_n(\omega) e^{-j\omega\tau_n(\theta)} \tag{7}$$

where N is the number of sensors of the array, $G_n(\omega)$ and $X_n(\omega)$ is the filter of the n^{th} sensor signal and its Fourier Transform, and $\tau_n(\theta_s)$ is the delay of the signal propagation from θ_s to the n^{th} sensor. The power of this signal is:

$$P(\theta) = \int_{-\infty}^{+\infty} |Y(\omega, \theta)|^2 d\omega \tag{8}$$

Since the direct computation of (8) is costly, it can be calculated utilizing the Generalized Cross-Correlation (GCC) function, called SRP, which can be written as

$$P(\theta) = 2\pi \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} R_{nm}(\Delta\tau_{nm}(\theta)) \tag{9}$$

where $\Delta\tau_{nm}(\theta) = \tau_n(\theta)$ - $\tau_m(\theta)$ and $R_{nm}(\tau)$ is the GCC between the n^{th} and m^{th} sensors:

$$R_{nm}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \Psi_{nm}(\omega) X_n(\omega) X_m^*(\omega) e^{j\omega t} d\omega \tag{10}$$

where $\Psi_{nm}(\omega) = G_n(\omega) G_m^*(\omega)$ is a weighting function, and * denotes the complex conjugate.

Equation (10), combined with PHAT $G_n(\omega) = 1/|X_n(\omega)|$ is commonly known as the SRP-PHAT algorithm [30].

REFERENCES

[1] D. Solomitckii, M. Heino, S. Buddappagari, M. A. Hein, and M. Valkama, "Radar scheme with raised reflector for NLOS vehicle detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9037–9045, Jul. 2022.

[2] Y. Schulz, A. K. Mattar, T. M. Hehn, and J. F. P. Kooij, "Hearing what you cannot see: Acoustic vehicle detection around corners," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2587–2594, Apr. 2021.

[3] C. G. Keller, T. Dang, H. Fritz, A. Joos, C. Rabe, and D. M. Gavrila, "Active pedestrian safety by automatic braking and evasive steering," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1292–1304, Dec. 2011.

[4] N. Scheiner et al., "Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using Doppler radar," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2065–2074.

[5] D. Faccio, A. Velten, and G. Wetzstein, "Non-line-of-sight imaging," *Nature Rev. Phys.*, vol. 2, no. 6, pp. 318–327, 2020.

[6] F. Naser et al., "Infrastructure-free NLoS obstacle detection for autonomous cars," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Nov. 2019, pp. 250–257.

[7] A. B. Yedidia, M. Baradad, C. Thrampoulidis, W. T. Freeman, and G. W. Wornell, "Using unknown occluders to recover hidden scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12223–12231.

[8] P. Sharma et al., "What you can learn by staring at a blank wall," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2310–2319.

[9] Q. Chen, Y. Xie, S. Guo, J. Bai, and Q. Shu, "Sensing system of environmental perception technologies for driverless vehicle: A review of state of the art and challenges," *Sens. Actuators A, Phys.*, vol. 319, Mar. 2021, Art. no. 112566.

[10] C.-P. Lai and R. M. Narayanan, "Ultrawideband random noise radar design for through-wall surveillance," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 4, pp. 1716–1730, Oct. 2010.

[11] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath Jr., "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, Dec. 2016.

[12] I. An, Y. Kwon, and S.-E. Yoon, "Diffraction- and reflection-aware multiple sound source localization," *IEEE Trans. Robot.*, vol. 38, no. 3, pp. 1925–1944, Jun. 2022.

[13] K. Asahi, H. Banno, O. Yamamoto, A. Ogawa, and K. Yamada, "Development and evaluation of a scheme for detecting multiple approaching vehicles through acoustic sensing," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2011, pp. 119–123.

[14] Z. Chen, Y. He, Q. Wang, and Y. Luo, "A sound source localization method under NLOS environment for vehicles," in *Proc. IEEE 4th Int. Conf. Electron. Technol. (ICET)*, May 2021, pp. 790–795.

[15] V. Singh, K. E. Knisely, S. H. Yönak, K. Grosh, and D. R. Dowling, "Non-line-of-sight sound source localization using matched-field processing," *J. Acoust. Soc. Amer.*, vol. 131, no. 1, pp. 292–302, Jan. 2012, doi: 10.1121/1.3664089.

[16] F. Grondin, J. Glass, I. Sobieraj, and M. D. Plumbley, "Sound event localization and detection using CRNN on pairs of microphones," 2019, *arXiv:1910.10049*.

[17] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound event detection: A tutorial," *IEEE Signal Process. Mag.*, vol. 38, no. 5, pp. 67–83, Sep. 2021.

[18] L. Ferranti et al., "SafeVRU: A research platform for the interaction of self-driving vehicles with vulnerable road users," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 1660–1666.

[19] A. Valada, L. Spinello, and W. Burgard, "Deep feature learning for acoustics-based terrain classification," in *Robotics Research*, vol. 2. Cham, Switzerland: Springer, 2018, pp. 21–37.

[20] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," *Proc. Comput. Sci.*, vol. 112, pp. 2048–2056, Jan. 2017.

[21] N. Bulatovic and S. Djukanovic, "Mel-spectrogram features for acoustic vehicle detection and speed estimation," in *Proc. 26th Int. Conf. Inf. Technol. (IT)*, Feb. 2022, pp. 1–4.

[22] S. Kawanaka, Y. Kashimoto, A. Firouzian, Y. Arakawa, P. Pulli, and K. Yasumoto, "Approaching vehicle detection method with acoustic analysis using smartphone for elderly bicycle driver," in *Proc. 10th Int. Conf. Mobile Comput. Ubiquitous Netw. (ICMU)*, Oct. 2017, pp. 1–6.

[23] G. Manhertz and A. Bereczky, "STFT spectrogram based hybrid evaluation method for rotating machine transient vibration analysis," *Mech. Syst. Signal Process.*, vol. 154, Jun. 2021, Art. no. 107583.

[24] B. McFee et al., "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, vol. 8, 2015, pp. 18–25.

[25] Y. Jang, J. Kim, and J. Kim, "The development of the vehicle sound source localization system," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2015, pp. 1241–1244.

[26] V. Kumar and S. K. Dhull, "Techniques of direction of arrival estimation: A review," *IUP J. Elect. Electron. Eng.*, vol. 9, no. 1, p. 48, 2016.

[27] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[28] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Amer.*, vol. 107, no. 1, pp. 384–391, Jan. 2000.

[29] L. Comanducci, M. Cobos, F. Antonacci, and A. Sarti, "Time difference of arrival estimation from frequency-sliding generalized cross-correlations using convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4945–4949.

[30] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, "Robust localization in reverberant rooms," in *Microphone Arrays: Signal Processing Techniques and Applications*. Berlin, Germany: Springer, 2001, pp. 157–180.

[31] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. AP-34, no. 3, pp. 276–280, Mar. 1986.

[32] J. P. Dmochowski, J. Benesty, and S. Affes, "Broadband music: Opportunities and challenges for multiple source localization," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, Oct. 2007, pp. 18–21.

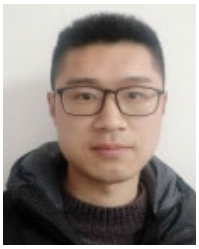
[33] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 300–311, 2021.

[34] S. Adavanne, A. Politis, J. Nikunen, and T. Virtanen, "Sound event localization and detection of overlapping sources using convolutional recurrent neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 1, pp. 34–48, Mar. 2019.

[35] E. Çakir, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.

[36] S. H. Bae, I. Choi, and N. S. Kim, "Acoustic scene classification using parallel combination of LSTM and CNN," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2016, pp. 11–15.

[37] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv:1904.08779*.
[38] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A Python package for audio room simulation and array processing algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 351–355.
[39] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.
[40] J. Libby and A. J. Stentz, "Using sound to classify vehicle-terrain interactions in outdoor environments," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2012, pp. 3559–3566.



Mingyang Hao received the M.Sc. degree from the Huaiyin Institute of Technology, Huai'an, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, China. His research interests include deep learning and intelligent acoustic perception.

Fangli Ning received the B.Sc. and M.Sc. degrees in mechanical design and theory and the Ph.D. degree in underwater acoustic engineering from Northwestern Polytechnical University, Xi'an, China, in 1996, 1999, and 2003, respectively. His research interests include microphone array signal processing, robots, and intelligent perception technology.

Ke Wang received the B.S. degree in energy and power engineering from Jilin University, Changchun, China, in 2019. He is currently pursuing the M.S. degree with the School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, China. His research interests include deep learning and intelligent acoustic diagnosis.

Shaodong Duan received the B.S. degree in mechanical design, manufacturing and its automation from Southwest Jiaotong University, Chengdu, China, in 2021. He is currently pursuing the M.S. degree with the School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, China. His research interests include TDOA sound source localization and DOA acoustic signal estimation.

Zhongshan Wang received the B.S. degree in mechanical engineering from Tiangong University, Tianjin, China, in 2022. He is currently pursuing the M.S. degree with the School of mechanical engineering, Northwestern Polytechnical University, Xi'an, China. His research interests include deep learning and acoustic target detection.

Di Meng received the B.S. degree in mechanical engineering from the Xi'an University of Technology, Xi'an, China, in 2020, and the master's degree in mechanical design and theory from Northwestern Polytechnical University, Xi'an, in 2023, where he is currently pursuing the Ph.D. degree. His research interests are acoustic measurement and array signal processing.

Penghao Xie received the B.S. degree in mechanical design manufacturing and automation from Yichun University, Yichun, China, in 2020, and the master's degree in mechanical engineering from Donghua University, Shanghai, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an, China. His research interests are acoustic measurement and array signal processing.