Non-Line-of-Sight Vehicle Localization Based on Sound

Mingu Jeon, Jae-Kyung Cho, Hee-Yeun Kim, Byeonggyu Park, Seung-Woo Seo, *Member, IEEE*, and Seong-Woo Kim, *Member, IEEE*

***Abstract*— Sound can be utilized to gather information about vehicles approaching a Non-Line-of-Sight (NLoS) region that remains hidden from Line-of-Sight (LoS) sensors due to its reflective and diffractive characteristics, like a radar. However, due to the inability to determine the location of NLoS vehicles in previous studies, it has not been possible to construct a sound-based active emergency braking system. This paper introduces a novel approach for localization of vehicles approaching in NLoS regions through sound. Specifically, a new particle filter method incorporating Acoustic-Spatial Pseudo-Likelihood (ASPLE) has been proposed to track objects using both acoustic and spatial information from the ego vehicle. Also, the Acoustic Recognition based Invisible-target Localization (ARIL) dataset, which is the firstly providing the location of the NLoS vehicle as ground truth using Bird's Eye View camera, is proposed. The proposed method is validated using two datasets: the ARIL dataset and the Occluded Vehicle Acoustic Detection Dataset (OVAD) dataset. The proposed method exhibited remarkable performance in localizing NLoS targets in both datasets, predicting the location of the vehicle in the NLoS region. Lastly, the analysis of how the reflection of sound affects to the proposed method, highlighting variations based on the spatial situations, and demonstrate the empirical convergence of the method is described. Our code and dataset is available at https://github.com/mingujeon/NLoSVehicleLocalization.**

***Index Terms*— Intelligent vehicles, non-line-of-sight detection, particle filter, sound source localization.**

I. Introduction

PERCEPTION is a fundamental function in autonomous driving, tasked with recognizing the surrounding environment and obstacles. Commonly employed sensors for perception in LoS region include those capable of accurately measuring distance or providing intuitive spatial information, such as LiDAR, cameras (RGB, Depth, Infrared), and ultrasonic sensors [1], [2], [3].

Received 21 March 2024; revised 12 June 2024 and 19 October 2024; accepted 28 November 2024. Date of publication 11 December 2024; date of current version 4 February 2025. This work was supported in part by the National Research Foundation of Korea (NRF) through the Ministry of Science and Information and Communications Technology (ICT) under Grant 2021R1A2C1093957. The Associate Editor for this article was H. Wu. *(Corresponding authors: Seung-Woo Seo; Seong-Woo Kim.)*

Mingu Jeon, Hee-Yeun Kim, and Seung-Woo Seo are with the Electrical and Computer Engineering, Seoul National University, Seoul 08826, Republic of Korea (e-mail: mingujeon@snu.ac.kr; hiyeun@snu.ac.kr; sseo@snu.ac.kr).

Jae-Kyung Cho is with SK Telecom, Seoul 04539, Republic of Korea (e-mail: jackyoung96.snu@gmail.com).

Byeonggyu Park and Seong-Woo Kim are with the Graduate School of Engineering Practice and Integrated Major in Smart City Global Convergence, Seoul National University, Seoul 08826, Republic of Korea (e-mail: pbk5485@gmail.com; snwoo@snu.ac.kr).

This article has supplementary downloadable material available at https://doi.org/10.1109/TITS.2024.3510582, provided by the authors.

Digital Object Identifier 10.1109/TITS.2024.3510582

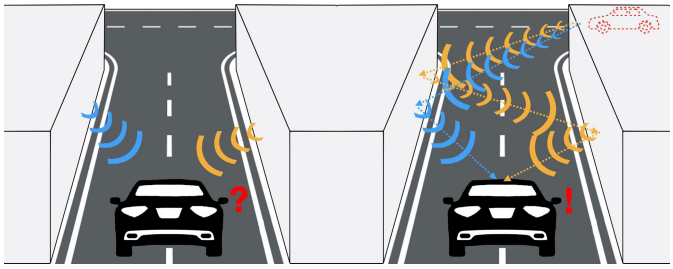


Fig. 1. **A schematic illustrating the challenging scenario when a vehicle is approaching from the NLoS region.** (a) Traditional methods fail to locate the NLoS vehicle due to the absence of spatial information. (b) The proposed method successfully identifies the location of the NLoS vehicle by calculating reflections and diffraction of sound, leveraging the spatial information.

However, an occluded area known as the NLoS region is inevitably created due to the straight path limitation of light-based sensors. Especially on narrow side roads, the NLoS region, which drivers cannot visually inspect, becomes wider. In such scenarios, a dynamic object suddenly emerging from the NLoS region is often unpredictable, posing a potential risk of accidents.

V2X communications, aimed at exchanging information to detect vehicles located in a NLoS region, have been explored for autonomous vehicles [4], [5]. Despite the intuitiveness of this approach, it is costly to implement and impractical for computer-less dynamic obstacles, such as pedestrians, bicycles, and motorcycles.

To solve the limitation of V2X communications, a radar has been utilized broadly to prevent accidents occurred by NLoS region. Radars acquire spatial information about their surroundings by transmitting and receiving reflected electromagnetic waves. A representative example of a system utilizing radars is the Active Blind Spot Assistant System, which detects blind spots during driving using radar and prevents accidents by assessing collision potential. Scheiner et al. [6] introduced a method for NLoS obstacle detection in an autonomous vehicle using a reflex radar signal and a data-driven approach. However, this method requires an additional active expensive device, namely radar.

In contrast, sound can be captured by passive, inexpensive and easily mountable microphone sensors. Additionally, spatial information can be deduced using echolocation even when the sound is outside the human audible frequency range. Moreover, sound, being less obstructed by obstacles due to its reflective and diffractive characteristics, can effectively convey information.

TABLE I
A COMPARISON OF NLOS VEHICLE DETECTION MODELS BASED ON SPATIAL AWARENESS STUDIES FOR OBJECT DETECTION

Reference	Description	Types	Frequency	Method	Outdoor vehicle	Passive sensor	Spatial information	Localization
Naser, *et al.* [7]	A detection model for NLoS vehicles based on detecting small changes in shadows and illumination has been proposed, and the model has been validated in relatively dark outdoor environments such as garage.	Visible Lights	400-800 THz	Rule-based	✓	✓		
Scheiner, *et al.* [6]	Radar signals reflected from the static objects are used to infer the location of obstacles located in NLoS region. The reflected radar signals are converted into an occupancy grid map, after which the positions of the objects are predicted by using a neural networks.	Radar	77 GHz	Data-driven, CNN	✓		✓	✓
Kim, *et al.* [4], [8]	The ego-vehicle receives information on unobservable moving objects from cooperative vehicles through V2X wireless communication.	WiFi	2-5 GHz	Rule-based	✓			
Choi, *et al.* [9]	Noise sources location in a multi-layer structure building is conducted by using a data-driven model with CNN architecture. Multi-channel sound is transformed to Mel spectrograms and used as input for the model.	Acoustic	∼20 kHz	Data-driven, CNN		✓		
Schulz, *et al.* [10]	The existance of NLoS vehicles are predicted through multi-channel sound by using the traditional beamforming method to extract features from the multi-channel sound and linear SVM to classify the final results. The OVAD dataset is also proposed.	Acoustic	∼1.5 kHz	Data-driven, SVM	✓	✓		
Hao, *et al.* [11]	A model using neural networks has been proposed not only for classifying the presence direction of NLoS vehicles at T-Junctions but also for classifying whether they are approaching or moving away. This model has been validated in an outdoor environment.	Acoustic	∼15 kHz	Data-driven, CNN	✓	✓		
Proposed	The location of NLoS vehicle are tracked by combining spatial information and multi-channel sound data. A particle filter method based on a novel pseudo-likelihood function is used to handle situations of sounds, and validated in ARIL dataset.	Acoustic	∼2 kHz	Rule-based, Particle filter	✓	✓	✓	✓

Methods for detecting NLoS objects based on sound have been extensively researched, primarily categorized into those utilizing spatial information and those leveraging acoustic information. Schulz et al. [10] proposed a data-driven classification framework and the OVAD dataset for detecting vehicles in NLoS regions. They trained a simple Support Vector Machine (SVM) based on beamforming response features to classify the likelihood of approaching vehicles in T-shaped alleys. Additionally, Hao et al. [11] proposed a framework utilizing neural networks to classify whether vehicles are approaching or moving away in NLoS regions and validated it using the OVAD dataset and directly acquired data. However, OVAD dataset provides no ground truth of vehicle position in NLoS area, which limits approaches into classification, not localization.

Researchers have also explored sound source localization using model-driven methods using cheap microphones. Measuring the Direction-of-Arrival (DoA) from an incoming sound wave using a microphone array to determine the azimuth direction of sound sources is one of the most common methods [12], [13], [14], [15]. Ward and Williamson [16] utilized a particle filter with sound measurements for sound source localization in an indoor environment. However, this method cannot be applied to the diverse spatial environments faced by autonomous vehicles since the spatial measurements were not combined with acoustic measurements. Table I summarizes a comparison of the existing research.

In this paper we propose a novel framework for localizing NLoS vehicle in outdoor by utilizing both passively obtained acoustic information and spatial information. Similar to the radar-based active blind spot assistant systems, to construct the same system based on sound, it is crucial to know the position information of the NLoS vehicle. The methods proposed by Schulz et al. [10] and Hao et al. [11] provide information about the presence and driving direction of NLoS vehicles, making them suitable for collision risk notice and alert systems. However, due to the lack of location information, it is not possible to construct an active emergency braking system, as shown in Table II. To tackle with this problem, we propose a particle filter with a novel likelihood function named ASPLE (**A**coustic-**S**patial **P**seudo-**L**ik**E**lihood).

TABLE II
FEASIBLE DRIVING ASSISTANCE SYSTEMS

Methods	**Providable Information**	**Feasible Assistance System**		
		Notice	Alert	Active Brake
Schulz *et al.* [10]	Presence	√		
Hao *et al.* [11]	Direction	√	√	
Proposed Method	Location	√	√	√

ASPLE is based on the recognition that sound can be perceived from multiple directions other than the actual sound source location due to reflection and diffraction. Fig. 1 illustrates a critical situation where a vehicle is approaching from the NLoS region at a T-junction. The traditional sound localization method struggles to accurately predict the sound source position due to the absence of spatial information. In contrast, by calculating the reflection and diffraction of sound based on spatial information, it becomes possible to predict the position of the sound source even when sound waves arrive from various directions. Beamforming is employed to compute the power of sound for all potential sound path directions, taking into account the reflection and diffraction of sound waves. Subsequently, the outcomes are converted into pseudo-likelihood by applying a likelihood-shaping function. Unlike conventional methods, ASPLE facilitates acoustic particle filtering in outdoor environments by leveraging spatial information solely through acoustic data. Regarding to the spatial information, when using OVAD dataset, the briefly

achieved spatial information was achieved through Google maps, and it shows reliable performance in NLoS vehicle localization.

We also propose the ARIL dataset, which is the first dataset providing the ground truth of the NLoS vehicle's position via Bird's Eye View (BEV) images. Acquired through the integration of a microphone array, radar, LiDAR, and camera systems, this dataset captures the dynamic movements of NLoS vehicles at T-Junctions from the viewpoint of the ego-vehicle. The ARIL dataset is expected to accelerate research in NLoS vehicle localization.

The contributions of this study are summarized as follows:

- A novel NLoS vehicle detection framework by combining acoustic information and surrounding spatial information through a particle filter is proposed.
- ARIL dataset, which is acquired for the localization of NLoS vehicle using bird's eye view image at the test bed where directly desinged and built, is first proposed.
- The acoustic pattern when a vehicle approaches from NLoS region by considering spatial information is analyzed.

The remainder of this paper is organized as follows. Section II provides a review of prior research focusing on the detection of objects in NLoS regions and the perception of spatial information through various wave types. Section III formulates the mathematical definition of the problem related to detecting obstacles in NLoS regions. Section IV presents a detailed explanation of particle filter-based methods that combine acoustic and spatial information. Subsequently, Section V outlines the datasets used to validate the proposed method, followed by the presentation of experimental results and discussions in Section VI and Section VII, respectively. Finally, Section VIII concludes the study by summarizing and evaluating the research content, along with indicating potential directions for future research.

II. Related Works

Recent autonomous vehicles utilize various sensors to acquire data in response to the increasing complexity of traffic. Consequently, numerous studies have been conducted to enhance the safety of driving using the acquired data. For example, methods have been developed to estimate the position of the ego-vehicle based on images acquired by cameras mounted on vehicles, thereby enhancing autonomous navigation [17]. Research has also focused on strengthening responses to uncertain situations faced by autonomous vehicles through the utilization of content-based image retrieval methods [18], [19]. Additionally, methods have been proposed to improve inaccurate GPS location information due to complex urban environments [20].

Unfortunately, autonomous vehicles rely heavily on LoS sensors such as LiDARs and cameras to obtain spatial information [21], [22]. For this reason, NLoS region inevitably occurs in which vehicles cannot be detected. Of course, there are methods for detecting NLoS vehicles using RGB images, such as the shadow formation or small illumination changes proposed by Naser et al. [7]. However, LoS sensors do not have a significant advantage in detecting NLoS objects due to their linearity. Yasuda and Ohama [23] showed that obtaining information on vehicles coming around blind corners greatly helps prevent collisions. In this section, we review the literature on the technological approaches for detecting obstacles in NLoS region and spatial awareness through analyzing various types of waves.

Sonar is a common approach for obtaining spatial information by transmitting sound waves in the kHz band and analyzing the reflected waves, which is similar to how a bat echolocates [24]. Railey et al. [25] detected and tracked unmanned underwater vehicles through sonar whereas on land, Rosique et al. [22] and Jang et al. [26] showed that an ultrasonic sensor (a kind of sonar) can be useful for short-distance detection in vehicle-assistance systems for activities such as parking and emergency braking. However, a sonar sensor has not been used for ground vehicle position tracking and far-distance detection on land because the sound waves bouncing off of dynamic objects cause high false-positives.

Instead, electromagnetic waves above the MHz bandwidth have been used for detecting obstacles in NLoS region on land. Peabody et al. [27] used an active radar sensor to obtain images of moving objects behind a wall by transmitting 2-4 GHz electromagnetic waves and analyzing the reflected signals. However, it is difficult to install this setup in a road vehicle because of the large radar transceiver. Similarly, Adib and Katabi [28] observed the movement of people behind a wall in a room by using the transmitted 20 MHz electromagnetic waves coming from the WiFi transmitter and capturing their reflection off of a human body. Human movement could be recognized by analyzing the reflected WiFi signals. However, it would be difficult to use this technique in an outdoor environment over a long distance because the WiFi signal strength is not strong enough and there is a lot of interference. Also, Palffy et al. [29] proposed a multimodal radar and camera fusion method for the detection of pedestrian behind parked vehicles.

Techniques for detecting NLoS region by sharing information rather than direct observation have been studied. Kim et al. [4], [8] proposed cooperative perception through V2X communications in which the ego-vehicle could receive information on unobservable moving objects through wireless communications, similar as Lee et al. [30]'s proposal using cellular networks. Safe and fast autonomous driving performance was shown through cooperative perception. However, it can only be used when a vehicle or infrastructure [31] capable of V2X communication and sensing exists.

Compared to the active methods that consist of a transmitter and a receiver, there is also a passive technology whereby only the surrounding sound information is received. Asahi et al. [32] proposed NLoS vehicle recognition method by calculating cross-power spectrum phase coefficient based on sound at the outdoor environment. Chen et al. [33] and Gan et al. [34] solved navigation tasks in the indoor environments through both camera images and sounds passively obtained by using a microphone. The authors trained Deep Reinforcement Learning (DRL) networks to make control decisions in the simulation based on camera images and sound data. Although it combined passively acquired sound and spatial information,

it did not enable the direct detection of objects in NLoS region. Mizumachi et al. [35] proposed a particle filter tracking method for intelligent vehicles based on sound obtained by using microphone arrays. Even though this method could predict the exact positions of other vehicles, it only detected objects in an open space like a highway, not in NLoS region. Chen et al. [36] proposed sound source localization method based on sound analysis and spatial information, however, it only considered sound diffraction and did not take into account sound reflections due to the surrounding terrain. Also, it was verified the feasibility only through simulations. Several studies have proposed localization methods that leverage both sound and spatial information. An et al. [37], [38] introduced a method for localizing objects using the reflective properties of sound and spatial information. However, the paper was conducted in a confined indoor environment of $7 \times 7\ m$, and it aims to localize objects in an open space rather than focusing on object localization in the NLoS area affected by obstacles, as addressed by the proposed method.

The state-of-the-art technology for detecting vehicles in NLoS region uses data-driven approaches. Choi et al. [9] estimated the noise source between floors beyond the wall using a fixed acoustic sensor in a data-driven manner. Scheiner et al. [6] analyzed electromagnetic waves from a radar that were double reflected from an object through surrounding walls. The authors trained the deep learning networks to infer the locations of obstacles in NLoS region from an occupancy grid map of the reflected radar signals. However, it is still an active method that requires both radar transmitter and radar receiver equipment. Schulz et al. [10] predicted the probability of the presence of vehicles in the front, left, and right sides of an alley in a T-shaped alley using a microphone array. After converting the recorded sounds to the output power responses with regard to the azimuth, linear SVM classified the vehicle positions based on the response feature vector. However, the method was only verified in a limited T-shaped alley environment. Moreover, Hao et al. [11] not only improve the performance of classification task, but also proposed the classification of NLoS vehicle approaching and leaving through the adaptation of neural networks. However, these data-driven methods require collecting lots of data. Since most autonomous vehicles use LoS sensors, datasets including NLoS information are rare. It is inevitable to artificially create data using multiple vehicles that can communicate.

In contrast to the previous methods, proposed method enables detecting objects in NLoS region by combining passively acquired acoustic information and spatial information. The next section shows the detailed pipeline of proposed method and explains it for combining acoustic and spatial information.

III. Target Scenario and Problem Definition

The target scenario addressed in this paper pertains to accidents occurring due to the inability to estimate the position of moving vehicles in the NLoS region of a T-Junction. Specifically, when a vehicle approaches the NLoS region of a T-Junction at a speed of 20 km/h, if the ego-vehicle is within 8 m from the intersection, human driver cannot avoid collision, irrespective of the driver's awareness of the NLoS vehicle. This is due to the necessity of a total stopping distance of 9 m comprising a thinking distance of 6 m and a braking distance of 3 m, to avert a collision, as illustrated in Fig 2. In such circumstances, minimizing the thinking distance of 6 m is essential for accident prevention. However, under the assumption of human judgment, there is no feasible method to minimize this distance.

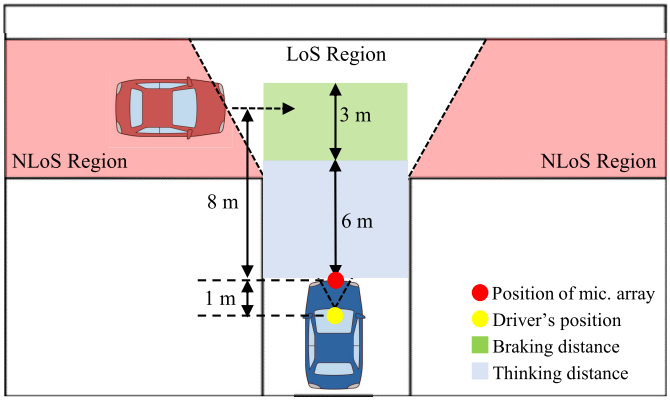


Fig. 2. **The target scenario, as a situation where a human driver inevitably gets into an accident when NLoS vehicle suddenly appeared.** When assuming a human driver is driving the vehicle at a speed of 20 km/h, the distance required to perceive and stop is a braking distance of 3 m (the green area), and a thinking distance of 6 m (the blue area), total stopping distance of 9 m. If a vehicle suddenly appears in the NLoS region as depicted in the red area, the remaining distance to collision is 8 m, which is shorter than the stopping distance, therefore, an accident would inevitably occur. However, if an active emergency braking system is utilized in this scenario, the braking distance becomes nearly equal to the stopping distance, thereby preventing accidents.

Hence, to prevent such accidents, active intervention by driving assist systems based on the assessment of collision potential is imperative to minimize the thinking distance. Similar to the radar-based Active Blind Spot Assistant System, a sound-based active emergency braking system requires three kinds of information: the presence of the NLoS vehicle, the driving direction of the NLoS vehicle, and the position of the NLoS vehicle. As the information regarding the presence and movement direction of the NLoS vehicle can be acquired using methods proposed by Schulz et al. [10] and Hao et al. [11], it is feasible to establish a system for notice and alert. However, due to the unknown location of the NLoS vehicle, it is not possible to assess the potential for collision, thereby rendering the construction of an active emergency braking system unfeasible. In other words, to develop a sound-based active emergency braking system, a method for localizing the NLoS vehicle is essential.

Therefore, the main target of this paper is to predict the location of NLoS vehicle $\hat{X}_t$ which is moving under 20 km/h speed on narrow road based on manually acquired measurements involving acoustic and spatial information. The problem can be formulated as follows:

$$\hat{X}_t = \text{argmax}\, p(X_t|O_{1:t}) = \text{argmax}\, p(X_t|A_{1:t}, M_{1:t}), \quad (1)$$

where X_t is the location of NLoS object at time step t. To get X_t, the most probable location of a NLoS object $p(X_t|O_{1:t})$ must be calculated based on observations $O_{1:t}$

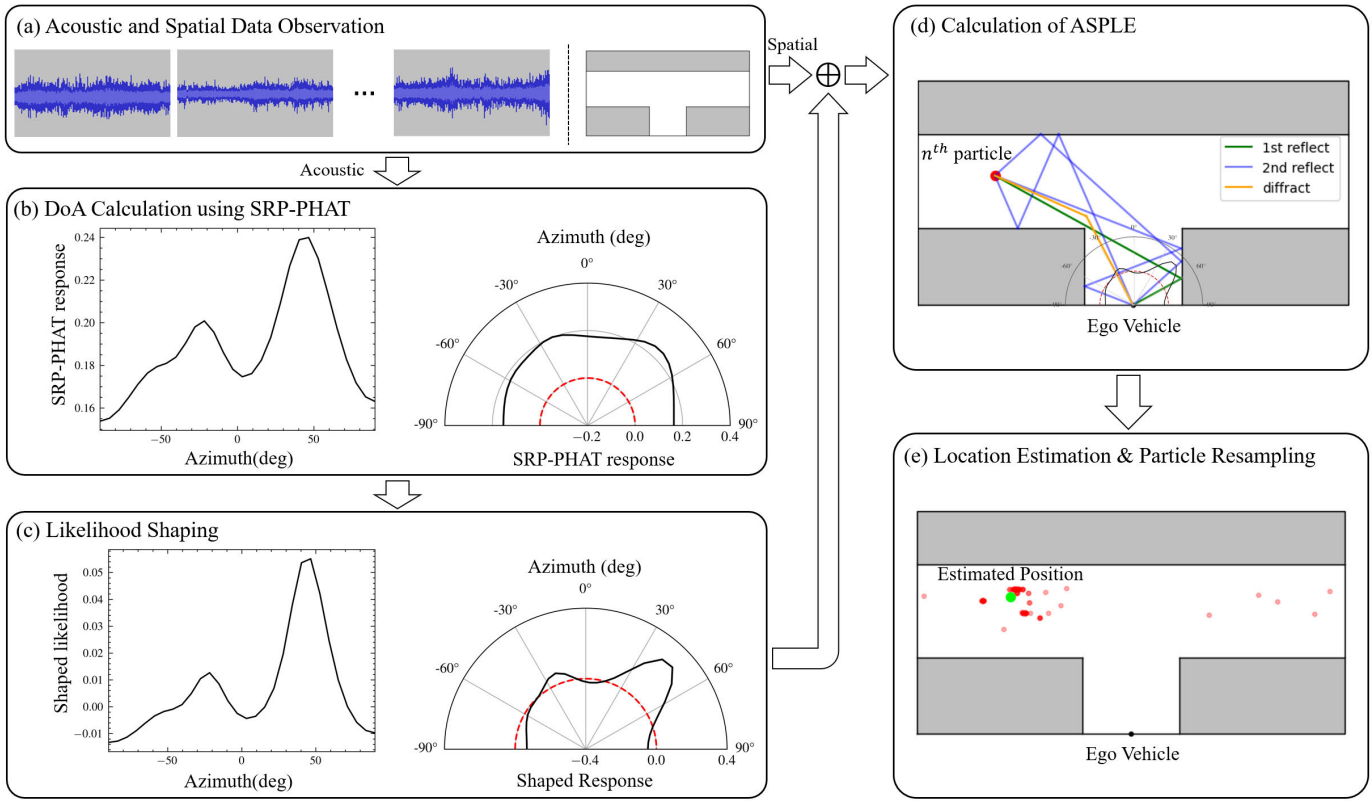


Fig. 3. **The overall framework for NLoS vehicle localization by using an acoustic-spatial particle filter.** (a) Acoustic observations from a multi-channel microphone array (left) and spatial observation (right), (b) Output power response calculations with respect to the azimuth angle by the Steered-Response Power Phase Transform (SRP-PHAT) algorithm. The maximum response direction commonly represents the Direction of Arrival (DoA), (c) The likelihood-shaping function emphasizing peak nodes and scaling to make the upper quartile value equal to zero, (d) Generation of multiple sound paths and calculation of ASPLE for each particles, (e) Estimate the position of NLoS vehicle and particle resampling. Red dots are particles and green dot represents the estimated position.

consisting of passively acquired sound information A_t, and spatial information M_t.

IV. Proposed Method

In this section, a novel acoustic-spatial particle filter framework designed to predict the position of a vehicle approaching from a NLoS region is introduced. The sound paths from each particle to the ego-vehicle undergo distortion due to interactions with spatial objects, and a new pseudo-likelihood and likelihood-shaping function are proposed to analyze these distorted sound paths. The overall framework is illustrated in Fig. 3, and the particle filter algorithm, encompassing these three steps, is summarized in Algorithm 1.

A. Acoustic-Spatial Particle Filter

Let the true state of a dynamic vehicle at time t be X_t. Our ultimate goal is to determine posterior density $p(X_t|O_{1:t})$, which enables us to calculate the most probable location of a dynamic vehicle based on the entire observations $O_{1:t}$. This can be calculated as

$$p(X_t|O_{1:t}) \propto p(O_t|X_t)p(X_t|O_{1:t-1}), \tag{2}$$

where $p(O_t|X_t)$ is the likelihood and $p(X_t|O_{1:t-1})$ is the prior distribution. This can be calculated through transition probability $p(X_t|X_{t-1})$ and the previous posterior as follows:

$$p(X_t|O_{1:t-1}) = \int p(X_t|X_{t-1})p(X_{t-1}|O_{1:t-1})dX_{t-1}. \tag{3}$$

Algorithm 1 Applying the Particle Filter for Vehicle Detection in a NLoS Region by Using Sound

Initialize particles $X_0 = \{X_0^{(i)},\ i = 1:N\}$
Initialize weights $w_0 = \{w_0^{(i)} = 1/N,\ i = 1:N\}$
for t=1:T **do**
  $\tilde{Y}_t \leftarrow$ *SRP-PHAT*(O_t)
  for n=1:N **do**
    $\tilde{X}_t^{(n)} \leftarrow$ *propagate*$(X_{t-1}^{(n)})$
  end for
  for n=1:N **do**
    calculate $\mathbf{L}(X_t^{(n)}, M_t)$
    $w_t^{(n)} \leftarrow$ *ASPLE*$(X_t^{(n)}, \tilde{Y}_t)$
  end for
  for n=1:N **do**
    $X_t^{(n)} \leftarrow$ *Resample*$(\tilde{X}_t, w_t)$
  end for
end for

Unfortunately, there is no closed-form solution for Equation (2) and Equation (3), as the distribution types of the likelihood, transition probability, and prior are all unknown. To address this challenge, Monte Carlo sampling is employed to approximate the posterior density in the particle filter. In this process, a set of particles, initialized randomly, are individually propagated and resampled based on the likelihood.

To apply the particle filter to the proposed task, three essential steps must be defined: propagating particles, defining the likelihood function, and resampling the particles. First, the particles are propagated based on a constant turn rate and velocity (CTRV) model, which is a simple vehicle dynamic model [39]. The second requirement is to define a proper likelihood function for calculating the importance weight for each particle. To obtain the likelihood distribution $p(O_t|X_t)$ when the distribution type is unknown, Ward and Williamson [16] proposed a pseudo-likelihood function based on the delay-sum beamformer (DSB) as follows:

$$p(O_t|X_t) = \phi(|\bar{Y}(\ell)|^2), \tag{4}$$

where the $|\bar{Y}(\ell)|^2$ is the output power of a DSB steered azimuth ℓ which is the angular direction to source location X_t, and $\phi(x)$ is the likelihood-shaping function for concentrating the distribution around the peaks, which can be formulated as:

$$\phi(x) = x^i, \ \ i = 2, 3, or\ 4. \tag{5}$$

where the parameter i was empirically determined, as referenced from Ward and Williamson [16].

However, the pseudo-likelihood proposed by Ward and Williamson [16] is the likelihood for the direction of the sound source, not for the location of the sound source. In order to find the exact source location, it is necessary to find the intersection of each direction through spread pairs of microphones over several locations. In the case of scenario where the microphone array is concentrated in the ego-vehicle, the location cannot be determined even if the direction can be known. Thus, a new likelihood function that can consider both the acoustic signal and spatial map information is required to handle situations where the microphone array is concentrated on the ego-vehicle in an outdoor environment.

First of all, all possible sound paths at time t from each particle to the microphone can be found based on map information $M_t \in O_t$. Only direct paths $\{\ell_{X_t}\}_{\text{dir}}$, single reflection paths $\{\ell_{X_t}\}_{1^{\text{st}}}$, double reflection paths $\{\ell_{X_t}\}_{2^{\text{nd}}}$, and diffraction paths $\{\ell_{X_t}\}_{\text{diff}}$ are considered. Since the method focuses on low-frequency sound, diffraction must be considered. The set of paths is converted to a set of azimuths, which is represented by

$$\mathbf{L}(X_t, M_t) = \{\ell_{X_t}\}_{\text{dir}} \cup \{\ell_{X_t}\}_{1^{\text{st}}} \cup \{\ell_{X_t}\}_{2^{\text{nd}}} \cup \{\ell_{X_t}\}_{\text{diff}}, \tag{6}$$

where $M_t \in O_t$ is the map information. The map consists of lines (walls) and an intersection of lines (corners). The reflection path is simply obtained by mirroring the source against the line. The diffraction path is obtained by finding the path through the corner. The largest hurdle against defining the ASPLE is that the size of the set $\mathbf{L}(X_t, M_t)$ for each particle is not the same. Hence the proposed novel pseudo-likelihood function, ASPLE, is used to handle this problem as follows:

$$\text{ASPLE}(\mathbf{L}, \bar{Y}_t) = \sum_{\ell \in \mathbf{L}(X_t, M_t)} \alpha(\ell) \cdot \phi_{\text{ASPLE}}(|\bar{Y}_t(\ell)|^2), \tag{7}$$

where $\alpha(\ell) \in \{\alpha_{\text{dir}}, \alpha_{1^{\text{st}}\text{-ref}}, \alpha_{2^{\text{nd}}\text{-ref}}, \alpha_{\text{diff}}\}$ is a weight parameter that depends on the type of sound path.

The last requirement is to resample the particles based on the ASPLE functions. The weight of each particle is assigned proportionately according to the ASPLE value in the previous stage. If the weight values of the particles are $[w^{(1)}, w^{(2)}, \ldots, w^{(N)}]$, the probability of resampling the j^{th} particle is $w^{(j)} / \sum_{n=1}^{N} w^{(n)}$. Moreover, random sampling noise with probability ϵ is added to ensure robustness and prevent the particles collapsing.

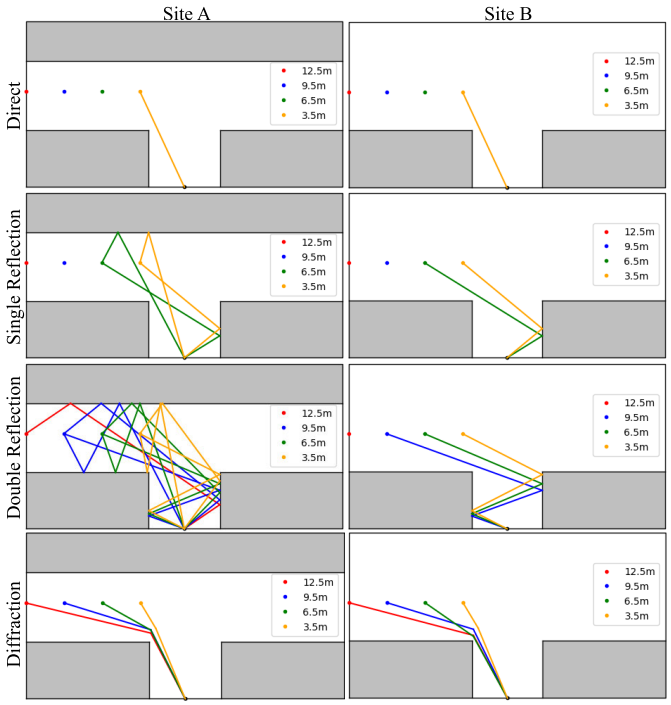


Fig. 4. **Comparison of possible paths of sound regarding the position of NLoS vehicle and spatial configuration in ARIL dataset.** At *Site A*, sound can be detected through both diffraction and double reflection paths, however, at the position 9.5 *m*, sound can be detected throught more double reflection path. Also, Compare with *Site A* and *Site B*, sounds coming from a distance of over 9.5 *m* can be detected not only through diffraction but also through double reflection, due to the presence of the opposite-side wall.

B. Decision of Weight Parameter $\alpha(\ell)$

The response of sound varies depending on the proximity of the sound source, the occurrence of reflections, and the degree of diffraction due to distance. In other words, the distance, the number of reflections, and the amount of diffraction based on distance all influence the probability of sound sources existing at each particle.

Furthermore, the types of detectable sound paths are limited depending on the location of the particle, as shown in Figure 4. For example, at the 6.5 *m* point of *Site A*, sound can be detected through single reflection, double reflection, and diffraction paths, unlike at the 9.5 *m* point where sound can be detected through double reflection, and diffraction paths. Additionally, sound cannot be detected through direct paths at any point other than the 3.5 *m* point. This limitation in the types of detectable sound paths depending on distance signifies their primary impact on estimating the location of the sound source.

Accordingly, the weight parameter $\alpha(\ell)$ is relatively set based on the possible sound paths generated depending on the particle's position, taking into account the contributions from

sound reflections and diffraction. To elaborate, the direct sound path is feasible when the sound source is in the LoS region and contributes the most to the probability of the sound source being present at that particle's location, as it is least affected by the physical environment. Additionally, for reflection sound paths, the contribution to the probability of a sound source being present for a single reflection is higher compared to double reflection, due to the decay of sound waves through reflection. Lastly, sound decaying due to diffraction increases significantly with distance, and decreases as the sound source approaches. Therefore, the contribution of diffraction to the probability of a sound source being present is lower for particles located far away, but increases as they approach. Taking all these factors into account, $\alpha(\ell)$ is set relative to reflect the possible sound paths generated depending on the particle's location, giving higher contributions to direct path, single reflection path, diffraction path, and double reflection path in that order, as the distance decreases. The values of the parameter set used in our experiment are listed in Section VI.

C. Likelihood-Shaping Function

Since the size of the azimuth set $\mathbf{L}(X_t, M_t)$ for each particle is different, it is difficult to apply the likelihood-shaping function in Equation (4) directly to ASPLE. A particle with a larger set size will have a larger ASPLE value because $|\bar{Y}(l)|^2$ is larger than zero. Therefore, particles with a large set size will be resampled more frequently regardless of the real likelihood of the measurement.

To handle this problem, a modified likelihood-shaping function is formulated as follows:

$$\phi_{\text{ASPLE}}(x) = x^4 - Q_3(x^4), \tag{8}$$

where the $Q_3(x^4)$ is the upper quartile value of x^4, which is empirically chosen. An exponential function fulfills a similar role to the original likelihood-shaping function by emphasizing the peak nodes. By subtracting the upper quartile value, only the azimuth bins that are larger than the upper quartile value increase the ASPLE value whereas the others decrease it. It is intended to decrease the likelihood of particles containing invalid azimuth values, regardless of the size of the azimuth and path set, so that the likelihood of particle sets containing only sound paths would increase.

D. Decreasing the Computation Burden for ASPLE

When calculating the ASPLE, the output power of a DSB Y_t should be calculated for each particle. This results in unnecessary calculations because lots of particles share similar azimuth elements in their own azimuth set. Furthermore, an expensive computation burden is required to handle the larger number of particles essential for making particle filter accurate. To solve this, output power of a DSB for n discrete azimuth bins ($\bar{Y}_t = [Y(l_1), Y(l_2), \ldots, Y(l_n)]$) can be calculate beforehand. Scheibler et al. [40] provided an Steered-Response Power Phase Transform (SRP-PHAT) algorithm to easily calculate this. Only the frequency bandpass of 50∼2000 Hz is used because it is the dominant frequency range of low-speed vehicles due to tire-road contaction [41]. Discrete bin values (from 0° ∼ 179° with a 1° interval) are linearly interpolated to obtain values for a continuous azimuth range; this trick reduces the computation burden dramatically because the number of sound paths for all of the particles is much larger than the number of bins.

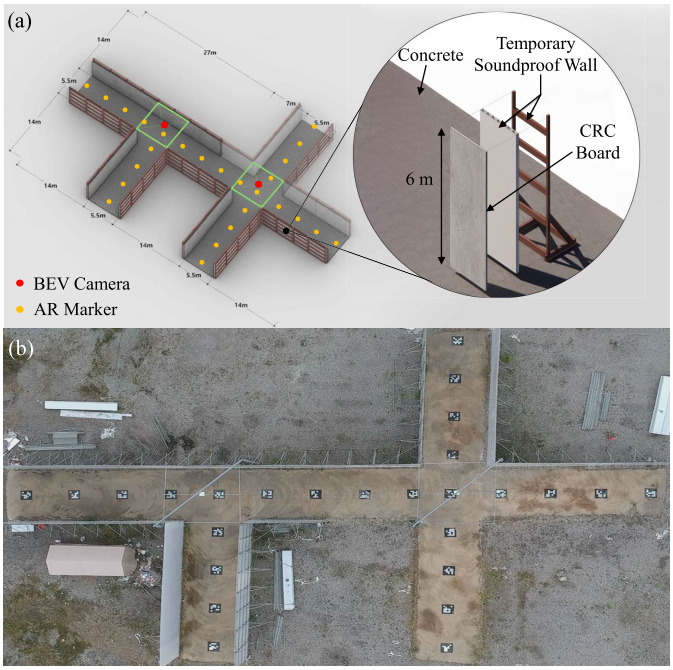


Fig. 5. **Blueprints, schematic layouts, and tangible photographs depicting the design and configuration of the test bed.** (a) A blueprint illustrating the configurations, including the attachment points for AR markers and BEV camera, dimensions of the test bed components, as well as the height and layout of the floor and temporary soundproof walls. (b) Aerial view photo of the test bed.

V. Datasets

To assess the validity of the proposed approach, we utilized two datasets, ARIL dataset and OVAD dataset, as follows. *Site A* denotes T-Junction with a wall environment, and *Site B* denotes T-Junction without a wall environment.

A. Acoustic Recognition Based Invisible-Target Localization (ARIL) Dataset

The ARIL dataset was procured to advance the study of a model dedicated to the detection and localization of NLoS objects. This dataset was amassed within a carefully constructed test bed designed to mirror the conditions encountered on actual roads, as depicted in Figure 5. An SUV outfitted with a microphone array, radar, LiDAR, and camera, served as the primary vehicle for data collection.

The dataset is centered on identifying NLoS objects, with a particular focus on vehicles, and encompasses variations in velocity (5, 10, 15, 20 km/h), the direction of travel for NLoS vehicles (left, right), and different spatial arrangements (*Site A* and *Site B*). Each scenario is documented through a series of data captures: one sound sequence, two point cloud sequences, one fisheye image sequence, and one BEV image sequence. The audio data was captured at a fidelity of 48,000 Hz, while

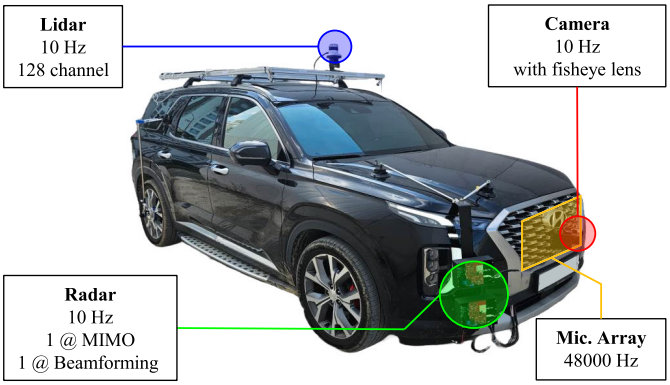


Fig. 6. **Sensors configuration of ego-vehicle.** To acquire data, two radars, two pairs of microphone arrays, a 12-megapixel camera with a fisheye lens, and one LiDAR were mounted on the ego-vehicle. The data acquisition frequency was set at 48,000 Hz for the microphone array and 10 Hz for the other sensors. The positions of the sensors were carefully arranged to mimic the real driving environment.

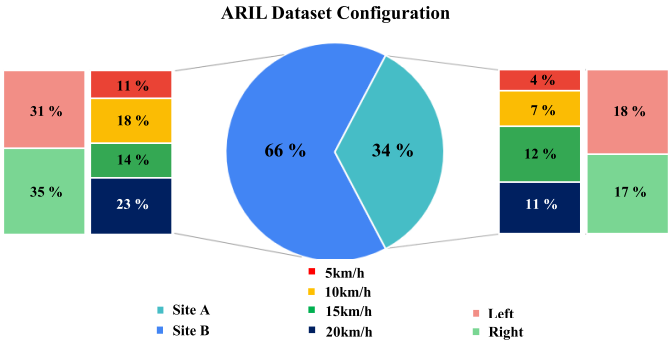


Fig. 7. **Distribution of acquired data.** Across an estimated timeframe of 2,680 seconds, a comprehensive set of 136 data entries was meticulously gathered. The accompanying chart delineates the collective distribution concerning the spatial layout, the trajectory and the velocity of the NLoS vehicle.

TABLE III
TOTAL AMOUNT OF ACHIEVED DATA FOR EACH CLASS

Spatial Configuration	Direction		Velocity [km/h]				Total
	Left	Right	5	10	15	20	
Site A	24	23	6	10	16	15	47
Site B	42	47	15	24	19	31	89
Total	66	70	21	34	35	46	136

the remaining data sets were collected at a frequency of 10 Hz. Detailed descriptions for each component are as follows.

1) Test Bed: The effectiveness of the ASPLE model was assessed through the establishment of a test bed, spanning 53.5 $\times$ 33.5 m. This environment was meticulously designed to simulate varied road conditions by incorporating both T-Junction configurations, with and without opposite-side wall. The drivable areas were constructed from concrete to closely mimic real road surfaces, while the walls were engineered to reflect the aesthetics and structural characteristics of actual buildings. This was achieved by erecting support structures adorned with temporary acoustic barriers, which were subsequently covered with cellulose-reinforced cement panels, comprising primarily of cement and cellulose. Whole layouts are depicted in Figure 5.

Furthermore, to ensure meticulous data collection and to encapsulate the entirety of the test bed within a singular view, a 12-megapixel camera equipped with a fisheye lens offering a 185° field of view was utilized as the BEV Camera. This camera was strategically mounted at an elevation of 7 m above the intersection's center within each spatial layout. To mitigate the distortional effects introduced by the camera's mounting angle and the fisheye lens, which complicates the precise tracking of NLoS objects, 26 Augmented Reality (AR) markers, each measuring 1 $\times$ 1 m, were positioned on the ground at intervals of 4 m. This arrangement facilitates the accurate determination of object positions.

To acquire data, two radars, two pairs of microphone arrays, a 12-megapixel camera with a fisheye lens, and one LiDAR were mounted on the ego-vehicle.

2) Hardware and Software: The data collection process involved the customization of an SUV to incorporate various sensors, with the objective of closely replicating authentic driving conditions, as illustrated in Figure 6. Initially, a microphone array utilizing the UMIK-X model by miniDSP was deployed. This setup included two arrays, each comprising 8 MEMS microphones, positioned at intervals of one meter at the grille of SUV. Furthermore, to ascertain the visibility status of objects as either NLoS or LoS, a fisheye camera was affixed to the vehicle's grille center. For the purpose of enhancing future studies, although not discussed within the current investigation, two radar units and a LiDAR system were installed. To address the challenges associated with processing the significant volumes of data collected, a laptop equipped with an i9-12900H CPU and 32 GB of RAM was utilized. The system operated on Ubuntu 20.04, employing ROS Noetic for the synchronized integration of data across all sensors.

3) Dataset Acquisition Conditions and Configurations: A cumulative total of 136 data points were systematically collected over a span of roughly 2,680 seconds, focusing on the low-velocity driving conditions of NLoS vehicles, not exceeding 20 km/h, on constrained roadways, as described in Section III. The dataset distribution is visually represented in Figure 7 and Table III. The positioning of the primary vehicle was determined to be 8 m from the center of the intersection, a decision influenced by the calculated stopping distance for the vehicle under these specific assumptions. The tracking of NLoS vehicles incorporated their emergence from either the left or right directions, with their initiation points set at 17 m away from their respective starting directions, proceeding towards their intended destinations. The speeds of the NLoS vehicles were standardized at intervals of 5, 10, 15, and 20 km/h, with the data collection process accommodating varying meteorological conditions, including periods of clear, overcast, and snowy weather.

B. OVAD Dataset and Baseline

Schulz et al. [10] developed a specialized dataset tailored for the T-junction alley scenario, where vehicles emerge unexpectedly from a NLoS area around a corner. This dataset encompasses synchronized sound data spanning one second across 56 channels, alongside corresponding camera imagery. Each data entry was annotated with one of

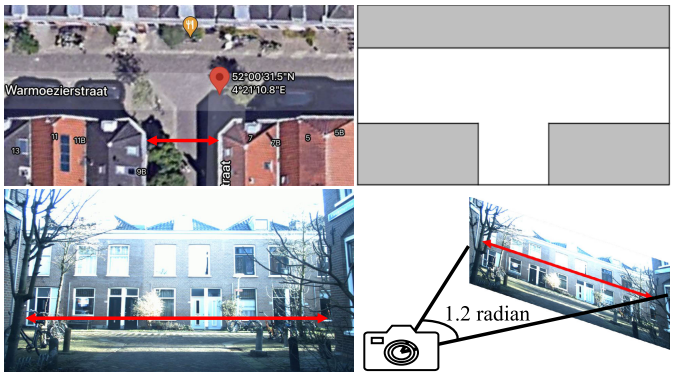


Fig. 8. **Extraction of spatial information of OVAD dataset using Google Maps.** Spatial information creation by using Google Maps, GPS positioning, and the camera configuration at OVAD dataset [10].

four directional labels–{$left, right, front, none$}–indicating the presumed approach direction of an external vehicle. The dataset encapsulates recordings from four distinct environmental settings: {SA, SB, DA, DB}. Here, S signifies a stationary environment with the ego-vehicle at a halt, while D denotes a dynamic environment with the ego-vehicle in motion, approximately at a 10 km/h speed. A represents a closed-walled T-junction alley, and B an open-walled T-junction alley. Data collection efforts were conducted from two locations for environment A and three locations for environment B. Schulz et al. [10] introduced a machine learning strategy leveraging a SVM trained on SRP-PHAT output response features derived from the acoustic data to classify the direction from which another vehicle would appear.

Owing to the absence of precise spatial data within the dataset, two-dimensional spatial information was extracted from Google Maps, leveraging GPS coordinates and the established camera setup, as depicted in Figure 8. For the purposes of comparative analysis, only the dataset representing static scenarios, denoted as {SA, SB}, was utilized. This limitation was due to the challenge of accurately determining the precise position and velocity of the ego-vehicle in dynamic scenarios.

VI. Experimental Results

Initially, the proposed method is applied to the ARIL dataset, which was generated by direct construction of test bed and collecting data through a 16-channel microphone array. This application aimed to confirm the effectiveness of the proposed NLoS vehicle localization method. Subsequently, it was conducted both classification and tracking tasks on the OVAD dataset [10], which features synchronized camera images and recorded 56-channel sound using a planar microphone array. This comprehensive evaluation allowed us to predict the direction and position of approaching vehicles from NLoS regions, thereby validating the broad applicability of proposed method.

A. Results on ARIL Dataset

In the ARIL dataset, experiments were conducted to validate the proposed method for the tracking task, and the experimental design is outlined as follows. Initially, Regions of Interest (RoI) were defined for each spatial configuration. This is due to the fact that the proposed method is specifically designed to consider sounds only coming through direct, single reflection, double reflection, and diffraction paths.

TABLE IV
Weight Parameter Settings in ARIL Dataset

Site	RoI [m]	Time [s] (Region)	α_{dir}	$\alpha_{1^{\text{st}}\text{-ref}}$	$\alpha_{2^{\text{nd}}\text{-ref}}$	α_{diff}
Site A	±12.5	0 ∼ 1.5 (*Far*)	0	0	1	0.8
		1.5 ∼ 2.0 (*Middle*)	1	0.7	0.5	0.5
		2.0 ∼ 7.0 (*Near*)	1	1	0.25	0.7
		7.0 ∼ 7.5 (*Middle*)	1	0.7	0.5	0.5
		7.5 ∼ 9.0 (*Far*)	0	0	1	0.8
Site B	±9.5	0 ∼ 0.7 (*Far*)	0	0.3	1	1
		0.7 ∼ 1.9 (*Middle*)	1	0.7	0.3	0.7
		1.9 ∼ 5.5 (*Near*)	1	0.6	0.2	0.7
		5.5 ∼ 6.9 (*Middle*)	1	0.7	0.3	0.7
		6.9 ∼ 8.0 (*Far*)	0	0.3	1	1

In further detail, in *Site A*, when the NLoS vehicle is situated beyond 12.5 m, sound cannot be detected at the microphone array location through direct, single reflection, or double reflection paths. Only paths involving three or more reflections and diffraction can result in detection. As this goes beyond the range of sound paths considered by ASPLE, in *Site A*, the RoI was established from the left 12.5 m to the right 12.5 m, utilizing the center of intersection as the reference point.

In *Site B*, the RoI was defined from the left 9.5 m to the right 9.5 m for analogous reasons. This more restricted range was selected due to the absence of a wall on the opposite side of the ego-vehicle, resulting in a diminished area where sound paths, including double reflection, could be identified. It is discussed in Section VII-A with more detail, and depicted in Figure 4.

Furthermore, even within the limited Region of Interest (RoI), detectable paths of reflected and diffracted sound differ as described in Section IV-B. Therefore, based on the understanding that primary sound paths vary across different areas of the ARIL dataset, weight parameters $\alpha(\ell) \in \{\alpha_{\text{dir}}, \alpha_{1^{\text{st}}\text{-ref}}, \alpha_{2^{\text{nd}}\text{-ref}}, \alpha_{\text{diff}}\}$ for each segmented RoI were assigned, as outlined in Table IV. In practical scenarios, characteristics of sound intensity were utilized for a rough estimation of location. Segmentation criteria were determined by the observation that sound tends to be quieter when the NLoS vehicle is distant and increases as it approaches, utilizing sound intensity in the acoustic spectrogram for segmentation.

Consequently, the spatial extent of NLoS vehicle presence based on the sound sequence was categorized into three distinct regions: *Far*, *Middle*, and *Near*. The *Far* region encompasses areas where sound can be detected through

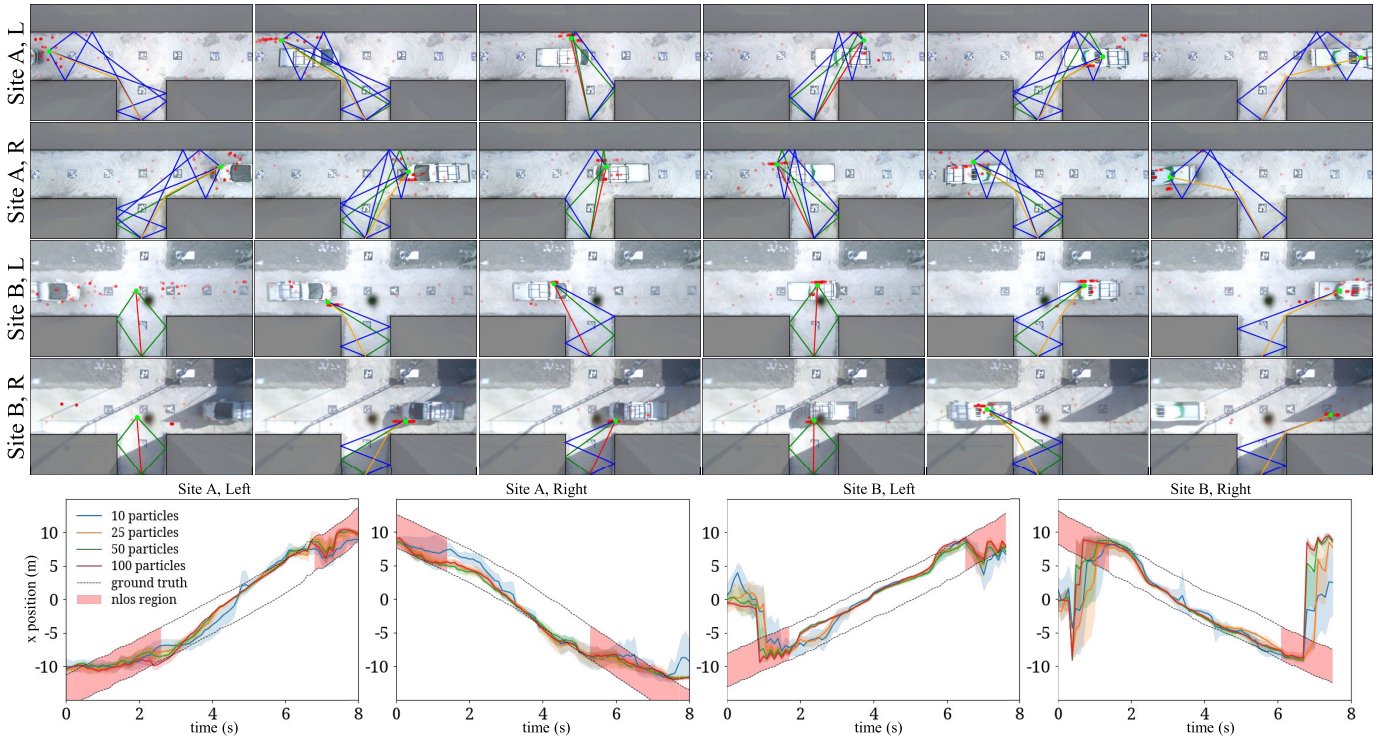


Fig. 9. **Qualitative and Quantitative results in the ARIL dataset.** At the upper images, the estimated position of the NLoS vehicle (green dot), dominant sound paths (red: direct, green: single reflection, blue: double reflection, orange: diffraction), and the BEV image as the ground truth for the qualitative comparison are presented for each scenario. It can be observed that the predictions fail in the early and late frames in *Site B* scenarios of the third and fourth rows, which is attributed to the left-right symmetry of the response map due to the position of the NLoS vehicle and the spatial characteristics. L and R denotes the left and right approaching direction of NLoS vehicle. The below graphs are the quantitative results, and in case of *Site B*, the results fluctuate in the early and late stages. The qualitative outcomes for these phenomenon can be observed in the first and last images of the third and fourth columns.

double reflection and diffraction paths. In the *Middle* region, sound detection is possible via single reflection, double reflection, and diffraction paths. The *Near* region assumes that sound can be detected through direct, single reflection, double reflection, and diffraction paths. These delineations served as the basis for conducting experiments.

In detail, the decision process was as follows: In the *Far* region of *Site A*, where the sound source is distant, causing significant decay due to diffraction, α_{diff} was set lower compared to $\alpha_{2^{\text{nd}}\text{-ref}}$. For the *Middle* region, considering that NLoS vehicles are not point-like in actual and may expose parts of the vehicle to the LoS area, sound can be detected through direct paths. Therefore, α_{dir} was set highest, followed by decreasing values for $\alpha_{1^{\text{st}}\text{-ref}}$ and $\alpha_{2^{\text{nd}}\text{-ref}}$. As the distance decreased, α_{diff} was set equal to $\alpha_{2^{\text{nd}}\text{-ref}}$. In the *Near* region, where the difference between direct paths and paths via single reflection is not significant, α_{dir} and $\alpha_{1^{\text{st}}\text{-ref}}$ were set equally.

In the case of *Site B*, in the *Far* region, the distance from the boundary of RoI to the corner is closer compared to *Site A*. Therefore, α_{diff} was set equal to $\alpha_{2^{\text{nd}}\text{-ref}}$, considering potential sound detection through single reflection paths in certain areas due to the size of the NLoS vehicle. For the *Middle* region, similar to *Site A* but without an opposite-side wall, α_{diff} was set higher than $\alpha_{2^{\text{nd}}\text{-ref}}$. Lastly, in the *Near* region, fewer reflection paths were considered due to its spatial characteristics.

With comparing a sound path map and a BEV image as shown at the upper images in Figure 9, it is evident that the proposed method adeptly tracks NLoS vehicles, especially when comparing it to the BEV image obtained concurrently with the sound path map. For quantitative analysis, the weighted average sum of particle positions X_t^i and their respective weights w_t^i, denoted as X_t^{avg}, was employed to estimate the location of NLoS vehicles. This estimation is formulated as follows:

$$X_t^{avg} = \frac{\sum_{i=1}^{N_p} w_t^i X_t^i}{\sum_{i=1}^{N_p} w_t^i}, \tag{9}$$

where N_p is the number of particles.

Additionally, to evaluate whether the proposed method successfully predicted the location of the NLoS vehicle throughout the entire sequence of each scenario and within predefined regions, the Prediction Success Rate (PSR) was used as a metric. The formulation is as follows:

$$PSR = \frac{N_{Succ}}{N_{Total}}, \tag{10}$$

where N_{Succ} is the number of frames in which the prediction was successful within the size of the NLoS vehicle, and N_{Total} is the total number of frames in the corresponding segment.

At the below graphs in Figure 9 illustrates the quantitative outcomes for four distinct scenario cases, categorized according to spatial configuration and the direction of the NLoS vehicle in the ARIL dataset. The red region delineates the NLoS area, and dashed lines are ground truth. In instances where the NLoS vehicle is moving to the right, the upper

TABLE V
QUANTITATIVE RESULTS IN ARIL DATASET[1,2]

Site	Dir.	Metrics	Number of Particles				Max DoA
			10	25	50	100	
Site A	Left	RMSE [m]	1.92	2.78	2.41	1.99	16.98
		Variance [m]	1.18	0.68	0.58	0.53	-
		Succ. Rate	68.9	75.6	76.1	79.3	44.3
	Right	RMSE [m]	3.30	1.83	2.02	1.87	17.81
		Variance [m]	1.10	0.74	0.50	0.44	-
		Succ. Rate	66.8	68.8	73.9	73.5	32.0
	Execution Time [ms]		70.3	109	180	314	0.01
Site B	Left	RMSE [m]	3.13	2.54	2.42	1.95	16.08
		Variance [m]	2.17	2.08	2.09	1.95	-
		Succ. Rate	68.8	75.7	76.6	76.1	33.3
	Right	RMSE [m]	2.80	2.47	2.36	2.09	16.38
		Variance [m]	1.68	1.59	1.44	1.37	-
		Succ. Rate	66.9	76.6	71.6	74.8	20.8
	Execution Time [ms]		58.9	71.24	109	171	0.01

TABLE VI
PSR RESULTS ACCORDING TO REGIONS IN THE ARIL DATASET

Site	Dir.	Regions	Number of Particles				Max DoA
			10	25	50	100	
Site A	Left	*Far*	61.8	67.2	59.3	68.3	5.0
		Middle	75.0	86.0	82.0	88.0	10.0
		Near	71.6	78.2	84.2	83.6	74.0
	Right	*Far*	41.0	51.5	63.9	64.1	0.0
		Middle	76.0	74.0	84.0	84.0	0.0
		Near	75.8	75.0	76.0	75.6	52.0
Site B	Left	*Far*	23.8	18.8	21.3	20.7	0
		Middle	67.7	77.0	78.9	77.0	0
		Near	89.5	100	99.5	100	72.2
	Right	*Far*	22.9	11.5	21.5	22.9	0
		Middle	77.0	90.0	83.1	86.2	0
		Near	76.7	92.3	82.8	86.7	44.4

TABLE VII
EXECUTION TIME OF SRP-PHAT WITH REGARD TO THE NUMBER OF MIC

Number of microphone	7	14	28	56
Execution time [ms]	12.06	19.49	37.22	93.64
Cosine similarity	0.993	0.998	0.999	1

dashed line corresponds to the vehicle's front, and the lower dashed line corresponds to the rear. Conversely, when the direction is to the left, the upper dashed line represents the vehicle's rear, and the lower dashed line corresponds to the front. This ground truth is defined as region because the proposed method relies on sound, which can originate from any part of the NLoS vehicle.

Table V offers an analysis of the root-mean-squared error (RMSE), variance, prediction success rate, and execution time concerning the ground truth as the center of the NLoS vehicle and X_t^{avg}. The baseline used for comparison is the maximum DoA localization method [13]. The number of particles influences the accuracy and reliability of the proposed method due to the randomness characteristic in the particle filter. Therefore, the RMSE, variance, prediction success rate, and execution time concerning the number of particles were measured.

The maximum DoA localization method was utilized to predict the location of the sound source in the direction where the output power value of SRP-PHAT is the largest. Although the baseline method cannot obtain the 2-D coordinate position, the x-coordinate position was obtained by fixing the y-coordinate as the ground truth.

For RMSE and variance, it was observed that the prediction errors consistently remained smaller than the size of the NLoS vehicle, which is 5.1 m, across all scenario cases, regardless of the number of particles in the particle filter. Furthermore, for accident prevention, the prediction success rate in the *Middle* and *Near* regions, where detection is crucial, appears to be around 80%, the minimal impact of the randomness in the particle filter suggests the potential usefulness of the proposed model, as presented in Table VI.

Table VII delineates the association between the quantity of microphones and the execution time of the SRP-PHAT algorithm. Notably, the overall execution time was influenced by the requisite number of microphones for SRP-PHAT algorithmic implementation. The investigations in [10] utilized data emanating from 56 microphones, albeit without considering practical constraints. It is foreseeable that a reduced number of microphones might be employed in practical implementations for commercial autonomous vehicles. The findings presented in Table VII underscore that the execution time of the SRP-PHAT algorithms exhibited only marginal sensitivity to variations in the number of microphones. The cosine similarity is computed between the output power vectors $\bar{Y}_t$, as elucidated in Section IV-C, using the following formulation:

$$Similarity(m) = \frac{\bar{Y}_t^{\,56} \cdot \bar{Y}_t^{\,m}}{|\bar{Y}_t^{\,56}||\bar{Y}_t^{\,m}|}, \tag{11}$$

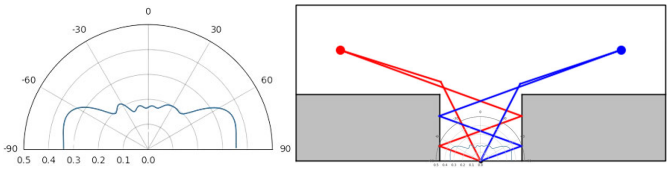
where the $\bar{Y}_t^{\,m}$ represents the output power vectors obtained from the m microphones. This suggests that a significantly reduced number of microphones (approximately 10) could be employed for real-time execution without compromising performance. Additionally, as depicted in Table V, a correlation is evident between the number of particles and the execution time of the proposed model. The results indicate that an escalation in the number of particles leads to a corresponding increase in the execution time of the proposed model.

In conclusion, the obtained results highlight a clear trade-off relationship between model accuracy and execution time. Employing more microphones to capture diverse sound data enables the acquisition of an accurate DoA response map. Combining this map with a higher number of particles enhances the precision of candidate location estimation. However, this increased computational load results in a longer execution time. Therefore, when implementing the model in real-world scenarios, careful consideration of hardware capabilities and hyperparameter settings is essential to strike a balance in this trade-off.

A notable observation concerns the variability in the accuracy of predicting NLoS vehicle positions during different phases of the scenario sequence, illustrated in the right bottom graphs of Figure 9. This fluctuation is further elucidated in

[1]This Is the Execution Time Excluding the Execution Time of the SRP-PHAT algorithm. The Entire Execution Time Can Be Obtained by Adding the Execution Time in Table VII together.

[2]The Execution Times Were Measured With an Intel i3-9100 CPU 3.6 GHz.



(a) (b)

Fig. 10. **DoA response map when NLoS vehicle is in *Far* region, and possible estimated position with sound paths that can be generated through it in *Site B* scenario.** (a) In the case of *Site B* scenario, based on the analysis of the sound path, areas outside the set RoI are exposed to sounds detected through double or more reflection paths. In such instances, the multi reflections become predominant paths for the detected sounds in the data. The resulting azimuth map exhibits symmetrical similarities, as these multiple reflections create left-right symmetrical patterns. (b) Consequently, this symmetry allows for the analysis of paths on both the left and right sides at similar distances, causing the predicted locations of NLoS vehicles by the proposed model to fluctuate horizontally.

Table VI, which details the prediction success rates across regions in scenarios involving *Site A* and *Site B*. In both scenarios, the success rate for predicting positions in the *Far* region is notably lower compared to the *Middle* and *Near* regions, with *Site B* scenario exhibiting a particularly significant decline in performance.

Upon closer analysis of the performance decline, in the *Far* region, the NLoS vehicle's front extends beyond the edge of the RoI. In the proposed method, the RoI is defined assuming the NLoS vehicle's position as a single point, yet in reality, vehicles have volume, causing sound to emanate from areas that extend beyond the RoI. For *Site B* scenario in the *Far* region, the sound originating from the NLoS vehicle spans from 14.5 m to 9.5 m. This aligns with the quantitative results in Figure 9 for *Site B* scenario, confirming the boundaries of our RoI set at ± 9.5 m. It is evident that when more than half of the NLoS vehicle exists within the RoI boundary, the model's estimated position remains stable. However, fluctuations in results are observed outside this boundary.

The observed phenomenon stems from the proposed method's inability to detect sound paths originating from specific areas. The method addresses direct, single reflection, double reflection, and diffraction paths. Due to the absence of an opposite-side wall as a spatial feature, reflections occur solely on walls adjacent to the ego-vehicle's position.

Particularly in scenarios where the NLoS vehicle is distant and the ego-vehicle is near a corner, sound paths exhibit steep angles of incidence from the sound source to these nearby walls. Consequently, the response map shows significant values beyond $+60°$ and below $-60°$, as depicted in Figure 10-(a). These angles predominantly capture sound reflections beyond the second order, which are not accounted for in our current model limited to second-order reflections. This limitation introduces inaccuracies in performance. Moreover, within the range of $[-70°, +70°]$, the DoA response map displays almost symmetrical patterns. This symmetry primarily facilitates similar-distance path analysis from both left and right sides, leading to fluctuations in NLoS vehicle position estimation and consequently compromising performance, as illustrated in Figure 10-(b).

As indicated in Table VI, instances of prediction failure were also observed in the *Middle* and *Near* regions. This can be attributed to the granularity and variable settings for the weight parameter $\alpha(\ell)$. The proposed method categorizes regions into *Far*, *Middle*, and *Near* based on sound intensity, determining feasible sound paths in each region and assigning relative weights through $\alpha(\ell)$. Consequently, the model may assign weights to paths that are impractical for sound acquisition from specific reflection paths, thereby reducing the accuracy of position predictions.

To address this issue, the model ultimately needs to be designed to consider sound paths beyond the 3rd reflection, refine the possible range of vehicle presence based on the intensity of the input sound, and adaptively apply $\alpha(\ell)$ for these positions. These improvements could enhance performance by allowing the model to more finely categorize the feasibility of sound paths and adjust the relative weights accordingly.

B. Results on OVAD Dataset

To assess the broad applicability of the proposed method, we employed the outdoor OVAD dataset obtained by Schulz et al. [10]. In this dataset, besides the original authors' classification task regarding the direction of NLoS vehicles, we conducted experiments for the tracking task using proposed method. Subsequent sections provide detailed explanations of the results for each task.

1) Classification Tasks: The particle filter method yields precise vehicle positions, tailored for tracking tasks rather than the classification of approaching vehicle directions. Initially, the exact directional position of the approaching vehicle was predicted. Subsequently, the directions $left, front, right$ were classified based on spatial information, distinguishing between NLoS on the left, LoS in the front, and NLoS on the right.

Two types of baselines, Baseline-Seen (B-S), and Baseline-Unseen (B-US), were compared. B-S utilized training and testing sets consisting of data from all possible environments, conducting classification in areas where it had already experienced sound patterns of a vehicle approaching in the NLoS region at least once. On the other hand, B-US used training and test sets that included different types of environments. For instance, when classifying the SA1 environment, B-US was trained using the dataset from $\{SB1, SB2, SB3\}$. This means that B-US conducted classification in areas where it had not been previously analyzed. The metrics used for evaluation included the Jaccard index for each class ($J^c = TP^c/(TP^c + FP^c + FN^c)$) and the total accuracy ($\sum TP^c/N$), where N represents the total number of data items, and TP, FP, and FN denote True Positive, False Positive, and False Negative, respectively. The weight parameters $\alpha(\ell) \in \{\alpha_{\text{dir}}, \alpha_{1^{\text{st}}\text{-ref}}, \alpha_{2^{\text{nd}}\text{-ref}}, \alpha_{\text{diff}}\} = \{1, 0.5, 0.5, 0.3\}$ and the random resampling noise parameter $\epsilon = 0.1$ were determined through a grid search. To ensure a fair comparison, 1-second-long segments of sound data just before appearing in the LoS were used, consistent with previous baseline experiments [10].

The proposed ASPLE method exhibited similar performance to the B-S model in environments $SA2$, $SB1$, and $SB2$.

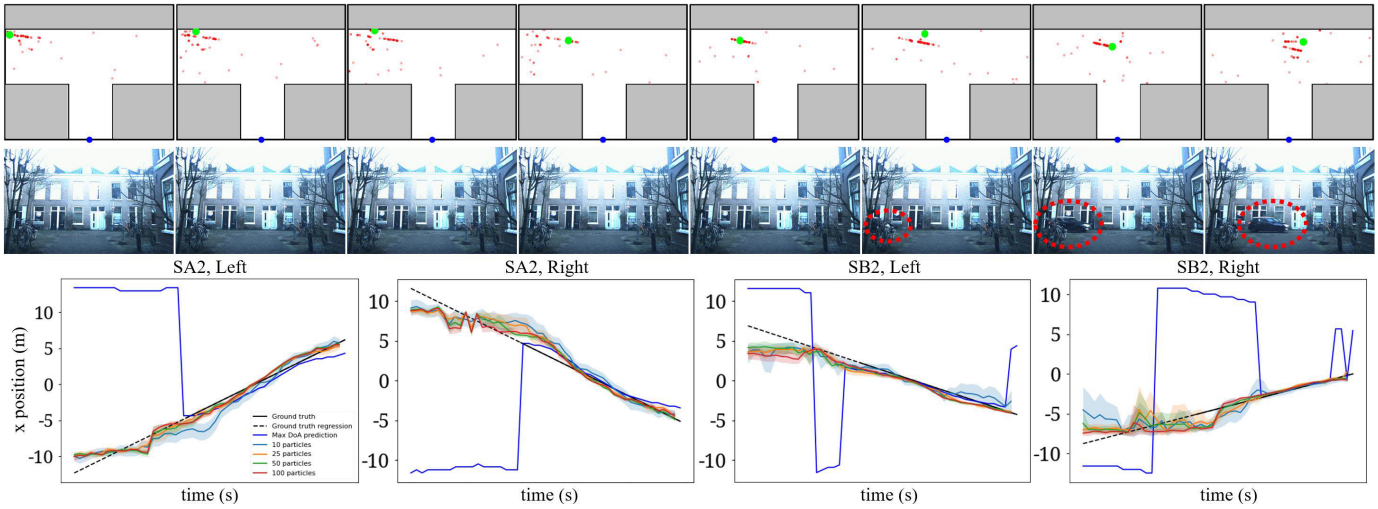


Fig. 11. **Qualitative and Quantitative results in the OVAD dataset.** In the upper row, the qualitative findings depict the scenario of an NLoS vehicle approaching from the left in *Site A* in the OVAD dataset. The particles, illustrated in red dots, actively converge from the left NLoS region and transition towards the right. The particles demonstrate proactive convergence from the left NLoS region, shifting towards the right. The blue point denotes the position of the ego-vehicle. Quantitative results in the the OVAD dataset is at the lower row. Starting from the left graph, the results correspond to NLoS vehicles coming from the left and right at *Site A*, and NLoS vehicles coming from the left and right at *Site B*, respectively.

TABLE VIII
CLASSIFICATION TASK RESULTS IN OVAD DATASET

Scnenario	Model	J^{left}	J^{front}	J^{right}	Accuracy
SA1	ASPLE	0.48	0.97	0.00	0.73
	B-US	0.00	0.63	0.00	0.50
SA2	ASPLE	0.92	0.87	0.68	0.92
	B-US	0.96	0.85	0.79	0.93
SB1	ASPLE	0.81	1.00	0.83	0.95
	B-US	1.00	0.95	0.92	0.98
SB2	ASPLE	0.69	0.88	0.83	0.90
	B-US	0.53	0.61	0.87	0.78
SB3	ASPLE	0.29	0.96	0.13	0.68
	B-US	0.76	0.73	0.73	0.86
B-S		0.89	0.93	0.91	0.95

TABLE IX
OVERALL TASK RESULT AND COMPARISON IN OVAD DATASET*

Method	DFC† [m]	J^{left}	J^{front}	J^{right}	J^{none}	Acc. [%]	ETTC‡ Avg. [s]
Hao *et al.* [11]	-	82.55	94.33	87.50	89.33	94.45	N/A
Schulz *et al.* [10]	-	71.25	90.82	78.08	83.90	89.63	N/A
Proposed method	$7 \backsim 9$	85.71	0	79.35	-	87.50	2.019
	$3 \backsim 7$	92.43	66.67	91.77	-	95.67	1.672
	$0 \backsim 3$	0	95.49	0	-	95.49	0.335
	Total	86.72	88.51	84.36	-	92.76	-

† Distance from the Center of T-Junction.
‡ Estimated Time to Collision.
* The results of baselines are referenced from Hao *et al.* [11], Table IV.

The results for the two baseline models and proposed model are presented in Table VIII, specifying the performance in different environments. Please note that the reason why J^{right} is zero at scenario SA1 is because during the data sequencing process, instances where the NLoS vehicle was on the right side were removed. It outperformed the B-US model in several unseen environments, demonstrating its adaptability without requiring a training process.while B-S outperformed the proposed ASPLE method, the latter still achieved reasonable results in the $SA2$, $SB1$, and $SB2$ environments. This is because the proposed method utilizes the reflection of sound paths rather than the spectrum of sound data.

2) Tracking Tasks: Tracking experiments were conducted in two environments as same as ARIL dataset. The particle filter loop was executed every 0.1 seconds, and the weight parameter $\alpha(\ell)$ and the random resampling noise parameter ϵ were set identical to those in the classification experiment. The predicted position of the obstacle in the NLoS region corresponds to the position of the particle with the highest weight value. At the upper row in Figure 11, the final particle filter results and synchronized camera images are displayed when a vehicle approached from the left NLoS region of the ego-vehicle. Notably, the red particles converge to the location of the vehicle even in cases where it was not in LoS region.

Lower row in Figure 11 presents quantitative results, showcasing mean values and 95% confidence intervals obtained from 20 repetitions. These results highlight the effectiveness of proposed method in accurately tracking the position of the approaching NLoS vehicle. Notably, the method demonstrates reasonable performance even when considering the variance, particularly in scenarios with a small number of particles.

As depicted in Table IX, quantitative result of experiments using sample data from the OVAD dataset were conducted, employing models proposed by Hao et al. [11] and Schulz et al. [10] as baselines. The target of proposed model was to develop a position estimation method of NLoS vehicle, therefore, experiments were not conducted on the *None* class data, which lacks NLoS vehicle instances. Additionally, the baselines primarily aim to classify the motion states, such as approaching and leaving, of NLoS vehicles using 1-second-long sound sequences. In contrast, the proposed method independently analyzes sound information acquired in each frame to estimate the position. Consequently, the classification task is designed to determine whether the particle

position in each frame is in the left or right NLoS region, and LoS region.

Unlike other methods, the proposed method is capable of not only classification but also localization of NLoS vehicles. However, there is no ground truth for the location of NLoS vehicles in OVAD dataset. Therefore, the Estimated Time to Collision (ETTC), as the evaluation metric, was first defined as follows:

$$ETTC = \frac{x^{E}_{vehicle} - x_{center}}{v_{vehicle}}, \tag{12}$$

where $x^{E}_{vehicle}$ is estimated location of NLoS vehicle, x_{center} is the center of T-Junction, and $v_{vehicle}$ is the velocity of NLoS vehicle. In OVAD dataset, since the exact $v_{vehicle}$ cannot be determined, it was estimated to be approximately 15 km/h based on the vehicle's movement in the LoS region from the front view video. Additionally, as a pseudo ground truth for the location of the NLoS vehicle, Distance from the Center of the T-Junction (DFC) was determined by considering the initial appearance time of the vehicle and $v_{vehicle}$. Subsequently, the time period corresponding to each DFC segment was determined, and experiments were performed on all audio frames within that time period.

In Table IX, a noteworthy point is the ETTC results. The ETTC was derived from $x^{E}_{vehicle}$, which was predicted from the sound frames of each segment. The model's performance in each segment was analyzed by averaging the ETTC values. By multiplying the average ETTC by $v_{vehicle}$, it can be confirmed that the results fall within the pseudo ground truth DFC segment. This indicates that the model successfully predicts the location of the NLoS vehicle. Regarding the Jaccard index, for the 7 to 9 m segment, the NLoS region results in J^{front} being zero. The 3 to 7 m segment contains both NLoS and LoS regions, allowing the calculation of the Jaccard index for all classes. For the 0 to 3 m segment, which is the LoS region, J^{left} and J^{right} are zero.

Comparing with total accuracy, while exhibiting higher performance compared to paper Schulz et al. [10], it was observed that there was a slight decline in performance relative to paper Hao et al. [11]. The decrease in performance compared to the State-of-The-Art (SOTA) method can be attributed to differences in the target task and approach. The primary objective of the SOTA method is to classify the motion states. To achieve this, they utilized neural networks to learn the characteristics of these sound sequences, including changes in sound intensity associated with the proximity of NLoS vehicles. In contrast, the proposed method independently analyzes sound information acquired in each frame to estimate the position, thus failing to account for sequential changes in sound intensity over time, which likely results in slightly inferior performance.

Furthermore, the proposed model needs to estimate sound reflection paths based on spatial information and determine RoI. However, due to the OVAD dataset being acquired for classification task, ground truth information is lacking, and the spatial information, obtained arbitrarily from an open map, is incomplete. Therefore, proposed model not only inherently suffers from reduced accuracy when classifying sounds of NLoS vehicles located outside the RoI, but also estimate the position of NLoS vehicles through inaccurate sound reflection paths, leading to a decline in performance. Additionally, when an NLoS vehicle is positioned at the boundary between NLoS and LoS regions, the sound can be correctly classified as either left or right and front. However, due to the necessity of direct comparison with the ground truth, this scenario negatively impacts performance.

TABLE X
THE TRACKING TASK RESULTS IN OVAD DATASET [10]

Metric [m]	Number of Particles				Max DoA
	10	25	50	100	
RMSE	4.49	4.19	4.10	4.07	14.168
Variance	15.23	6.45	4.22	2.87	0
Execution time [ms]	34.69	89.55	182.69	368.22	0.001

The prediction results were subsequently compared with the ground truth regression results. Given the characteristics of the T-junction, only the x-coordinate prediction results corresponding to the left and right directions in the classification experiment were compared. Similar to the experiments conducted in the ARIL dataset, the maximum DoA localization method was used as the baseline for tracking performance evaluation. Metrics such as RMSE, variance, and execution time were utilized for quantitative comparison. The results in Table X indicate that proposed method consistently outperformed the Max DoA method across different particle quantities. It's worth noting that while the RMSE exhibited a slight decrease with a larger number of particles, the execution time showed a significant decrease. This suggests that only a few particles are required for efficient NLoS detection within a 100 ms timeframe, making it suitable for real-time applications.

VII. DISCUSSION

A. Decision of RoI

Sound exhibits a maximum detectable distance depending on the types of reflection counts and spatial characteristics, as depicted in Figure 4. In the case of *Site B* in the ARIL dataset, as exemplified in Figure 4, when the distance between the NLoS vehicle's position and the center of the intersection exceeds 9.5 m, sound cannot be detected through direct, single reflection, or double reflection. Instead, it can only be detected through triple or more reflections and diffraction at the position of the microphone. Conversely, in the case of *Site A* in the ARIL dataset, illustrated in Figure 4, due to the presence of the opposite-side wall, sounds originating from a distance of over 9.5 m can be detected not only through diffraction but also through double reflection. Consequently, for *Site A*, the RoI was set from the left 12.5 m to the right 12.5 m. For *Site B*, the RoI was set from the left 9.5 m to the right 9.5 m.

Certainly, as the target sound paths of ASPLE encompass direct, single reflection, double reflection, and diffraction, the distance of 12.5 m, where sound can be detected through diffraction, can also be considered as an RoI. However, based on experimental results, the pseudo-likelihood forms peaks at angles lower than -60° and higher than +60° when an NLoS

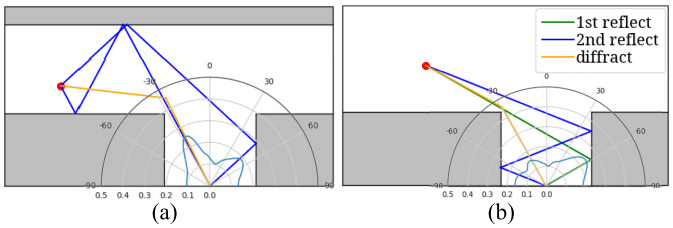


Fig. 12. **Predominant sound paths in T-Junction environments.** (a) Sound reflection is primarily influenced by the opposite-side wall as the vehicle approaches in *Site A*. Conversely, (b) the side wall in the direction of the approaching vehicle plays a predominant role in sound reflection in *Site B*.

vehicle is present at that location. This suggests that there are more sounds detected through triple or more reflections or sounds detected through reflection after diffraction than sounds detected through diffraction alone. Therefore, it falls beyond the scope of ASPLE. Accordingly, when validating ASPLE in *Site B*, the RoI was limited to the left and right 9.5 m, and the performance was assessed within this constrained area. Ultimately, when utilizing this model, the distance at which model performance is ensured may vary depending on the spatial configuration and the sound paths.

B. Predominant Sound Paths According to the Environment

Fascinating insights have emerged regarding the predominant sound paths when utilizing ASPLE in T-Junction environments with and without walls. Figure 12 illustrates scenarios where a vehicle approaches from behind the left corner in each spatial configuration. As depicted in Figure 12-(a), the largest pseudo-likelihood occurs at $-45°$, attributed to single reflection from the topside wall and double reflection from the left wall. This reflected sound predominates in most cases at *Site A*. Consequently, sounds are notably audible on the opposite side of the road to the approaching vehicle at the closed-walled T-junction alley.

Conversely, as depicted in Figure 12-(b), the predominant sound path is attributed to single reflection from the left wall and double reflection from the right wall in most cases at *Site B*. This leads to the sound being audible from the opposite direction compared to *Site A*. This difference arises due to the absence of a topside wall capable of reflecting sound into the alley where the ego-vehicle is situated.

This phenomenon was verified through the results of the maximum DoA method presented at the quantitative results illustrated in Figure 11. At the case of *Site A* in OVAD dataset, the maximum DoA consistently pointed in the opposite direction to the actual approaching direction in the NLoS region. This occurred because the upper sidewall, creating the first reflection, facilitated secondary reflections from the opposite-side wall even when the vehicle was too distant to form a reflection solely on the opposite-side wall. In contrast, the maximum DoA varied depending on the distance in *Site B*. This discrepancy arises because only the sound reflected twice can be detected when the approaching vehicle is far away, while a more robust sound, reflected only once, becomes detectable as the vehicle comes closer.

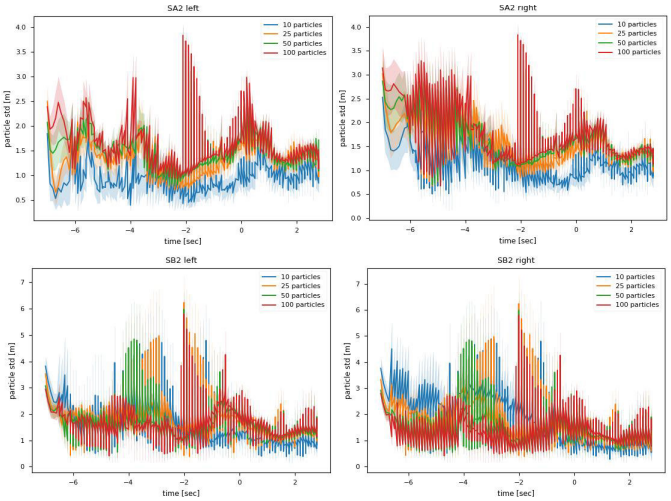


Fig. 13. **The standard deviations of the x positions of particles in the SA2 (upper row) and SB2 (bottom row) environments.** The overall standard deviation converged at near 2 m, which is a reasonable when considering the size of the approaching vehicle.

C. Convergence of ASPLE

An acknowledged limitation of ASPLE is that convergence is not assured, given the utilization of pseudo-likelihood, an approximation of the actual likelihood. Consequently, an empirical analysis is conducted to assess the dispersion of particles by measuring the standard deviation in their positions.

Figure 13 illustrates the standard deviations of the particles' x-coordinate positions in each environment. The zero-second point denotes the moment when the front end of the approaching vehicle is initially observed by the ego-vehicle's camera. It is observed that the overall standard deviation converges to around 2 m rather than 0 m. It's important to note that the sound source is considered as a point in the particle filter. The vehicle used in Schulz et al. [10], a Skoda Fabia 1.2 TSI (2010), was 4000 mm long and 1640 mm wide. As sound can emanate from any part of the vehicle, the particles will converge within a range approximately the size of the vehicle, rather than a single point. For a standard deviation of 2 m, about 68% of the particles converge within a 4 m range, assuming a Gaussian distribution. These results qualitatively indicate that the proposed method converges effectively.

The convergence test revealed two noteworthy phenomena. Firstly, there is an increase in standard deviation fluctuation with the number of particles, emphasizing the importance of choosing an appropriate number of particles for stable results. Secondly, a significant increase in standard deviation at the -2 second point was observed across all environment types. This corresponds to the scenario where only the sound path through diffraction is formed when the detected vehicle is far from the ego-vehicle, while sound paths generated by first-order and second-order reflections occur when the vehicle is close enough. Consequently, accuracy is not guaranteed even when the particles converge before the -2 second point.

To address this challenge, one potential solution is to incorporate higher-order reflection paths to account for more sound paths from the source to the ego-vehicle. However, considering higher-order sound paths introduces more measurement noise

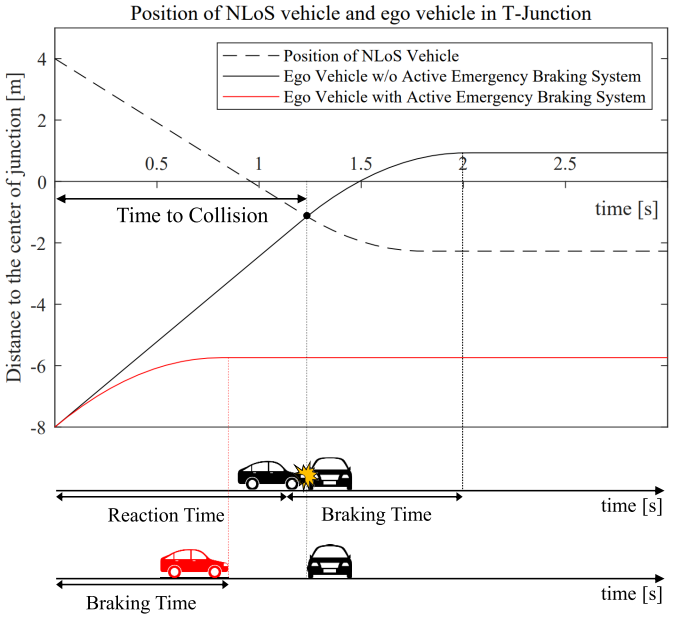


Fig. 14. **The effect of the application of the active emergency braking system on the ego vehicle's sudden braking-induced positional changes and the risk of collision with NLoS vehicles.** When the drivers of the NLoS vehicle and the ego vehicle first perceive each other, the Time to Collision (TTC) between the two vehicles is 1.3 seconds. In the scenario where the active emergency braking system is not applied, assuming the ego vehicle's driver is a keen and alert driver with a reaction time of 1.2 seconds, the braking time of 0.81 seconds results in a total stopping time of 2 seconds, making collision avoidance impossible. However, with the application of the active emergency braking system, the reaction time is eliminated, reducing the stopping time to less than TTC, thereby enabling collision avoidance.

due to sound diffusion during reflection, and it also increases the computational cost of ASPLE calculations. Therefore, predicting invisible objects faster than two seconds remains a potential avenue for future research.

D. Potential Applications in Intelligent Transportation Systems

Through the sound-based NLoS vehicle position prediction method proposed in this study, it is feasible to establish a sound-based active emergency braking system capable of preventing accidents caused by the sudden appearance of NLoS vehicles, as outlined in Table II. When applying the proposed method to the Target Scenario assumed in Section III, an NLoS vehicle approaching at 15 km/h and an ego vehicle approaching at 20 km/h will each detect the other when they are 4 m and 8 m from the center of the intersection, respectively. At this moment, the Time to Collision (TTC) is approximately 1.3 seconds. Even a keen and alert driver has a reaction time of 1.0 to 1.2 seconds [42], [43], [44], [45], [46], during which both vehicles will have already moved 5 m and 6.67 m, respectively. Even if both vehicles perform full braking after this, their respective braking distances are 1.27 m and 2.26 m, making collision avoidance impossible. However, if an active emergency braking system is developed using the proposed method, the driver's reaction time would effectively approach 0 second, resulting in the ego vehicle moving only 2.26 m during the 0.81 seconds of braking time. This allows efficient accident prevention through a stopping time shorter than the TTC, as shown in Figure 14.

Additionally, the proposed method can definitively estimate the position of the NLoS vehicle as being 10 m from the center of the intersection at *Site A* and 8 m at *Site B*. In this case, the TTC with the NLoS vehicle approaching at 15 km/h is approximately 3 seconds at *Site A* and 2.3 seconds at *Site B*. During this time, it is possible to determine whether the ego vehicle can make a right turn or pass through, thus minimizing traffic flow disruption due to stopping.

However, due to variations in ego-vehicle's physical attributes, road conditions, and spatial configurations, the stopping distance may differ. Therefore, to build a more reliable system, it is crucial to ascertain the essential braking distance according to the ego-vehicle's driving situation and the correlation between the initial detection position of the NLoS vehicle ensuring safety. Additionally, investigating the correlation between the model's accuracy in predicting the NLoS vehicle's position and the corresponding safe initial detection timing is also essential.

The proposed model utilized spatial information extracted from open maps or acquired directly. In real-world driving environments, direct acquisition of spatial information in all driving conditions is impractical, and receiving such information via open maps can be unreliable due to communication network issues or failures. From a perception perspective of autonomous vehicle, integrating spatial information recognition technologies enables the development of more robust autonomous driving systems. Therefore, integrating ego-vehicle's own spatial information inference technologies based on perception systems such as cameras, radars, and LiDARs currently used in vehicles contributes to the development of more reliable autonomous driving systems. This integration will help mitigating the challenges associated with unreliable or unavailable spatial information from external sources, thereby enhancing overall system stability and performance.

VIII. Conclusion and Future Works

The proposed work introduces a novel approach named ASPLE for detecting vehicles approaching from NLoS regions. The method has been successfully validated using the proposed ARIL dataset, which focuses on acoustic recognition based invisible target localization within a built testbed. ASPLE not only demonstrated comparable detection performance to conventional classification methods, it also outperformed the conventional maximum DoA localization method through the SRP-PHAT algorithm in vehicle tracking tasks.

The study also delves into the reflection and diffraction patterns of sound based on surrounding spatial conditions in the context of the proposed approach. The anticipated impact of this method is a significant reduction in collision risks with vehicles emerging from NLoS regions by localizing them through the integration of sound and spatial information. Also, ASPLE can be the corner stone of sound based Active Blind Spot Assistant with active emergency braking systems. To address this, which will be the focus of future work,

analysis and model refinement based on sounds acquired in the real-world environment of an ego-vehicle driving on an actual road will be necessary.

For future work, addressing the trade-off between time and performance identified in this study is crucial. The proposed method aims to consider sound from all directions, calculating DoA every 1° within the azimuth range of $[-90°, +90°]$ based on the vehicle's heading direction. However, certain angle ranges may not need to be considered depending on the type of detected sound path. To reduce computational costs, future research can explore adaptive sound detection ranges based on the ego-vehicle's position and surrounding spatial characteristics. This adaptive approach would tailor the range of angles considered for sound detection, optimizing computational resources by excluding unnecessary calculations depending on the type of sound path being detected. Furthermore, the execution time increased with an increase in the number of particles and the number of microphones. Therefore, by adapting the model that uses spectrograms for classifying the presence and direction of NLoS vehicles, we can integrate it to sense the direction of NLoS vehicles. Subsequently, applying particles adaptively to those specific areas can optimize the performance and enable real-time processing when using an optimized number and placement of microphones.

Using a multimodal approach with other sensors to enhance performance is also a future direction. The proposed model is based on sound, which can diminish sharply in intensity as distance increases, potentially leading to undetected signals. Therefore, integrating radar, known for its directivity and reflective properties, with the proposed method could facilitate research into multimodal localization methods for NLoS objects, leveraging radar actively utilized in current driving assistant systems.

ACKNOWLEDGMENT

The authors would like to thank anonymous reviewers for their valuable suggestions and comments, and architect Yujeong Soh for helping with the construction of test bed. The research facilities for this work was provided by the Institute of Engineering Research, Seoul National University.

REFERENCES

[1] S. Campbell et al., "Sensor technology in autonomous vehicles: A review," in *Proc. 29th Irish Signals Syst. Conf. (ISSC)*, Jun. 2018, pp. 1–4.

[2] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3D object detection methods for autonomous driving applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3782–3795, Oct. 2019.

[3] J. Fayyad, M. A. Jaradat, D. Gruyer, and H. Najjaran, "Deep learning sensor fusion for autonomous vehicle perception and localization: A review," *Sensors*, vol. 20, no. 15, p. 4220, Jul. 2020.

[4] S.-W. Kim et al., "Multivehicle cooperative driving using cooperative perception: Design and experimental validation," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 663–680, Apr. 2015.

[5] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina, "V2X access technologies: Regulation, research, and remaining challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1858–1877, 3rd Quart., 2018.

[6] 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. 2068–2077.

[7] 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.

[8] S.-W. Kim, W. Liu, M. H. Ang, E. Frazzoli, and D. Rus, "The impact of cooperative perception on decision making and planning of autonomous vehicles," *IEEE Intell. Transp. Syst. Mag.*, vol. 7, no. 3, pp. 39–50, Fall 2015.

[9] H. Choi, H. Yang, S. Lee, and W. Seong, "Classification of inter-floor noise type/position via convolutional neural network-based supervised learning," *Appl. Sci.*, vol. 9, no. 18, p. 3735, Sep. 2019.

[10] 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.

[11] M. Hao et al., "Acoustic non-line-of-sight vehicle approaching and leaving detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 8, pp. 9979–9991, Aug. 2024.

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

[13] J. H. DiBiase, *A High-accuracy, Low-latency Technique for Talker Localization Reverberant Environments Using Microphone Arrays*. Providence, RI, USA: Brown Univ., 2000.

[14] J. P. Dmochowski, J. Benesty, and S. Affes, "A generalized steered response power method for computationally viable source localization," *IEEE Trans. Audio, Speech Language Process.*, vol. 15, no. 8, pp. 2510–2526, Nov. 2007.

[15] B. Kwon, Y. Park, and Y.-s. Park, "Analysis of the GCC-PHAT technique for multiple sources," in *Proc. ICCAS*, Oct. 2010, pp. 2070–2073.

[16] D. B. Ward and R. C. Williamson, "Particle filter beamforming for acoustic source localization in a reverberant environment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, May 2002, pp. II-1777–II-1780.

[17] A. K. Aggarwal, "On the use of artificial intelligence techniques in transportation systems," *Int. J. Soft. Comput. Eng.*, vol. 5, no. 5, pp. 21–24, 2015.

[18] S. Unar, Y. Su, X. Zhao, P. Liu, Y. Wang, and X. Fu, "Towards applying image retrieval approach for finding semantic locations in autonomous vehicles," *Multimedia Tools Appl.*, vol. 83, no. 7, pp. 20537–20558, Aug. 2023.

[19] K. Arora and A. Kumar, "A comparative study on content based image retrieval methods," *Int. J. Latest Technol. Eng., Manage. Appl. Sci.*, vol. 6, no. 4, pp. 77–80, 2017.

[20] A. Aggarwal, "A hybrid approach to GPS improvement in urban canyons," *Int. J. Eng. Sci. Res. Technol.*, vol. 4, no. 10, pp. 358–363, 2023.

[21] J. Van Brummelen, M. O'Brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transp. Res. C, Emerg. Technol.*, vol. 89, pp. 384–406, Apr. 2018.

[22] F. Rosique, P. J. Navarro, C. Fernández, and A. Padilla, "A systematic review of perception system and simulators for autonomous vehicles research," *Sensors*, vol. 19, no. 3, p. 648, 2019.

[23] H. Yasuda and Y. Ohama, "Toward a practical wall see-through system for drivers: How simple can it be?" in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, Nov. 2012, pp. 333–334.

[24] J. A. Simmons and R. A. Stein, "Acoustic imaging in bat sonar: Echolocation signals and the evolution of echolocation," *J. Comparative Physiol. A*, vol. 135, no. 1, pp. 61–84, 1980.

[25] K. E. Railey, "Demonstration of passive acoustic detection and tracking of unmanned underwater vehicles," Ph.D. dissertation, Dept. Mech. Eng., Joint Program Oceanogr./Appl. Ocean Sci. Eng., Woods Hole Oceanograph. Inst., Massachusetts Inst. Technol., Cambridge, MA, USA, 2018.

[26] 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.

[27] J. E. Peabody, G. L. Charvat, J. Goodwin, and M. Tobias, "Through-wall imaging radar," *Lincoln Lab. J.*, vol. 19, no. 1, pp. 62–72, 2012.

[28] F. Adib and D. Katabi, "See through walls with WiFi!" in *Proc. ACM SIGCOMM Conf.*, 2013, pp. 75–86.

[29] A. Palffy, J. F. P. Kooij, and D. M. Gavrila, "Occlusion aware sensor fusion for early crossing pedestrian detection," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2019, pp. 1768–1774.

[30] S. Lee, Y. Jung, Y.-H. Park, and S.-W. Kim, "Design of V2X-based vehicular contents centric networks for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13526–13537, Aug. 2022.
[31] B. Rebsamen et al., "Utilizing the infrastructure to assist autonomous vehicles in a mobility on demand context," in *Proc. TENCON IEEE Region Conf.*, Nov. 2012, pp. 1–5.
[32] 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.
[33] C. Chen et al., "SoundSpaces: Audio-visual navigation in 3D environments," in *Proc. 16th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 17–36.
[34] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 9701–9707.
[35] M. Mizumachi, A. Kaminuma, N. Ono, and S. Ando, "Robust sensing of approaching vehicles relying on acoustic cues," *Sensors*, vol. 14, no. 6, pp. 9546–9561, May 2014.
[36] 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.
[37] I. An, M. Son, D. Manocha, and S.-E. Yoon, "Reflection-aware sound source localization," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 66–73.
[38] 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.
[39] R. Schubert, E. Richter, and G. Wanielik, "Comparison and evaluation of advanced motion models for vehicle tracking," in *Proc. 11th Int. Conf. Inf. Fusion*, Cologne, Germany, Jul. 2008, pp. 1–6.
[40] 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)*, Calgary, AB, Canada, Apr. 2018, pp. 351–355.
[41] U. Sandberg, L. Goubert, and P. Mioduszewski, "Are vehicles driven in electric mode so quiet that they need acoustic warning signals," in *Proc. 20th Int. Congr. Acoust.*, 2010.
[42] K. G. Hooper and H. W. McGee, "Driver perception-reaction time: Are revisions to current specification values in order?" Transp. Res. Board, Washington, DC, USA, Tech. Rep. HS-036 165, 1983.
[43] P. L. Olson, "Driver perception response time," *SAE Trans.*, vol. 98, pp. 851–861, Jan. 1989.
[44] N. H. Gartner, C. J. Messer, and A. Rathi, "Traffic flow theory-a state-of-the-art report: Revised monograph on traffic flow theory," Transp. Res. Board, Washington, DC, USA, Special Rep. 165, 2002.
[45] (2018). *Speed-Measuring Device Operator Training. Core Module. Participant Manual*. NHTSA. [Online]. Available: https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/core_participant_manual-smd-2018.pdf
[46] K. Čulík, A. Kalašová, and V. Štefancová, "Evaluation of driver's reaction time measured in driving simulator," *Sensors*, vol. 22, no. 9, p. 3542, May 2022.

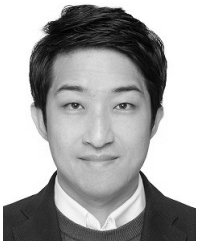
Mingu Jeon received the B.S. degree in electrical engineering from Pohang University of Science and Technology, Pohang, South Korea, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University. His current research interests include computer vision, machine learning, machine vision, smart manufacturing, and autonomous driving.

Jae-Kyung Cho received the B.S. degree in mechanical engineering and the master's degree in electrical and computer engineering from Seoul National University, in 2020 and 2023, respectively. He is currently with SK Telecom, as a LLM Researcher. His current research interests include LLM, RLHF, and reinforcement learning.

Hee-Yeun Kim received the B.S. degree in mechanical engineering from Korea University, Seoul, South Korea, in 2024. She is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Seoul National University. Her research interests include robot learning and autonomous driving.

Byeonggyu Park received the B.S. degree in software engineering from Kookmin University, Seoul, South Korea, in 2024. He is an Intern with Seoul National University. His current research interests include computer vision, multi modal, machine learning, robotics, and autonomous driving.

Seung-Woo Seo (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, South Korea, in 1987 and 1989, respectively, and the Ph.D. degree in electrical and computer engineering from The Pennsylvania State University, University Park, PA, USA, in 1993. He was a Faculty Member with the Department of Computer Science and Engineering, The Pennsylvania State University. He was a member of the Research Staff with the Department of Electrical Engineering, Princeton University, Princeton, NJ, USA. In 1996, he joined the School of Electrical Engineering, Institute of New Media and Communications, and the Automation and Systems Research Institute, Seoul National University, as a Faculty Member. He is currently a Professor of electrical engineering with Seoul National University and the Director of the Intelligent Vehicle IT (IVIT) Research Center funded by the Korean Government and automotive industries.



Seong-Woo Kim (Member, IEEE) received the B.S. and M.S. degrees in electronics engineering from Korea University, Seoul, South Korea, in 2005 and 2007, respectively, and the Ph.D. degree in electrical engineering and computer science from Seoul National University in 2011. He was a Post-Doctoral Associate with Singapore-MIT Alliance for Research and Technology. In 2014, he joined Seoul National University, where he is currently an Associate Professor with the Graduate School of Engineering Practice. He received the Best Student Paper Award at the 11th IEEE International Symposium on Consumer Electronics; and the Outstanding Student Paper Award at the First IEEE International Conference on Wireless Communication, Vehicular Technology, Information Theory, and Aerospace and Electronic Systems Technology. He was a Guest Editor of a Special Issue on Applications and Systems for Collaborative Driving in the IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS.