SonicVision: Acoustic Object Detection for Autonomous Driving

RO57035: RO MSc Thesis Exam

Ziang Liu

TU Delft

SonicVision: Acoustic Object Detection for Autonomous Driving

RO57035: RO MSc Thesis Exam

by

Ziang Liu

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on Friday, September 26, 2025, at 11:00 AM.

Student number:	5978866
Thesis committee:	Dr. J.F.P. Kooij
	Dr. H. Caesar
	Dr.ir. R. Hendriks (EWI)
	S. Wang, MSc

TU Delft

SonicVision: Acoustic Object Detection for Autonomous Driving

Ziang Liu
ziangliu@tudelft.nl

Abstract

Autonomous driving relies heavily on cameras and LiDAR for 3D perception, yet these vision-based sensors face limitations under poor illumination, adverse weather, or occlusion. Inspired by human hearing, we explore whether microphone arrays can enhance vehicle perception. We propose SonicVision, the first bird's-eye-view (BEV) acoustic detection framework that jointly localizes and classifies traffic participants using sound alone. Our method employs a horizontally arranged 32-channel microphone array and transforms raw waveforms into short-time Fourier transform (STFT) features augmented with positional embeddings. A ResNet-based architecture is trained with novel Gaussian label representations to predict class-conditioned direction–distance distributions. To support this study, we collect three datasets (simulation, test track, and real road) with synchronized audio and LiDAR, where LiDAR detections serve as pseudo-labels. Experiments show that SonicVision significantly outperforms beamforming-based baselines, achieving accurate localization and classification performance. In some cases, our approach is able to identify objects that are missed by LiDAR, suggesting its potential as both an independent sensor and a complementary modality. These results provide the first evidence that low-cost microphone arrays can meaningfully contribute to 3D perception for autonomous vehicles.

1. Introduction

Autonomous driving systems rely heavily on accurate 3D detection of surrounding traffic participants to ensure safe navigation and decision making. Traditionally, this task has been addressed using vision-based sensors, primarily cameras and LiDAR. Cameras provide rich semantic cues, while LiDAR offers precise 3D geometry. Their combination has driven major advances in 3D detection, making them the core backbone of today's autonomous vehicles.[13]

Although vision-based perception systems have enabled significant progress in autonomous driving, they still face inherent limitations. Recent studies indicate that only about 30 % of autonomous vehicles currently deploy LiDAR[28], meaning that many systems still rely solely on cameras. In such cases, perception becomes highly sensitive to illumination changes and suffers in low-light or nighttime conditions. Even when LiDAR is available, its performance can still deteriorate under adverse weather, such as heavy fog, or suffer from issues like overheating and line dropouts.

However, for humans, these challenges are rarely problematic. We not only rely on our eyes to observe, but also on our ears to listen. For instance, when walking along a dark street at night, we can roughly estimate the position of an approaching car behind us solely from its sound, even without turning around. Inspired by this natural ability, in our study, we equip autonomous vehicles with a microphone array to explore a key question: **Could microphone arrays contribute to 3D perception for autonomous vehicles?**

1.1. Research Questions

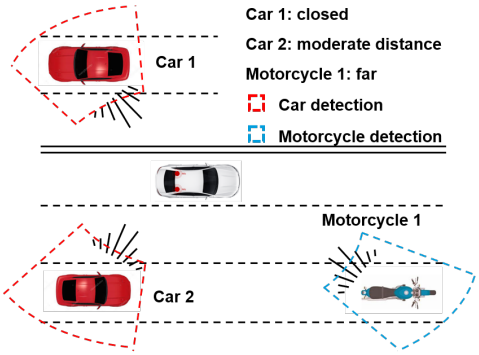


Figure 1. **Problem definition**: Acoustic perception for autonomous driving

To investigate whether sound can support 3D detection tasks, we retain the same task formulation as in a conventional detection framework, focusing on the localization and classification of surrounding objects. However, considering that sound cannot directly predict bounding boxes, as it does not provide information about object size. We restrict localization to position estimation. Accordingly, our goal is to design a neural network, as illustrated in Figure 1, that accurately localizes nearby traffic participants while simultaneously identifying their categories.

In this task, two key aspects need to be considered: the design of informative features and the choice of an appropriate neural network architecture. Leveraging our microphone array, we first seek to capture the acoustic cues present in the driving environment and investigate suitable preprocessing strategies to filter out the most useful information. The next challenge is to formulate the problem and design a neural network that can effectively transform these features into accurate detection results, jointly encoding both the position and category. Based on this reasoning, we propose the following research questions:

- What types of information are conveyed by sound, and in what ways have they been exploited in prior research on autonomous driving?
- How can auditory information be effectively encoded, and what output representations are most suitable for formulating the detection problem?
- Which neural network architectures are best adopted to address this problem and to jointly perform both localization and classification?

To address the above questions, Chapter 2 reviews related work, from which we derive key insights, refine the research questions, and identify gaps that motivate our approach. Chapter 3 details the proposed methodology, while Chapter 4 presents experiments designed to evaluate the defined tasks and answer the research questions. Finally, Chapter 5 concludes by revisiting our central inquiry: whether microphones can enhance 3D perception in autonomous driving environments.

2. Related Work

In line with our research questions, our work can be uniquely positioned within four related research directions: acoustic perception in autonomous driving, acoustic localization, sound classification and acoustic datasets in the field of robotics.

2.1. Acoustic Perception in Autonomous Driving

Sound offers unique advantages as a sensing modality. Unlike vision, its propagation is independent of light, allowing perception under poor illumination or occlusion. In the context of autonomous driving, the use of microphones can be categorized into two primary tasks: in-line-of-sight (ILOS) detection and non-line-of-sight (NLOS) object detection.

ILOS perception refers to the ability to detect objects that are geometrically within the sensor's field of view but cannot be reliably recognized due to adverse sensing conditions. For ILOS problem, Jannik et al.[46] first mounted microphones on a vehicle and, using a camera-based detector as supervision, performed beamforming-like acoustic segmentation of frontal vehicles. Later, Chakravarthula et al.[7] introduced a large 1,024-microphone array on the vehicle's fender and, by combining audio with visual data, achieved robust performance across diverse lighting and weather conditions. Similar efforts were made by Chuang[11], though their setup only placed microphones on a static pole and thus lacked dynamic driving scenarios. More recently, Valverde et al. [39] explored a horizontal microphone layout—an approach closely related to our problem—but their study suffered from test-set leakage, which undermines the reliability of their detection results.

Another interesting study comes from Dai et al[9]. Although they did not mount microphones on a vehicle, they employed four sets of human ear-shaped microphones and placed the array roadside. This setup enabled semantic segmentation and depth estimation within a 360-degree camera view of the environment, thereby demonstrating that sound carries information about both the location and the category.

In addition, many studies have focused on addressing the NLOS problem using acoustic information. For example, Schulz et al.[34] employed a microphone array mounted on the roof of a vehicle, through beamforming and SVM methods, they successfully predicted the presence of vehicles emerging from behind a narrow T-junction. Following this, Hao et al.[15] reproduced the study using neural network–based methods, further validating the effectiveness of acoustic cues for NLOS detection. However, these studies were limited to predicting only the direction of oncoming vehicles. Building on this, Jeon et al.[18] incorporated a particle filtering approach and leveraged sound reflections to accurately estimate the positions of vehicles located around street corners.

Research Gaps: Most existing studies have focused on detection from a front-view perspective, whereas autonomous driving requires the more comprehensive perspective offered by a bird's-eye view. In addition, prior work has largely addressed specific challenges, such as NLOS conditions, while little attention has been given to exploring the intrinsic potential of sound in autonomous driving — specifically, whether sound alone can serve as a reliable modality for detection and localization.

2.2. Acoustic Localization

Acoustic localization constitutes a broad application area related to sound. Active techniques, such as echolocation, operate by transmitting sound signals and analyzing the corresponding reflections for localization, navigation, or prey detection, a mechanism commonly observed in animals like bats and dolphins[19]. For instance, sonar systems perceive the surrounding environment by actively emitting and receiving sound signals[41], and are mainly applied in underwater and robotics domains[26, 37]. More recently, learning–based approaches have also been explored in active acoustic localization. For example, Brunetto et al.[6] mimicked a sonar system by mounting a loudspeaker on a robot

to emit short "beep" sounds, and employed a U-Net architecture to achieve indoor depth estimation.

2.2.1. Sound Source Direction Estimation

Besides active techniques, passive acoustic localization methods have been widely applied to the problem of locating naturally sound-emitting objects. Previously, most approaches focus primarily on estimating the direction of the sound source. Classical physical methods, such as beamforming[36] and MUSIC[33], infer the direction of arrival by exploiting the phase differences (or equivalently, time delays) of sound waves captured across microphones.

However, traditional physical methods face the following problem: without prior knowledge of the number of sound sources, it is difficult to accurately predict the directions of multiple sources.[20] Consequently, learning-based methods have emerged. Xiao et al.[42] achieved sound source direction estimation using a simple MLP network. Building on this, He et al.[16] optimized both the network architecture and the encoding method, employing a CNN-based neural network with a specially processed GCC-PHAT as input to achieve more accurate sound source detection. This approach was further applied to the real robot Pepper, enabling it to interact more effectively with users issuing voice commands. Moreover, considering the sequential nature of sound, several works[1, 17, 22] have adopted CRNN-based approaches for direction estimation and likewise achieved promising results.

2.2.2. Sound Source Distance Estimation

Evidently, knowing only the direction of a sound without its distance does not allow accurate localization of the source. However, research on distance estimation from sound remains scarce. Some studies[5, 31] combined machine learning model with traditional signal processing features, but these often fail to generalize to the new environment.

More recently, deep learning–based methods have been explored for sound distance estimation, where some approaches discretize spatial information and output a range (e.g., 5–10 m) through classification[43], while others adopt regression strategies to directly predict the distance[24]. Nevertheless, almost all studies related to distance prediction have primarily focused on indoor acoustic tasks, leaving a research gap when it comes to outdoor environments, which are more spacious, dynamic, and noise-prone[14].

Research Gaps: Research on sound source direction estimation has mainly focused on indoor environments, where conditions are simpler and noise levels are lower. Distance estimation itself remains an emerging area, and to the best of our knowledge, no prior work has addressed audio localization that jointly considers both direction and distance.

2.3. Sound Classification

In autonomous driving, a complete object detection pipeline requires both localization and classification, enabling the system to distinguish among categories such as cars, motorcycles, and buses[3]. For the classification task, vision-based approaches typically rely on cues such as object shape and appearance, whereas acoustic-based methods instead exploit frequency characteristics and sound intensity to differentiate between classes[10].

In the acoustic domain, several studies have explored traditional machine learning approaches for classification, including Support Vector Machines (SVM)[38, 40], Hidden Markov Model (HMM)[12, 44, 45] and K-Nearest Neighbor (KNN) classifiers[4, 8]. Other studies have adopted learning-based approaches. Su et al.[35] introduced CNN for sound classification using spectrogram inputs, showing that CNN outperforms traditional machine learning methods. Building on this, Salamon et al.[29] incorporated data augmentation techniques to further enhance CNN performance. In addition, to avoid information loss from preprocessing, Sang et al.[32] instead used raw waveforms with cRNN structure, achieving higher recognition accuracy.

Research Gaps: Research on sound source classification has reached a relatively mature stage, and existing methods have shown strong generalization across diverse acoustic scenes. However, only a few studies have explored its integration with localization.

2.4. Acoustic Dataset in Robotics

Autonomous driving scenario: In the field of autonomous driving, publicly available datasets for acoustic detection remain scarce. Among the existing efforts, some multimodal datasets incorporate sound, but they primarily concentrate on the camera's field of view, detecting vehicles located in front of the ego vehicle [7, 39]. Others are designed to address specific challenges such as non-line-of-sight (NLOS) scenarios, aiming to detect vehicles in regions where the direct line of sight is blocked [18, 34].

Indoor robotics research: In contrast, acoustic-related datasets are more abundant in indoor robotics research. Many studies employ microphone arrays for tasks such as estimating a user's location or orientation and enhancing human–robot interaction [16, 27]. Other datasets target specific tasks, including sound classification [30] and depth estimation from acoustic echoes [6].

Research Gaps: To date, no dataset has explored the use of microphone arrays for 3D object detection.

2.5. Research Gaps and Contributions

Through the above literature, we have learned how sound has been applied in the context of autonomous driving and how it supports localization and classification in indoor

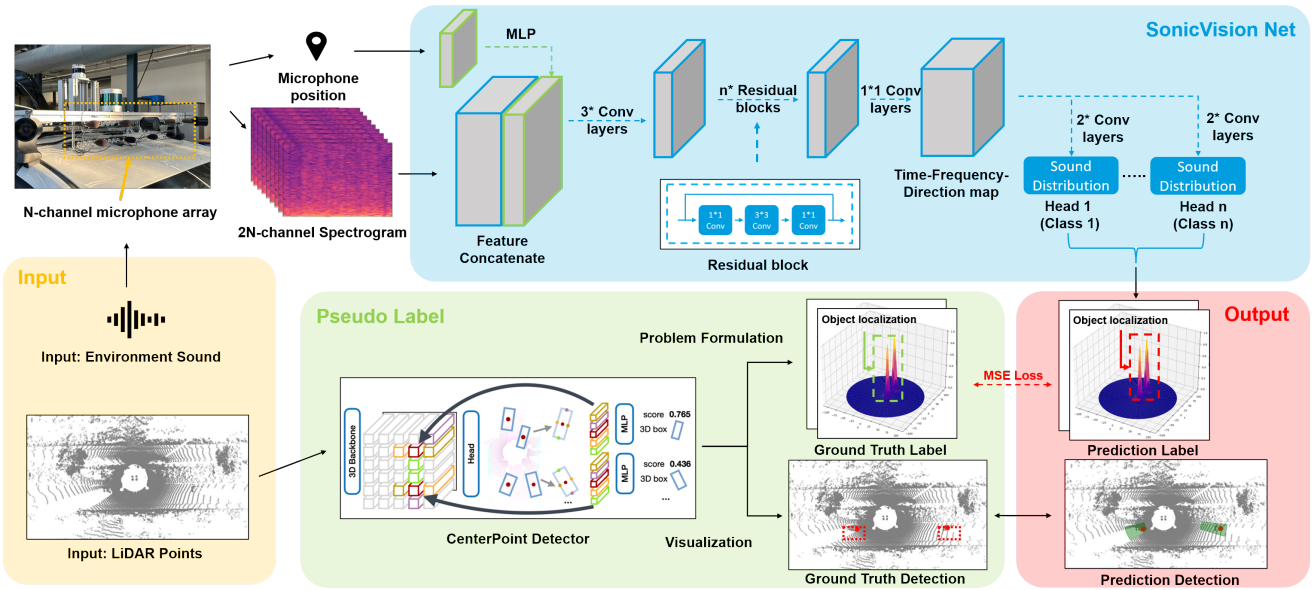


Figure 2. The framework of acoutic object detection. We collect synchronized environmental audio and LiDAR point clouds, where multi-channel microphone signals are transformed into STFT features and fused with microphone positional embeddings as input to our network. Meanwhile, CenterPoint LiDAR detections serve as pseudo labels to supervise the audio-based localization task, and consequently the model outputs class-specific distance–direction distributions in a bird's-eye view (BEV), where colored bounding boxes indicate localized sound sources of different object categories.

robotics. Nevertheless, the existing studies in the audio domain lack a unified approach that enables both localization (direction + distance) and classification of sound sources. Similarly, in the context of autonomous driving, there is a lack of research that leverages acoustic sensing from a bird's-eye-view perspective to enable vehicles to fully perceive their surrounding environment.

To address these limitations, we design a novel bird's-eye-view acoustic detection framework for vehicle localization and classification using sound. We further construct our own dataset for evaluation, demonstrating that our method not only achieves the proposed functionalities but also outperforms baseline methods. Overall, the main contributions of this study are as follows:

- To the best of our knowledge, we are the first to **address bird's-eye view localization and classification purely from audio signals**, showing that sound provides both positional and categorical information to complement other perception modalities and improve autonomous driving safety.
- We adapt a Res-net architecture and propose a **2D Gaussian labeling scheme** for bird's-eye-view sound source **localization**, as distributional labels are more suitable for sound-based perception.
- We **collect and annotate a new dataset** specifically for this task, which we expect will advance research in acoustic and multi-modal perception for autonomous driving.

3. Methodology

As outlined earlier, our research focuses on the localization and classification of traffic participants in autonomous driving environments using a microphone array. Accordingly, this chapter first introduces the microphone array employed in our study. We then describe how the recorded sounds are transformed into neural network inputs, along with the design of our tailored output representation. Finally, we present the proposed model and the datasets used for its training and evaluation. We outline the framework of our proposed approach in Figure 2.

3.1. Microphone Array Design

To enable bird's-eye view detection, we require our vehicle to perceive sounds from all 360° directions. Therefore, we designed a horizontally arranged microphone array installed around the vehicle, as shown in Figure 3. On each side, a PMMA plate was fabricated to host eight microphones, resulting in a total of 32 microphones covering four directions. Each plate features a central groove that allows precise alignment with the camera, simplifying the computation of transformation matrices between sensors for multimodal perception. The microphones are mounted around the vehicle's roof frame without obstructing LiDAR. As our focus is on bird's-eye-view detection, the array was not designed for high vertical resolution, which would require a larger aperture and risk blocking other sensors.

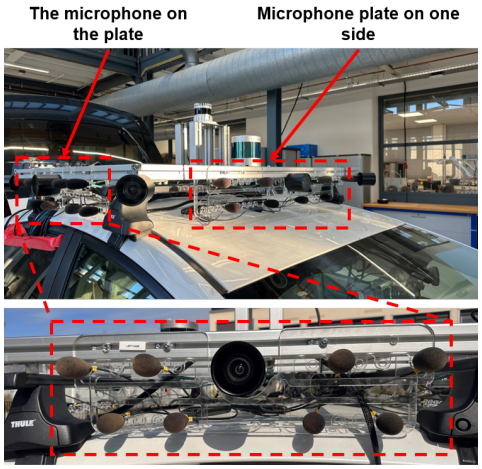


Figure 3. Microphone Design

3.2. Feature Extraction

With our designed microphone array, we are able to capture raw signals from n audio channels. These signals can be encoded in different ways depending on the application: raw audio preserves complete information but is often contaminated with noise, whereas processed representations such as the STFT reduce noise and simplify neural network learning, albeit at the cost of some information loss.

After comparison (see Appendix), we adopt STFT as our feature representation. It decomposes the signal into frequency components over short time windows, producing a time–frequency representation that captures frequency, intensity, and temporal structure—key cues for localizion and classification. The image-like nature of STFT outputs also makes them well-suited for CNN-based feature extraction.

For each channel $c = 1, 2, \ldots, n$, the short-time Fourier transform (STFT) is computed as:

$$X_c(m,k) = \sum_{t=0}^{T-1} x_c[mH + t]\, w[t]\, e^{-j\frac{2\pi k}{K}t}, \tag{1}$$

where $x_c[\cdot]$ is the discrete-time signal of channel c, $w[t]$ is the analysis window of length T, m is the frame index with hop size H, and k is the frequency bin among K DFT points. This formulation follows the standard definition of the short-time Fourier transform (STFT).[2]

In addition, to represent the complex spectrum, we use its real and imaginary parts rather than phase and magnitude, thereby avoiding the discontinuity problem at π:

$$R_c(m,k) = \mathrm{Re}\{X_c(m,k)\}, \tag{2}$$
$$I_c(m,k) = \mathrm{Im}\{X_c(m,k)\}. \tag{3}$$

3.3. Task Formulation

Unlike visual data, sound does not directly encode spatial information, which makes a pure regression formulation for network outputs particularly challenging. To overcome this, we adopt a tailored labeling strategy for localization and classification, which provides a more suitable and effective supervision signal, as described in the following.

Sound Source	Direction (°)	Distance (m)
Car1	-44	16
Car2	69	29

Table 1. Objects with annotated directional and distance labels.

3.3.1. Localization (direction + distance) label

For the sound localization problem, as mentioned earlier, our goal is to detect the exact position (with uncertainty) of an object by estimating its direction and distance. In the following, we use the example in Table 1. to illustrate our label design methodology. Our design is based on the label strategy proposed by He et al[16]. As shown in Figure 5a, a 360-dimensional direction vector is employed, in which multiple Gaussian distributions represent the angular distribution of sound sources, formally defined in (4) and (5).

$$\Delta(i,\theta_k) = \min\big(|i-\theta_k|,\ 360 - |i-\theta_k|\big), \tag{4}$$

$$y[i] = \sum_{k=1}^{K} \exp\left(-\frac{\Delta(i,\theta_k)^2}{2\sigma^2}\right), \tag{5}$$

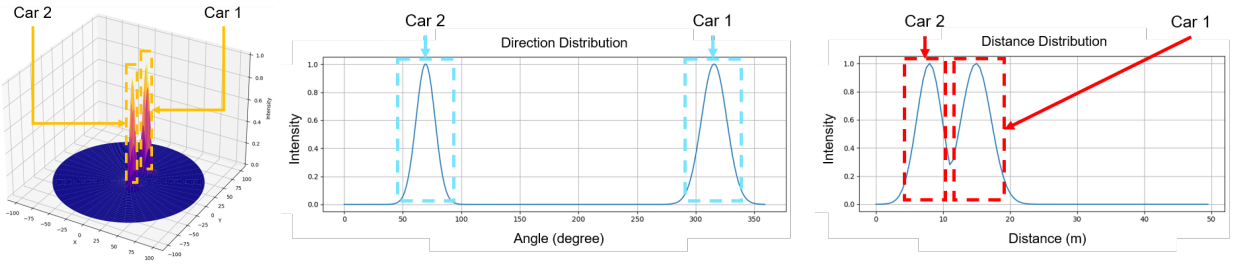
where $y \in \mathbb{R}^{360}$ is the direction label vector, $i = 0, 1, \ldots, 359$ is the discretized direction index, θ_k denotes the true direction of the k-th source (in degrees), and σ controls the angular spread of the Gaussian distribution.

Building upon this direction encoding, we design two methods to further incorporate distance labels, namely a 1D Distance-Weighted Direction Gaussian and a 2D Direction–Distance Gaussian. In the following, we provide a detailed explanation of both methods.

1D Distance-Weighted Direction Gaussian: A simple approach is to encode distance directly into the intensity of the Gaussian distribution, as formulated in (6). In this formulation, nearer sound sources yield higher intensities, resulting in Gaussians with taller peaks. As illustrated in Figure 5b, Car 1, being closer, exhibits a higher peak than Car 2, which is farther away, while the spread of their Gaussians remains identical to those in Figure 5a.

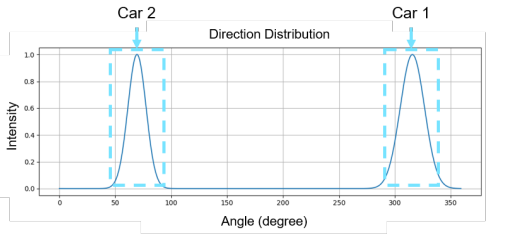
$$y[i] = \sum_{k=1}^{K} \frac{1}{d_k} \exp\left(-\frac{\Delta(i,\theta_k)^2}{2\sigma^2}\right), \tag{6}$$

where d_k denotes the distance of the k-th source, serving as an intensity weight (closer sources produce higher peaks).

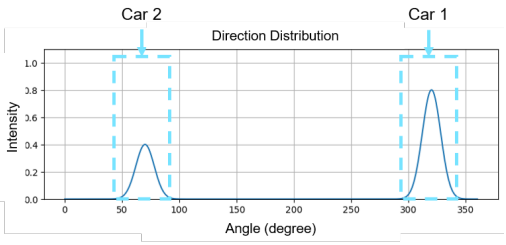


(a) 2D lable for localization (b) Direction-view Gaussian label (c) Distance-view Gaussian label

Figure 4. 2D Direction–Distance Gaussian for multiple sound source localization, with two Gaussians applied in polar coordinates along the direction and distance dimensions.



(a) 1D Gaussian direction label for direction estimation



(b) 1D Distance-Weighted Direction Gaussian

Figure 5. Illustration of different Gaussian label representations for sound source localization

2D Direction–Distance Gaussian: Figure 4 presents the labels for two targets in space (converted from polar to Cartesian coordinates), along with the corresponding views in the direction and distance domains. It can be seen that the direction dimension is identical to the labels in Figure 5a, with an additional dimension introduced to represent distance, as formulated in (7). The variance (σ) of the Gaussian distribution can be interpreted as the uncertainty in sound source localization. For direction estimation, closer objects occupy a larger angular range and therefore exhibit greater uncertainty. For distance prediction, farther objects are more difficult to estimate accurately, resulting in higher uncertainty. Overall, compared to closer objects, farther objects have greater uncertainty in distance and lower uncertainty in direction (i.e., a wider Gaussian in distance dimension and a narrower Gaussian in direction dimension).

$$y[i,j] = \sum_{k=1}^{K} \exp\left(-\frac{\Delta(i,\theta_k)^2}{2\sigma_\theta^2} - \Phi(j,d_k) \right), \tag{7}$$

$$\Phi(j,d_k) = \frac{(j-d_k)^2}{2\sigma_d^2}. \tag{8}$$

where j denotes the discretized distance bin ($j = 0,\ldots,D_{\max}$), θ_k and d_k are the direction and distance of the k-th source, and σ_θ and σ_d control the spread (uncertainty) in direction and distance, respectively.

3.3.2. Classification label

For the further classification task, we adapt the localization network into a multi-head architecture. Each head is dedicated to one sound category (e.g., car, truck, motorcycle) and outputs a class-conditioned localization label of the same form as in our localization task (e.g., a 1D direction distribution or a 2D direction–distance Gaussian). Intuitively, this label is a probability/intensity map over space indicating where sources of that class are likely to be. The detailed architecture will be presented later.

3.4. SonicVision Net

In the previous subsection, we defined the STFT real and imaginary parts $R_c(m,k)$ and $I_c(m,k)$ as inputs. We aim to combine them with the microphone positional encoding,

$$\mathbf{e}_c = \mathrm{PE}(\mathbf{p}_c), \tag{9}$$

and feed the concatenated features into the neural network f_Θ (as formulated in 10) to predict the target distribution.

$$\hat{y}[i,j](\hat{y}[i]) \;=\; f_\Theta\big(\{R_c(m,k),\, I_c(m,k),\, \mathbf{e}_c\}_{c=1}^{C}\big)\,. \tag{10}$$

In the introduction chapter, we have already presented the overall neural network architecture(Figure 2), while Figure 6 provides a more detailed version of the network.

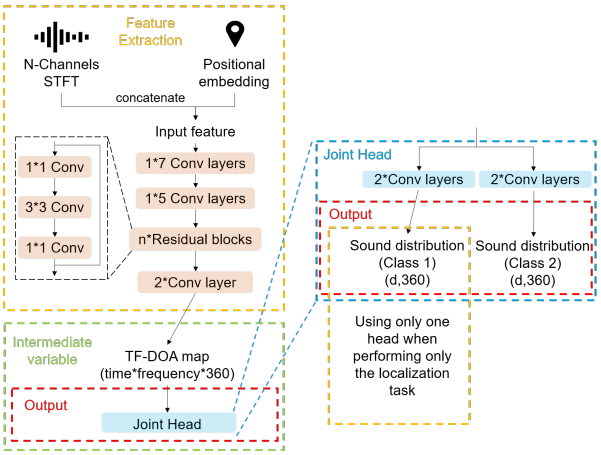


Figure 6. **SonicVision Design:** The backbone is based on a ResNet structure that extracts features from STFT real/imaginary inputs combined with microphone positional encoding. On top of the backbone, an ae joint head performs both classification and localization depending on the training objective. If only localization is performed, a single head is used.

As mentioned earlier, we adopt the STFT as the network input, which is particularly well-suited for processing by CNN-based architectures. And therefore, for the backbone, we employ the classical ResNet architecture, which is capable of extracting deep hierarchical features from image-like representations. Its residual connections not only help preserve essential low-level spectral information but also enable the learning of more complex time–frequency patterns from the STFT. Furthermore, by alleviating the vanishing gradient problem, ResNet facilitates the training of deeper networks, resulting in more robust feature extraction.

Passing through the backbone, we obtain a TF-DOA map with dimensions time × frequency × 360 (direction), which encodes the sound intensity for each time–frequency–direction bin. This representation enables the network not only to detect the presence of sound from a given direction, but also to infer its category and approximate distance by leveraging frequency and intensity cues.

In the output stage, If multiple sound types need to be classified, as shown in **Joint head** in Figure 6, we adopt a multi-head architecture in which each head shares the same structure as the localization network and outputs the BEV distribution for a specific sound type (e.g., motorcycle sounds). The the output has dimensions of $(d, 360)$, where the dimension d represents distance bins. For example, with a detection range of 30 m, we set $d = 60$, which corresponds to a distance resolution of 0.5 m.

If no classification task is required, we simply use one of the classification head alone to produce the sound distribution in the bird's-eye-view (BEV) domain.

3.5. Loss Function

For the above neural network architecture, we adopt two types of loss functions for training. Taking the localization network as an example, the target label is represented as a Gaussian distribution in a two-dimensional space of size $(n, 360)$, where all values lie within $[0, 1]$. One option is to use the mean squared error (MSE) loss to directly compare the predicted $(n, 360)$ matrix with the ground truth:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n \times 360} \sum_{i=1}^{n} \sum_{j=1}^{360} \big(\hat{y}[i,j] - y[i,j]\big)^2. \tag{11}$$

Alternatively, we can apply the binary cross-entropy (BCE) loss to each element individually:

$$\mathcal{L}_{\text{BCE}} = \frac{1}{n \times 360} \sum_{i=1}^{n} \sum_{j=1}^{360} \ell_{\text{BCE}}(\hat{y}[i,j], y[i,j]), \tag{12}$$

where the element-wise BCE term is defined as

$$\ell_{\text{BCE}}(\hat{y}, y) = -\Big(y \log \hat{y} + (1-y) \log(1-\hat{y})\Big). \tag{13}$$

We will compare the performance of these two approaches, (11) and (12), in the following experiment.

3.6. Dataset

As mentioned in the first chapter, to develop and validate the performance of our neural network, we have collected our own three datasets: simulation, test track, and real road data (shown in Table 2 and Figure 7).

- **Simulation data** Data are generated with the Acoular package by modeling the microphone array geometry and simulating point sources at specified locations. Real recordings of cars and motorcycles are imported into the simulation as source signals, and the resulting microphone signals are then used for training and evaluation.
- **Test track data** Data are collected at an open test track, where a car is parked at the center and one or more speakers play sounds of various traffic participants, such as cars and motorcycles. Although the car is stationary, wind speeds during data collection are typically between 5 and 10 m/s, partially simulating the effect of wind noise.
- **Road dataset** Data are collected either by parking the car at the roadside or by driving at 10 km/h, 30 km/h, and 50 km/h, thereby capturing the pass-by sounds of surrounding traffic participants. Both the environmental conditions and the sounds are entirely real, enabling evaluation of the network's performance in highly realistic scenarios. Due to time constraints, the classification network is not tested in this setting, as acquiring sufficient samples of diverse sound types would require substantial effort.

For annotating the real-world data, we use a LiDAR-based detector (CenterPoint) to obtain the positions of

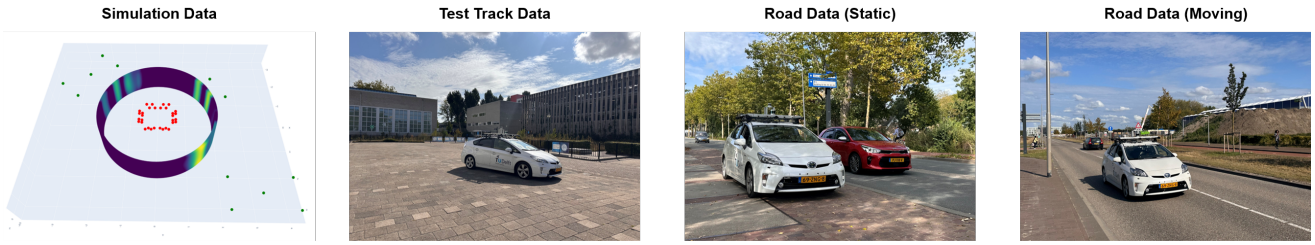


Figure 7. Dataset scenario across simulation, and real environments

Dataset	Number of samples	Sound generation	Task
Simulation	25000	Predefined virtual sound sources	Localization & Classification
Test Track	7500	Loudspeakers carried by pedestrians	Localization & Classification
Road Data	16000	Real traffic participants (cars, motorcycles, trucks)	Localization

Table 2. Overview of the datasets from simulation and real environment.

sound sources around the ego vehicle. On the test track, the detector identifies the positions of pedestrians carrying speakers (serving as the sound sources). On public roads, it provides the positions of surrounding traffic participants, where we focus on cars, motorcycles, trucks, and buses as the detection targets.

Since all tasks tested on the simulation data were also evaluated on the more realistic datasets (test track and road data), the results on the simulation data are presented in the appendix. For the test track and road datasets used in our experiments, an overview is provided in Figure 8 below.

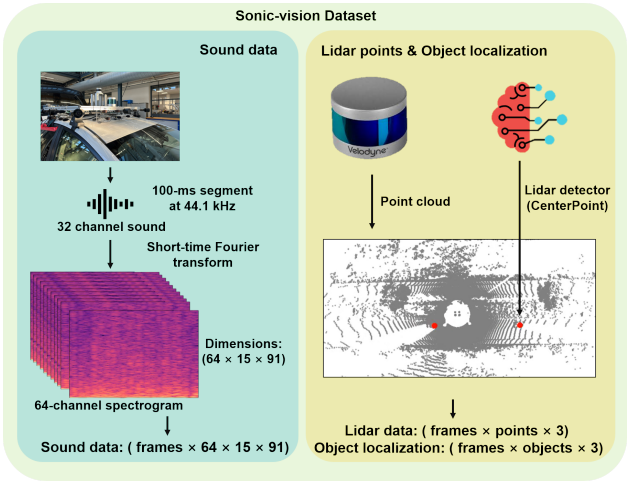


Figure 8. Sonic-vision dataset

4. Experiment

In this chapter, we investigates the two research questions introduced earlier, focusing on localization and joint localization–classification. We begin by describing the experimental setup in detail. Next, we present the results of our experiments. Finally, we conduct ablation studies to gain deeper insights into the proposed approach.

4.1. Experiment Setup

Label refinement: As mentioned earlier, we use CenterPoint to obtain the 3D positions of objects in the environment. Taking the road dataset as an example, since we are only interested in sound-emitting traffic participants, we filter out silent or irrelevant objects based on their velocity (e.g., stationary vehicles) and category (e.g., pedestrians on the roadside). In addition, we define a detection range for the microphone array, and any vehicles outside this range are also discarded. Finally, because the detector's predictions are not always entirely accurate, we perform manual label correction to the results, which amounted to nearly 20 hours of annotation work.

Labeling schemes: In the previous chapter, we introduced two approaches for localization labeling and two loss functions. In this chapter, we adopt the 2D Direction–Distance Gaussian encoding to represent distance and use MSE loss for training. The reasons for these choices are further analyzed in the ablation study.

Task-specific datasets: For the localization tasks, we used data collected on real roads for both training and testing. **For the classification task** specifically, due to time constraints, it was hard to record sounds from various types of vehicles on real roads. Instead, we conducted data collection on a test track, where motorcycle and car sounds were played through loudspeakers for training and testing.

Since we focus on an autonomous driving scenario, we evaluated its performance under both stationary conditions and while the vehicle was moving at different speeds (10 km/h, 30 km/h, and 50 km/h). Specifically, we compared the recognition of surrounding objects across these scenar-

Method	Acc@5°-5m(%)↑[1]	Acc@10°-5m(%)↑[2]	OA@-5°(%)↑[3]	OA@-10°(%)↑[4]	TP-ADE(m)↓[5]	TP-AAE(°)↓[6]
Beamforming	-	-	33.5	41.6	-	5.4
Beamforming + MLP	-	-	36.8	45.4	-	3.8
Iterative Beamforming	-	-	43.4	49.7	-	3.8
SonicVision (ours)	**58.9**	**67.4**	**62.1**	**69.1**	**2.3**	**3.2**

Table 3. Comparison of sound source localization performance across different methods. The proposed *SonicVision* achieves consistently higher accuracy and orientation scores, while yielding lower average distance and angle errors, demonstrating its effectiveness over beamforming-based baselines.

ios to assess the ability of the microphone array and the neural network to handle wind noise.

4.2. Baseline Method

Since both acoustic localization algorithms and sound-based localization–classification approaches are relatively novel in the audio domain, and even research on distance estimation from sound is still scarce, we only selected a baseline method—beamforming—for comparison in the direction estimation task. Moreover, as our focus is on multi-target recognition, and given the inherent limitations of conventional beamforming, we introduced two enhancements to improve its performance.

- **Beamforming MLP network** The output of beamforming is a prediction of sound intensity at a set of predefined spatial locations. We use this intensity prediction as the input to an MLP network, whose output is the estimated direction of the potential sound source.
- **Iterative beamforming** In each iteration, beamforming is performed to obtain the direction of the dominant sound source. The signal from this direction is then suppressed, and the beamforming algorithm is executed again. By repeating this process multiple times, the directions of multiple sound sources can be obtained.

4.3. Experiment 1: Sound Source Localization

4.3.1. Static Scenario

The first experiment addresses localization on the road dataset using microphone-recorded sound signals. In this part, **we first performed experiments on the simpler static dataset (with the car parked at the roadside)** and compared the results with the baseline method. The recognition performance is illustrated in Figure 9, and the metrics are summarized in Table 3.

From the figure and the table, it can be seen that the neural network achieves robust localization of passing vehicles: the green boxes represent the predicted positions, while the red dots denote the ground-truth locations. In terms of direction estimation, the proposed model also outperforms the baseline methods. Furthermore, as illustrated in the right panel of Figure 9, there are instances where the microphone detects a vehicle earlier than the LiDAR detector, which at that moment fails to register it. This highlights the complementary role of microphones in relation to vision-based detectors. Within a cooperative neural network framework, for example, the detection of a sound from a particular direction could guide the vision-based network to allocate more attention and computational resources to that region.

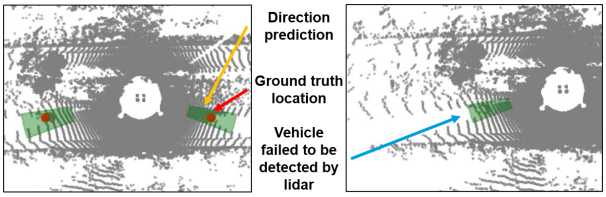


Figure 9. Localization performance under static motion. The green box represents the predicted sound source location, while the red dot indicates the ground truth position.

Moreover, from the experiment, we observe that our Beamforming+MLP method achieves more accurate angle predictions, which results in a lower TP-AAE. However, since it struggles with multi-target scenarios, the orientation accuracy shows little improvement.

4.3.2. Moving Scenario

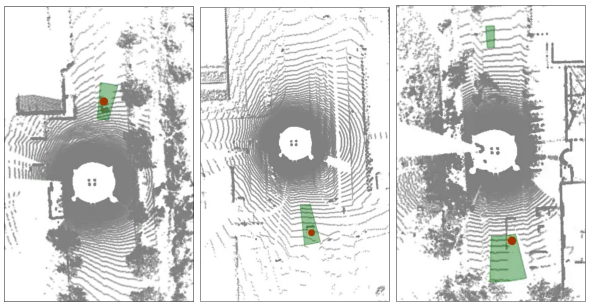
Considering that our application targets microphone-based recognition for autonomous driving, it is insufficient to evaluate performance only in stationary conditions. Therefore, we also conducted tests while the vehicle was in motion. Since wind noise varies with vehicle speed, we examined the performance of the microphone array under different conditions: 10 km/h (a representative low-speed environment), 30 km/h (the legal speed limit in Dutch residential areas), and 50 km/h (the legal speed limit on main roads).

[1] Accuracy with 5° and 5 m tolerance
[2] Accuracy with 10° and 5 m tolerance
[3] Orientation accuracy within 5°
[4] Orientation accuracy within 10°
[5] True positive sample average angle error
[6] True positive sample average distance error

Class	Acc@5° -5m(%)↑	Acc@10° -5m(%)↑	F1 score @5°-5m↑[1]	F1 score @10°-5m↑[2]	TP-ADE(m)↓	TP-AAE(°)↓
Car	82.0	85.9	0.89	0.92	1.4	1.7
Motorcycle	75.6	78.8	0.82	0.86	1.9	1.0

Table 4. Comparison of sound-based classification and localization performance across object classes. The proposed method achieves consistently higher accuracy and F1 scores for cars, while motorcycles show relatively lower performance.

To improve efficiency, the model initially trained in the stationary environment was subsequently adapted through fine-tuning with new data collected under different velocities. The resulting recognition performance and evaluation metrics are presented in Figure 10 and Table 5 below.



(a) 10 km/h (b) 30 km/h (c) 50 km/h

Figure 10. Localization performance of the proposed system under three different driving speeds: 10 km/h, 30 km/h, and 50 km/h.

Speed (km/h)	Acc@10°-5m(%)↑	TP-ADE(m)↓	TP-AAE(°)↓
0	83.2	1.7	2.5
10	81.2	1.9	2.8
30	77.8	2.3	3.2
50	63.1	3.5	4.8

Table 5. Localization performance of the proposed method under different vehicle speeds. (The test set here is relatively simple, so the overall performance is better than the static performance.)

From the experimental results above, we can see that in dynamic scenarios, the performance at 10 km/h and 30 km/h is strong. These two speeds cover typical use cases in urban residential areas, demonstrating that sound can be effectively utilized in real-world environments. However, at 50 km/h, the neural network performs slightly worse, which can be largely attributed to the increased severity of wind noise at higher speeds. This is consistent with our observations: under such conditions, traditional beamforming methods almost fail to operate. To further improve robustness, we plan to collect a larger and more diverse dataset in future work, which we expect will enhance the system's ability to cope with adverse noise conditions.

However, apart from improvements in the model and the data, we also considered that for such high-speed scenarios, under severe wind noise, a more effective solution should be pursued from a hardware design perspective. For example, professional microphones used in film production often employ specialized designs to suppress wind noise and enhance directional sensitivity. Moreover, in our current setup, the microphones are directly exposed on the outside of the vehicle; designing a semi-enclosed housing could further reduce wind noise and improve robustness.

4.4. Experiment 2: Sound Source Classification

Through the above experiments, we have demonstrated that vehicles can localize surrounding targets under both stationary and moving conditions. Nevertheless, in the broader context of 3D object detection for autonomous driving, it is essential not only to localize objects but also to recognize their categories (e.g., car, truck, motorcycle).

In this task, we adopt the multi-head neural network architecture previously illustrated in Figure 6, where each head outputs the localization of a specific sound category. Since road data makes it difficult to collect a sufficient number of samples for both motorcycle and car at the same time, we base our experiments on the available dataset. The detection performance and evaluation metrics are presented in Figure 11 and Table 4.

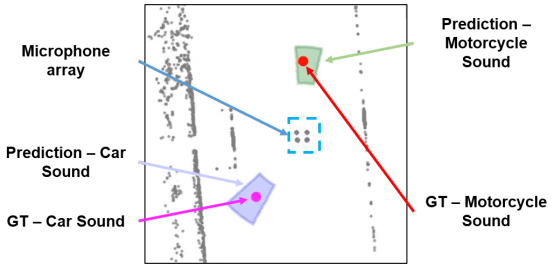


Figure 11. Illustration of classification and localization performance. The green bounding box indicates the predicted motorcycle position, while the purple bounding box indicates the predicted car position. The red dot marks the ground-truth motorcycle location, and the magenta dot marks the ground-truth car location.

[1]F1 score with 5° and 5 m tolerance

[2]F1 score with 5° and 5 m tolerance

The experimental results show that our neural network is capable of both localizing objects using sound and classifying their types. However, the network attains higher accuracy for cars than for motorcycles. This discrepancy likely arises from frequency differences, as car sounds align more closely with the aperture characteristics of the microphone array. As a result, given the same amount of training data, the model performs better for cars.

Regarding the previously mentioned challenge of limited road data, we also recognize that this issue may partly arise from the design of our current model. Specifically, the network trains the motorcycle head only when motorcycle sounds are detected, and the car head only when car sounds are detected. A potential improvement would be to design a labeling strategy that enables classification of different sound sources on top of localization. Such an approach could reduce the dependence on large amounts of class-specific training data.

4.5. Ablation Study

In the previous section, we demonstrated that our neural network can address both object localization and classification tasks in autonomous driving scenarios. As described in the Methodology section, several design choices regarding the loss function and training strategies were introduced. In this chapter, we conduct an ablation study to analyze the impact of these choices in detail and discuss practical tricks and hyperparameter settings used in the network design.

4.5.1. Comparison of Labeling Strategies for Localization

In the output section, we propose two approaches for labeling localization: 1D Distance-Weighted Direction Gaussian (1D Gaussian, Figure 5b) and 2D Direction–Distance Gaussian (2D Gaussian, Figure 4a). To compare the effectiveness of the two labeling schemes, we conducted experiments on the test track dataset.

Method	Acc@10°-5m(%)↑	TP-AAE(°)↓	TP-ADE(m)↓
1D Gaussian	74.1	**1.7**	2.0
2D Gaussian	**79.4**	1.1	**1.2**

Table 6. Comparison of labeling strategies for sound localization

From Table 6, it can be observed that although the two labeling methods yield similar performance in direction estimation, the 2D Gaussian approach achieves nearly 1 meter lower error in distance prediction compared to the 1D Gaussian method, and also provides higher recognition accuracy.

Our observation is that encoding distance as a 1D Gaussian with a relatively low peak (e.g., 0.3) is inherently harder to learn than a 2D Gaussian with a peak of 1. One reason, as illustrated in Figure 12, is that in the 2D case the sigmoid activation essentially learns a binary-like decision — whether there is sound at a given location — which aligns well with its natural role as a probability estimator.

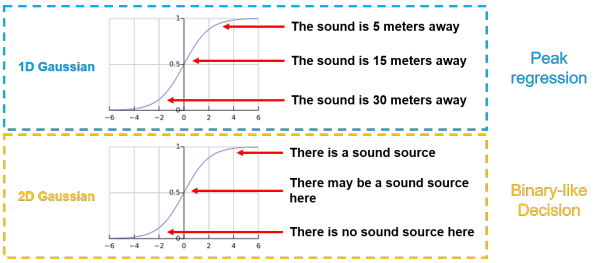


Figure 12. Comparison of 1D and 2D Gaussian labeling strategies, illustrating the different behaviors of the sigmoid activation.

By contrast, in the 1D case the sigmoid must not only decide if sound is present but also regress the precise peak magnitude, which is far more challenging. This difficulty is further amplified because intermediate values (e.g., 0.3) are harder for the network to approximate reliably, and smaller peaks provide weaker gradient signals during training.

4.5.2. Loss Function Selection

For the neural network labels (a 2D Gaussian distribution defined over the $n \times 360$ space), we pproposed two training strategies: (1) MSE loss, and (2) BCE loss applied at each position. Their performance on the localization task is compared, with the results summarized in Table 7.

Loss function	Acc@10°-5m(%)↑	TP-AAE(°)↓	TP-ADE(m)↓
MSE	**61.76**	**3.2**	1.9
BCE	60.29	3.4	**1.7**

Table 7. Comparison of performance using different loss functions.

Through comparison, we observe that the two loss functions achieve nearly identical training performance. The MSE-based approach yields slightly higher accuracy and smaller angular error, whereas BCE produces a lower distance error. This finding is consistent with the results reported by Perotin et al.[25], although their study focused solely on the impact of the two loss functions for direction estimation. They also demonstrated that BCE and MSE achieve similar effectiveness when applied to Gaussian labels, such as those illustrated in Figure5a.

4.5.3. Error Analysis for Acoustic Localization

In the problem definition section, we proposed two assumptions for setting the labels, which are as follows:

- The uncertainty in distance (represented by the Gaussian distribution's σ) increases linearly with distance:

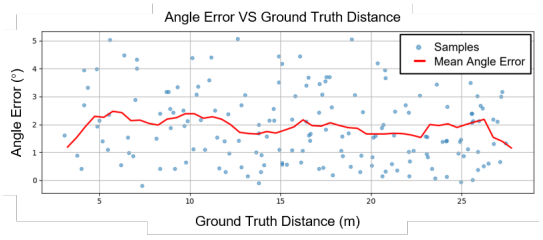
$$\sigma_d(d) = \sigma_d^{\min} + \left(\sigma_d^{\max} - \sigma_d^{\min}\right) \frac{d}{D_{\max}}, \quad (14)$$

- The uncertainty in direction (represented by the Gaussian distribution's σ) decreases linearly with distance:

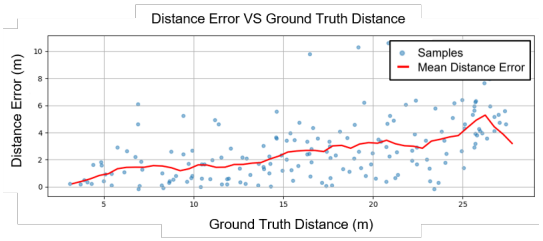
$$\sigma_\theta(d) = \sigma_\theta^{\max} - \left(\sigma_\theta^{\max} - \sigma_\theta^{\min}\right) \frac{d}{D_{\max}}, \quad (15)$$

where d is the source distance and $D_{\max}$ is the maximum distance bin.

These two points stem from the intuition that as an object becomes farther away, it occupies a smaller angular range and, from the perspective of human hearing, its position also becomes more difficult to estimate (especially the distance). To examine this hypothesis, we conducted the following experiment, analyzing how both distance error and direction error vary with increasing distance.



(a) Relationship between direction error and ground truth distance



(b) Relationship between distance error and ground truth distance

Figure 13. Distance-dependent error analysis

Direction error analysis: We compared the magnitude of direction error under different ground-truth settings, as illustrated in Figure 13a. Contrary to our expectation, we found that the direction error remains almost unchanged with respect to the ground truth. In other words, although a 3-meter-long vehicle subtends only about 6° when located 30 m away and nearly 35° when only 5 m away, this variation does not affect the recognition performance.

Therefore, we conducted an additional experiment, in which we compared the cases with distance-dependent uncertainty (in the direction dimension) against those with constant uncertainty. The results are summarized in Table 8, which indicate that the dynamic uncertainty encoding approach leads to improved recongnition accuracy and reduced angle error

Method	OA@-5°(%)↑	OA@-10°(%)↑	TP-AAE(°)↓
Static Label	56.5	64.3	3.3
Dynamic label	**58.9**	**67.4**	**3.1**

Table 8. Comparison of static and dynamic direction labeling strategies. The dynamic label adjusts directional uncertainty with distance and achieves better overall performance.

Distance error analysis. We analyzed distance error under different ground-truth settings(distance), as shown in Figure 13b. As expected, the distance error increases with source distance, partially validating our hypothesis. Nevertheless, further experiments are required to determine whether this growth follows a linear or exponential trend.

4.5.4. Positional Embedding

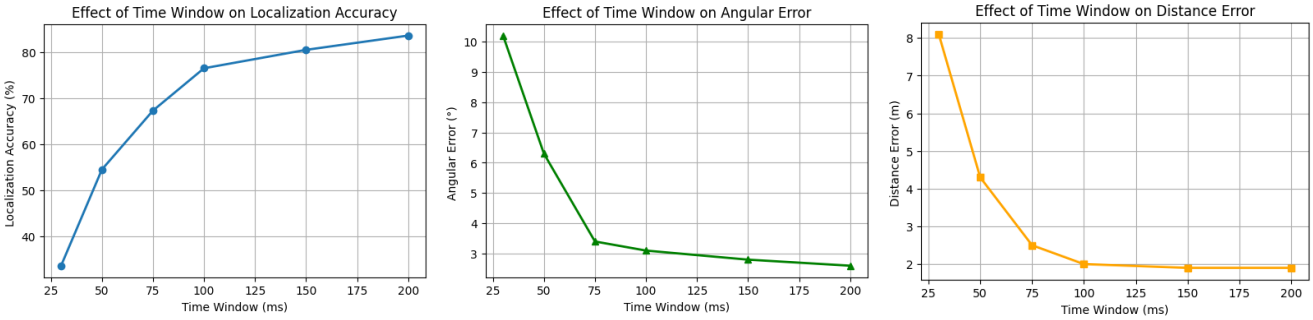
In our network design2, we incorporate positional embeddings after each microphone signal to help the model capture the spatial relationships. Therefore, in this ablation study, we compare the performance with and without positional embeddings, as shown in Table 9.

Positional Embedding	Acc@10°-5m(%)↑	TP-AAE(°)↓	TP-ADE(m)↓
×	65.6	2.9	1.9
✓	**68.9**	**2.7**	**1.7**

Table 9. Effect of Positional Embedding on Performance.

It can be observed that positional embedding yields a modest improvement in recognition performance. This is an interesting finding, as we neither changed the microphone positions within the array nor altered the input ordering. The key difference lies in the learning objective: with positional embeddings, the network is encouraged to learn the absolute position of each channel with respect to a spatial reference, whereas without embeddings, it must rely only on relative ordering.

Two factors may explain this improvement. First, positional embeddings provide the model with a shared coordinate system, allowing it to better associate phase differences with the physical microphone geometry, rather than inferring such relations solely from channel indices. Second, by supplying explicit positional cues, the model reduces the burden of mapping from the acoustic domain to the geometric domain, thereby lowering the learning difficulty.



(a) Effect of time window on localization accuracy (b) Effect of time window on angular error (c) Effect of time window on distance error

Figure 14. Effect of time window length on localization performance. Improvements continue with longer windows, but the growth becomes gradual after 100 ms.

4.5.5. STFT Time Windows Length

In the input stage, we adopt 100 ms audio segments converted into STFT representations as our baseline, following prior work on direction estimation. Longer windows capture richer information and provide higher frequency resolution, but they also increase model complexity and may introduce time delay and localization errors due to target motion. Conversely, shorter windows reduce the parameter size and are more robust to moving sources, but at the cost of losing spectral detail. To better understand this trade-off, we investigate the impact of different STFT window lengths in this ablation study. We evaluate multiple window sizes ranging from 30 ms to 200 ms and compare their performance on the static road dataset, as illustrated in Fig. 14.

Through our observations, we found that as the time window increases, recognition performance consistently improves. However, we must also take into account the latency introduced by a larger window—for example, a 200 ms window implies an additional 100 ms delay before recognition can be made. Our results show that once the window length exceeds about 80-100 ms, the improvements in recognition accuracy and error(direction and distance) reduction become relatively marginal. This indicates that our earlier choice of a 100 ms window length appears to be reasonably justified, as it strikes a balance between recognition performance, network parameter efficiency, and latency.

5. Discussion

Through the two experiments above, our approach demonstrates the capability to localize and classify traffic participants in the environment using a microphone array. We have also successfully addressed our initial research questions: based on the STFT experiments and the comparisons presented in Appendix A, we identified a suitable encoding strategy for auditory information. Furthermore, two ablation studies on label design confirm the effectiveness of our proposed representation, which is specifically tailored for BEV acoustic localization in autonomous driving. This labeling scheme not only enables more robust learning but also bridges the gap in the robotics audio perception domain, where direction and distance have typically been treated in an isolated manner. In addition, two studies on network design choices—specifically, positional embedding and the loss function—further support the **effectiveness** of our network design.

Furthermore, although microphone performance may degrade in complex environments, **they remain valuable for enhancing other sensors through joint detection.** For example, in our experiments, we observed instances where acoustic sensing successfully identified objects that LiDAR failed to detect. Such failures were typically caused by partial occlusions in the LiDAR's field of view or by sparser point clouds resulting from sensor overheating. In these cases, acoustic information provided complementary cues that enabled cross-validation and more reliable recognition. Specifically, when microphones detect sound from a region, the system can assign higher confidence to that location, effectively favoring it as a true positive even when LiDAR evidence is ambiguous.

5.1. Limitations and future work

Despite achieving encouraging results, our work is still subject to the following limitations.

- The dataset was collected primarily in relatively simple road environments, without highly congested scenarios.
- The current experiments did not fully explore the vision-independent nature of sound; future work could investigate more challenging cases, such as narrow T-junctions.
- The hardware setup relied on simple, fully exposed microphones, making it more susceptible to wind noise.

To tackle these limitations, we outline the following directions for future work: we plan to expand the dataset to include highly congested downtown scenarios, enabling a more thorough evaluation of model performance under

complex traffic conditions. Another promising direction is to investigate the non-line-of-sight (NLOS) problem by combining microphone array data with environmental information, such as maps or LiDAR points. In such cases, knowledge of the road layout combined with acoustic cues from the environment can help recognize vehicles hidden behind corners. Finally, we aim to improve the microphone design by introducing stronger wind-noise suppression covers, exploring more advanced types such as directional or ear-shaped microphones, and adopting semi-enclosed layouts to reduce direct wind exposure and enhance robustness in real-world deployments.

6. Conclusion

In the previous chapter, we provided answers to the research questions at the beginning. In summary, we proposed an acoustic detection method that leverages low-cost microphone arrays for autonomous driving to detect critical surrounding objects. This approach successfully extends prior research in indoor robotics from simple direction estimation to full localization and classification, and introduces acoustic perception into the autonomous driving domain from a bird's-eye-view perspective. Our experiments demonstrate that the proposed method can not only reliably detect traffic participants in the environment, but in certain cases can even identify objects that vision-based sensors fail to recognize—highlighting its value both as an independent sensor and as a complementary modality. **Overall, our findings provide strong evidence that microphone arrays could contribute to 3D perception for autonomous vehicles**. Looking ahead, we hope that acoustic sensing will be applied in broader scenarios, enabling vehicles to perceive the world more like humans do—through both eyes and ears.

References

[1] Sharath Adavanne, Archontis Politis, and Tuomas Virtanen. Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network. In *EUSIPCO*, 2018. 3

[2] J. B. Allen. Short term spectral analysis, synthesis, and modification by discrete fourier transform. *IEEE transactions on acoustics, speech, and signal processing*, 1977. 5, 16

[3] Abhishek Balasubramaniam and Sudeep Pasricha. Object detection in autonomous vehicles: Status and open challenges. *arXiv*, 2022. 3

[4] Vasileios Bountourakis, Lazaros Vrysis, and George Papanikolaou. Machine learning algorithms for environmental sound recognition: Towards soundscape semantics. In *Proceedings of the audio mostly 2015 on interaction with sound*. 2015. 3

[5] Andreas Brendel and Walter Kellermann. Distance estimation of acoustic sources using the coherent-to-diffuse power ratio based on distributed training. In *IWAENC*, 2018. 3

[6] Amandine Brunetto, Sascha Hornauer, X Yu Stella, and Fabien Moutarde. The audio-visual batvision dataset for research on sight and sound. In *IROS*, 2023. 2, 3

[7] Praneeth Chakravarthula, Jim Aldon D'Souza, Ethan Tseng, Joe Bartusek, and Felix Heide. Seeing with sound: Long-range acoustic beamforming for multimodal scene understanding. In *CVPR*, 2023. 2, 3

[8] Selina Chu, Shrikanth Narayanan, C-C Jay Kuo, and Maja J Mataric. Where am i? scene recognition for mobile robots using audio features. In *ICME*, 2006. 3

[9] Dengxin Dai, Arun Balajee Vasudevan, Jiri Matas, and Luc Van Gool. Binaural soundnet: predicting semantics, depth and motion with binaural sounds. *TPAMI*, 2022. 2

[10] Ali Dalir, Ali Asghar Beheshti, and Morteza Hoseini Masoom. Classification of vehicles based on audio signals using quadratic discriminant analysis and high energy feature vectors. *arXiv*, 2018. 3

[11] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *CVPR*, 2019. 2

[12] Oguzhan Gencoglu, Tuomas Virtanen, and Heikki Huttunen. Recognition of acoustic events using deep neural networks. In *EUSIPCO*, 2014. 3

[13] Alireza Ghasemieh and Rasha Kashef. 3d object detection for autonomous driving: Methods, models, sensors, data, and challenges. *Transportation Engineering*, 2022. 1

[14] Pierre-Amaury Grumiaux, Sran Kitić, Laurent Girin, and Alexandre Guérin. A survey of sound source localization with deep learning methods. *JASA*, 2022. 3

[15] Mingyang Hao, Fangli Ning, Ke Wang, Shaodong Duan, Zhongshan Wang, Di Meng, and Penghao Xie. Acoustic non-line-of-sight vehicle approaching and leaving detection. *T-ITS*, 2024. 2

[16] Weipeng He, Petr Motlicek, and Jean-Marc Odobez. Deep neural networks for multiple speaker detection and localization. In *ICRA*, 2018. 3, 5

[17] Hugo Jallet, Emre Cakır, and Tuomas Virtanen. Acoustic scene classification using convolutional recurrent neural networks. *DCASE*, 2017. 3

[18] Mingu Jeon, Jae-Kyung Cho, Hee-Yeun Kim, Byeonggyu Park, Seung-Woo Seo, and Seong-Woo Kim. Non-line-of-sight vehicle localization based on sound. *T-ITS*, 2024. 2, 3

[19] Gareth Jones. Echolocation. *Current Biology*, 2005. 2

[20] Dongjin Kim, Sung Jin Um, Sangmin Lee, and Jung Uk Kim. Learning to visually localize sound sources from mixtures without prior source knowledge. In *CVPR*, 2024. 3

[21] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 2003. 16

[22] Abdullah Küçük and Issa MS Panahi. Convolutional recurrent neural network based direction of arrival estimation method using two microphones for hearing studies. In *MLSP*, 2020. 3

[23] Beth Logan et al. Mel frequency cepstral coefficients for music modeling. In *Ismir*, 2000. 16

[24] Michael Neri, Archontis Politis, Daniel Aleksander Krause, Marco Carli, and Tuomas Virtanen. Speaker distance estimation in enclosures from single-channel audio. *TASLP*, 2024. 3

[25] Lauréline Perotin, Alexandre Défossez, Emmanuel Vincent, Romain Serizel, and Alexandre Guérin. Regression versus classification for neural network based audio source localization. In *WASPAA*, 2019. 11

[26] Yvan Petillot, I Tena Ruiz, and David M Lane. Underwater vehicle obstacle avoidance and path planning using a multi-beam forward looking sonar. *IEEE-JOE*, 2001. 2

[27] Archontis Politis, Sharath Adavanne, and Tuomas Virtanen. A dataset of reverberant spatial sound scenes with moving sources for sound event localization and detection. *arXiv*, 2020. 3

[28] Reuters. China's hesai to halve lidar prices next year, sees wide adoption in electric cars, 2024. https://www.reuters.com/technology/chinas-hesai-halve-lidar-prices-next-year-sees-wide-adoption-electric-cars-2024-11-27/. 1

[29] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE SPL*, 2017. 3

[30] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *ACMMM*, pages 1041–1044, 2014. 3

[31] Prasanga N Samarasinghe, Thushara D Abhayapala, MA Polettfi, and Terence Betlehem. On room impulse response between arbitrary points: An efficient parameterization. In *IS-CCSP*, 2014. 3

[32] Jonghee Sang, Soomyung Park, and Junwoo Lee. Convolutional recurrent neural networks for urban sound classification using raw waveforms. In *EUSIPCO*, 2018. 3

[33] Ralph Schmidt. Multiple emitter location and signal parameter estimation. *TAP*, 1986. 3

[34] Yannick Schulz, Avinash Kini Mattar, Thomas M Hehn, and Julian FP Kooij. Hearing what you cannot see: Acoustic vehicle detection around corners. *RA-L*, 2021. 2, 3

[35] Yu Su, Ke Zhang, Jingyu Wang, and Kurosh Madani. Environment sound classification using a two-stream cnn based on decision-level fusion. *Sensors*, 2019. 3

[36] Elisabet Tiana-Roig, Finn Jacobsen, and Efrén Fernández Grande. Beamforming with a circular microphone array for localization of environmental noise sources. *JASA*, 2010. 3

[37] Robert J Urick. Principles of underwater sound-2. 1975. 2

[38] Burak Uzkent, Buket D Barkana, and Hakan Cevikalp. Non-speech environmental sound classification using svms with a new set of features. *IJICIC*, 2012. 3

[39] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *CVPR*, 2021. 2, 3

[40] Jia-Ching Wang, Hsiao-Ping Lee, Jhing-Fa Wang, and Cai-Bei Lin. Robust environmental sound recognition for home automation. *T-ASE*, 2008. 3

[41] Wikipedia contributors. Sonar. https://en.wikipedia.org/wiki/Sonar, 2025. Accessed: 2025-08-11. 2

[42] Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li. A learning-based approach to direction of arrival estimation in noisy and reverberant environments. In *ICASSP*, 2015. 3

[43] Mariam Yiwere and Eun Joo Rhee. Sound source distance estimation using deep learning: An image classification approach. *Sensors*, 2019. 3

[44] Yi Zhan and Tadahiro Kuroda. Wearable sensor-based human activity recognition from environmental background sounds. *AIHC*, 2014. 3

[45] Haomin Zhang, Ian McLoughlin, and Yan Song. Robust sound event recognition using convolutional neural networks. In *ICASSP*, 2015. 3

[46] Jannik Zürn and Wolfram Burgard. Self-supervised moving vehicle detection from audio-visual cues. *RA-L*, 2022. 2

A. Appendix: Feature Extraction

In this thesis, several commonly used feature extraction methods were reviewed and compared, including raw waveforms, GCC-PHAT[21], cepstral features (MFCCs)[23], and spectral features such as STFT[2]. Table 10 summarizes their respective merits and limitations. Below we briefly introduce each method.

Raw Waveform. The raw audio signal $x(t)$ contains the full temporal information without any preprocessing. Its main advantage is completeness, but neural networks must learn task-specific representations directly from $x(t)$, which typically requires deep models (e.g., CNNs, RNNs, Transformers).

GCC-PHAT. Generalized Cross-Correlation with Phase Transform estimates the time delay τ between two microphone signals $x_1[n]$ and $x_2[n]$:

$$R_{12}(\tau) = \mathcal{F}^{-1}\left\{ \frac{X_1(k)X_2^*(k)}{|X_1(k)X_2^*(k)|} \right\}, \tag{16}$$

where $X_1(k)$ and $X_2(k)$ are Fourier transforms of the signals. This emphasizes phase differences while reducing sensitivity to noise and reverberation.

Cepstral Features (MFCCs). MFCCs approximate human auditory perception by mapping the spectrum to the Mel scale, applying log compression, and then computing the Discrete Cosine Transform (DCT):

$$c[n] = \sum_{m=1}^{M} \log(S_m) \cos\left[\frac{\pi n}{M}(m - 0.5) \right], \tag{17}$$

where S_m is the Mel-scaled spectral energy. MFCCs are compact and efficient, but sensitive to noise.

Spectral Features (STFT). The Short-Time Fourier Transform provides a time–frequency representation:

$$X(\tau, \omega) = \int_{-\infty}^{\infty} x(t)\, w(t - \tau) e^{-j\omega t}\, dt, \tag{18}$$

where $w(\cdot)$ is a window function. STFT captures both magnitude and phase information, which are critical for direction-of-arrival (DOA) estimation.

Based on the review above, raw waveforms preserve the entire acoustic signal, but they inevitably include noise and redundant information, making effective feature extraction more demanding for the neural network. GCC-PHAT addresses this by focusing on inter-microphone time-delay cues and is robust in reverberant environments, yet it discards much of the spectral content that can be valuable for richer perception tasks. MFCCs provide a compact representation inspired by human auditory perception and are widely used in classification, but their compression of spectral and temporal details limits their applicability to localization.

Compared with the above features, STFT retains both magnitude and phase information while suppressing part of the noise through its windowed transformation, providing a richer time–frequency representation that is well suited to our tasks of classification and localization, and is therefore adopted in this work.

B. Appendix: Microphone Array Design

As discussed in our related works, most current acoustic research in autonomous driving focuses on microphone layouts arranged in a vertical configuration. In contrast, we aim to perform recognition from a bird's-eye-view perspective, which requires the design of a custom horizontally arranged microphone array. In the preliminary stage of this thesis, we employed the Acoular library in Python to design a horizontal microphone array tailored for autonomous vehicles and evaluated its performance in simulation using beamforming methods (Shown in Figure 15).

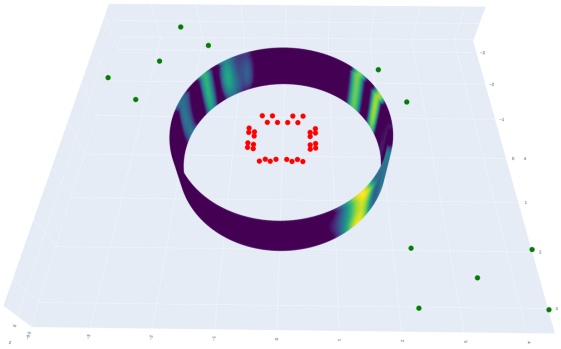


Figure 15. Beamforming simulation with the proposed microphone array.

In this simulation setup, the proposed horizontal microphone array is visualized by the red points, while the green points represent the positions of the predefined sound sources. Beamforming is performed over the circular band surrounding the array, with the highlighted region corresponding to the estimated direction of arrival. The alignment between the detected direction and the ground truth source positions demonstrates the effectiveness of the array design for sound localization tasks.

After finalizing the microphone placement design, we utilized the mounting rack on top of our autonomous vehicle to install eight microphone plates around the vehicle, with two plates on each side. Each plate is equipped with four microphones, resulting in a total of 32 channels. The plates

Method	Typical Application	Strengths / Weaknesses
Raw Waveform	Classification, Distance Estimation	• **Strength:** Preserves full information without preprocessing. • **Weakness:** Requires deep architectures; less interpretable.
GCC-PHAT	Direction of Arrival Estimation	• **Strength:** Encodes inter-microphone time delay, robust to reverberation. • **Weakness:** Loses frequency and intensity information; scales poorly with many microphones.
MFCCs	Speech / Sound Classification	• **Strength:** Compact, efficient, aligned with human hearing. • **Weakness:** Sensitive to noise; limited temporal information.
Spectral Features (STFT)	Classification, DOA Estimation	• **Strength:** Captures rich frequency–time patterns and phase information, critical for DOA tasks. • **Weakness:** Computationally more demanding.

Table 10. Comparison of feature extraction methods considered.

are fabricated from PMMA material, which offers a balance between rigidity and lightweight properties. A schematic illustration is provided in Fig. 16.

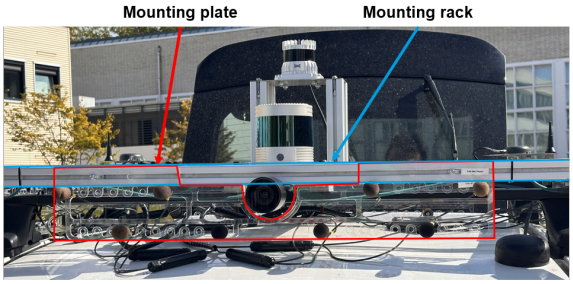


Figure 16. Horizontal microphone array installation.

C. Appendix: Simulation Test

We used the Python Acoular package to design simulations and perform an initial validation of our neural network design. In the simulation, the microphone array geometry can be specified, point sound sources can be generated at arbitrary spatial locations, and real recordings can be injected into these sources. The setup is illustrated in Figure 17

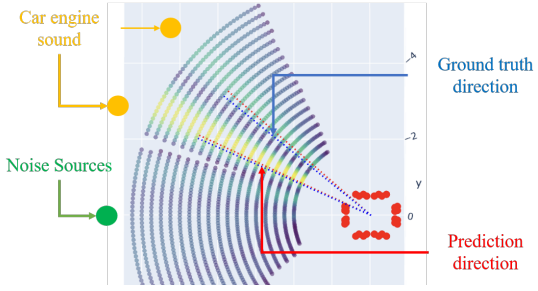


Figure 17. Simulation setup.

As an initial step, the simulation experiments targeted direction estimation and classification, with distance modeling introduced in later parts(test-track and road data) of this work. In the direction estimation task, the neural network achieved the performance summarized in Table 11.

Method	Acc@ 5°(%)↑	Acc@ 10°(%)↑	TP-AAE(°)↓
SonicVision	85.2	92.3	0.9

Table 11. Direction Estimation Performance with Simulation Data

The results in the Table above demonstrate that our neural network effectively handles sound source direction estimation, maintaining strong performance even in multi-source scenarios. In addition, the performance of the sound source classification task is presented in Table 12.

Target	Overall Acc (%)↑	Class-wise Acc (%)↑	TP-AAE(°)↓
Car	84.2	86.7	1.0
Motorcycle		81.3	1.6

Table 12. Classification Performance with Simulation Data

The table shows that our network can also perform the classification task effectively, although recognition of car sounds is slightly stronger than that of motorcycles. This is partly because car sounds align more closely with the aperture characteristics of our microphone array.

In summary, the simulation data allowed us to verify the feasibility of our neural network for direction estimation and joint classification with direction. However, distance estimation was not validated, and compared to real-world data, the simulations lack environmental noise and cannot incorporate obstacles that may affect sound propagation. As a result, they fall short in realism and require further validation on real data.