Seeing With Sound:
Long-range Acoustic Beamforming for Multimodal Scene Understanding

Praneeth Chakravarthula[1] Jim Aldon D'Souza[2] Ethan Tseng[1] Joe Bartusek[1] Felix Heide[1,2]
[1]Princeton University [2]Algolux

Abstract

Mobile robots, including autonomous vehicles rely heavily on sensors that use electromagnetic radiation like lidars, radars and cameras for perception. While effective in most scenarios, these sensors can be unreliable in unfavorable environmental conditions, including low-light scenarios and adverse weather, and they can only detect obstacles within their direct line-of-sight. Audible sound from other road users propagates as acoustic waves that carry information even in challenging scenarios. However, their low spatial resolution and lack of directional information have made them an overlooked sensing modality. In this work, we introduce long-range acoustic beamforming of sound produced by road users in-the-wild as a complementary sensing modality to traditional electromagnetic radiation-based sensors. To validate our approach and encourage further work in the field, we also introduce the first-ever multimodal long-range acoustic beamforming dataset. We propose a neural aperture expansion method for beamforming and demonstrate its effectiveness for multimodal automotive object detection when coupled with RGB images in challenging automotive scenarios, where camera-only approaches fail or are unable to provide ultra-fast acoustic sensing sampling rates. Data and code can be found here[1].

1. Introduction

Autonomous mobile robots of today predominantly rely on several electromagnetic (EM) radiation-based sensing modalities such as camera, radar and lidar for diverse scene understanding tasks, including object detection, semantic segmentation, lane detection, and intent prediction. The most promising approaches rely on fused data input from these camera, lidar and radar sensor configurations [7, 42, 50] and robust data-driven perception algorithms using convolutional neural networks or vision transformers. However, existing camera/radar/lidar stacks do not return signal for objects with low reflectance and in conditions where light-based sensors struggle, such as severe scattering due to fog. All existing EM radiation-based sensor systems (active or passive) are fundamentally limited by the propagation of EM waves.

Acoustic waves are an alternative and complementary sensing modality that are not subject to these limitations. Every automotive vehicle generates noise due to engine/transmission, aerodynamics, braking, and contact with the road. Even electric vehicles are required by law to emit sound to alert pedestrians [36]. However, acoustic sensing is not without challenges. Spatially resolving the acoustic spectrum at meter wavelengths (e.g., a $1\,\mathrm{kHz}$ sound wave has a wavelength of about $35\,\mathrm{cm}$ in air) has limited existing approaches to low-resolution tracking of 3D spatial coordinates [11–13, 32, 44].

In this work, we show that acoustic sensing is complementary to existing EM wave-based sensors, robust to challenging scenarios, and achieves improved performance when combined with existing vision-only approaches. To this end, we captured a large multimodal dataset with a prototype vehicle equipped with a 1024 (32x32 grid) microphone array and a plethora of vision sensors, and had them labeled by human annotators, which we release as the *first multimodal long-range beamforming dataset*. To the best of our knowledge, there is no such large and diverse multimodal acoustic beamforming dataset, as also illustrated in Table 1. We additionally propose a neural acoustic beamforming method for small aperture microphone arrays via learned aperture expansion. The aperture-expanded beamforming maps recover spatial resolution typically lost in sound measurements, and facilitate fusion with visual inference tasks. We assess multimodal visual and acoustic vision tasks in diverse real-world driving scenarios. We validate that *visual and acoustic signals can complement each other* in challenging automotive scenarios and can enable future frame predictions at kHz frequencies. We also demonstrate that object detection using vision and acoustic signals outperform that of vision-only signals in challenging low-light scenarios. Furthermore, we show the applicability of acoustic sensing in non-line-of-sight and partially occluded scenes where purely vision-based sensing fails.

Specifically, we make the following contributions:

[1] light.princeton.edu/seeingwithsound

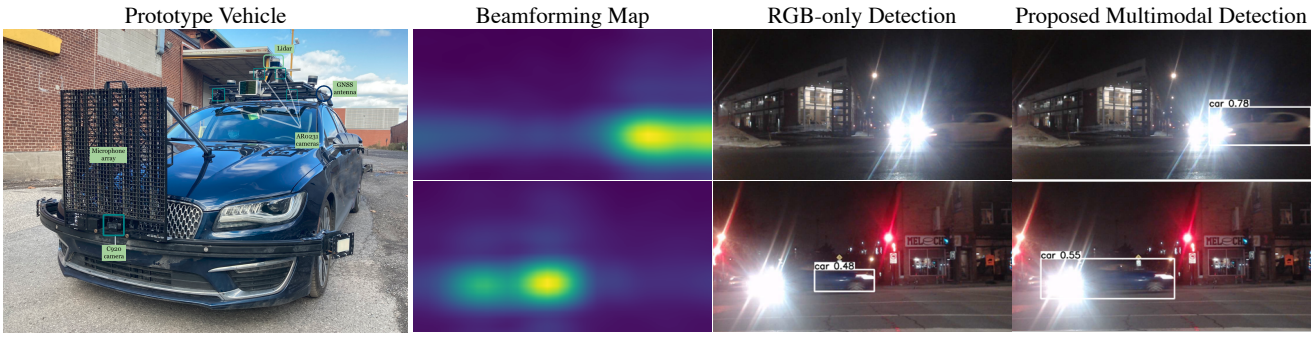


Figure 1. We capture a large dataset of acoustic pressure signals at several frequencies from roadside noise using our prototype test vehicle (left). Of the available 250-5000 Hz frequency bands in the dataset, we visualize beamformed signals at the 4000 Hz octave band here (middle left). Using RGB-only result in missed and inaccurate detections at night (middle right). The complementary nature of acoustic signals, on the other hand, helps robustly detect the objects in challenging night scenarios (right).

Table 1. Existing beamforming works and datasets are limited to just a few hundred processed beamforming maps and a single RGB camera data. In stark contrast, our dataset is very large with acoustic signals captured at 40kHz frequency across 11 frequency bands in diverse urban scenarios.

Dataset	Michel et al. [35]	Zunino et al. [51]	Guidati [21]	Proposed
Ego Motion	Static	Static	Static	**Dynamic**
Frequency Bands	1	1	1	**11**
Frequency Range	✗	500 - 6400 Hz	✗	**1 Hz - 20 kHz**
Processed Beamforming Frames	✗	151	✗	**42250**
RGB Cameras	1	1	✗	**5**
RGB Frames	✗	151	✗	**3.2 Mio**
Lidar Point Clouds	✗	✗	✗	**480,000**
Annotated Frames	✗	✗	✗	**16,324**

- We introduce long-range acoustic beamforming of road noise as a *complementary sensing modality* for automotive perception, and introduce the first annotated long-range acoustic beamforming dataset comprising of sound measurements from planar microphone array, lidar, RGB images, GPS and IMU data, in urban driving scenarios.
- We propose neural acoustic beamforming for small aperture microphone arrays via learned aperture expansion. We validate that this beamforming approach can learn features with a spatial resolution that allows for fusion with existing RGB vision tasks.
- We validate that the proposed method complements existing modalities and outperforms existing RGB-only and audio-only detection methods in challenging scenarios with occlusion or poor lighting.

Scope As the proposed acoustic sensing modality relies on passive sound from traffic participants, beamforming measurements are fundamentally limited to sound-producing vehicles. Beamforming of quieter traffic participants such as pedestrians and bicycles is challenging. However, we show that infusing existing vision stacks with acoustic signals can enable robust scene understanding in challenging scenarios such as night scenes and under severe occlusion.

2. Related Work

Acoustic Localization and Applications Acoustic localization is an often observed phenomenon in nature. Active techniques like echolocation, where sound signals are transmitted and the corresponding reflected signals are analyzed for localization, navigation and prey detection is commonly observed in animals such as bats and dolphins. Systems such as sonar (sound navigation and ranging) [46] which are common for underwater and robotics applications also operate on the active echolocation principle [27, 43]. Passive techniques, on the other hand, involve analyzing ambient sound signals using an array of microphones via acoustic beamforming [6, 11–13, 32, 44]. Beamforming techniques locate sound sources based on the timing differences in the sound received by various microphones.

Apart from sound source localization, recent smart home speakers use several microphones for speech recognition accuracy from multi-channel inputs [40] and tasks such as sound source separation [37]. Existing attempts to locate sound sources from visual inputs by associating image pixels to an object [2–4, 22, 24, 28] making a particular sound

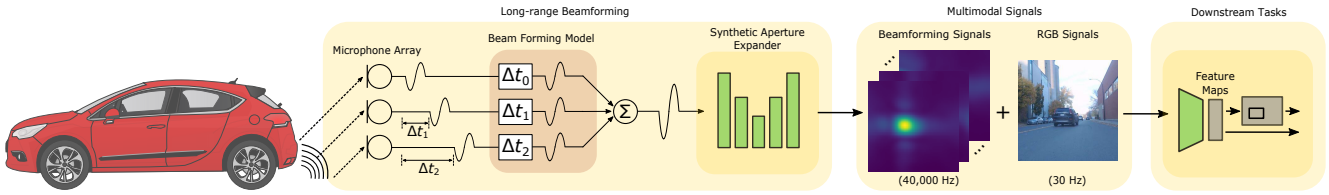


Figure 2. Overview of the pipeline: Roadside noise is measured by a microphone array sensor and a beamforming map of acoustic signals is computed as complementary modality to existing sensor stacks. A trained neural network translates multimodal signals to interpretable traffic scene information which can be used for downstream tasks such as object detection and predicting a future RGB camera frame.

often make such audio-visual correlations and localization via trained neural networks [17, 41, 49]. In contrast, we take a different approach where we learn high-resolution feature maps via beamforming that can be fused with visual inference models.

Beamforming Datasets and Applications While existing work employs stationary or hand-held microphone arrays for beamforming, we consider dynamic automotive scenarios for the first time where the array is mounted on a moving vehicle. For example, Michel et al. [35] and Zunino et al. [51] use a stationary array for beamforming on pass-by vehicles. However, their dataset contains only a single sequence with 151 beamforming maps of a motorcycle passing by. Guidati [21], on the other hand, used near-field acoustic holography to generate beamforming maps for engine noise analysis of a stationary car. Such applications have recently been also used in commercial products for fault detection applications. However, existing work has not proposed beamforming for dynamic automotive scenarios to the best of our knowledge. Additionally, we are not aware of any existing beamforming datasets with annotated data. Our annotated dataset, in stark contrast to existing datasets, comprises of 3.2 million RGB frames with 5 cameras, 480k lidar pointclouds with IMU/GNSS annotations, raw microphone files and 42250 processed beamforming maps for each of the 11 frequency bands that we capture.

Multisensor Detection Over the last decade, several critical tasks such as object detection [8, 9, 18, 26, 30, 47], lane detection [23], traffic light detection [25] depth estimation [1, 15, 29] and end-to-end driving models [5, 48] have been explored. Recent autonomous driving applications rely on multimodal sensor stacks, including camera, radar, lidar and gated near-infrared imaging sensors [19, 20]. These multi-sensor feeds are generally fused to jointly understand the cues in measurements [34] to allow for redundancy in the presence of distortions [38], thereby enabling vision tasks [14, 47]. Many proposed multi-sensor methods [9, 26, 31, 47] such as AVOD [26] and MV3D [9] incorporate multiple sensor streams that are processed separately in the feature extraction stage. In our work, we propose acoustic beamforming as an extended sensing modality to complement these existing methods. While researchers have utilized acoustic signals before [10, 16], this work employs acoustic beamforming to extract high-resolution spatial information from ambient roadside noise.

3. In-the-Wild Acoustic Beamforming

In this work, we measure environmental sound from ambient sources and active road participants using a planar microphone array, along with other sensor modalities, as shown in Fig. 1. We interchangeably refer to the environmental sound as acoustic signals here on. In this section, we introduce acoustic wave propagation and beamforming in-the-wild.

Acoustic Wave Propagation Sound propagation is governed by the time-domain acoustic wave equation

$$\nabla^2 p(\vec{x},t) - \frac{1}{c_s^2}\frac{\partial^2}{\partial t^2} p(\vec{x},t) = f(\vec{x},t), \tag{1}$$

where ∇^2 is the Laplacian, $p(\vec{x},t)$ is the pressure at location $\vec{x}$ and time t, c_s is the speed of sound in homogeneous media (typically $343\,\mathrm{ms}^{-1}$), and $f(\vec{x},t)$ is the forcing function corresponding to the source. The forcing function represents the sources of disturbances in the air pressure, i.e., the sound sources, as measured by the microphone sensor at a given space and time. For a monopole source q located at $\vec{x}_s$, the forcing function $f(\vec{x},t) = q(\vec{x}_s,t)\delta(\vec{x}-\vec{x}_s)$, where the Dirac delta function represents the geometric location of the acoustic source. The pressure resulting from this source at any given location $\vec{x}$ can be computed using the free space Green's function [6] as

$$p(\vec{x},t) = \frac{q(\vec{x}_s, t - |\vec{x}-\vec{x}_s|/c_s)}{4\pi|\vec{x}-\vec{x}_s|}. \tag{2}$$

Note that the acoustic pressure decays here inversely with the distance from the source. Also, since the acoustic pressure signal propagates at a constant speed c_s in a given medium, the measured pressure at any instant at a given location is from the acoustic pressure produced by the sound source at a previous instant $\Delta t = |\vec{x}-\vec{x}_s|/c_s$.

Beamforming Model Consider a planar microphone array consisting of M microphones that are spatially located at different positions $\vec{x}_m$. Given a pressure signal $p(\vec{x}_m,t)$ that

has originated from a source $q(\vec{x}_s, t - \Delta t)$, each sensor of the acoustic camera's microphone array spatially samples the incoming pressure wave as $\vec{y}_m = p(\vec{x}_m, t)$. We wish to use these measurements to construct a spatial map locating the sound source q via *beamforming*. For a single sound emitter at $\vec{x}_s$, the beamforming spatial map BF can be constructed following Eq. (2) as

$$\begin{aligned} \text{BF}(t, \vec{x}_s) &= \frac{1}{M} \sum_{m=0}^{M} y_m(t - \Delta t_m) \\ &= \frac{4\pi}{M} \sum_{m=0}^{M} p_m(\vec{x}_s, t + \Delta t_m) |\vec{x}_m - \vec{x}_s|, \end{aligned} \tag{3}$$

where y_m are measurements from the microphone array and Δt_m are unknown time delays induced by travel times from the sound source to the microphone array. The final beamforming maps of multiple sound sources is obtained by scanning through a range of time delays and superposing those acoustic signals corresponding to constructive interference of each individual sound source on the focal plane of the microphone array. Please refer to the Supplementary Material for additional details on the measurement model.

Measuring Environment Sounds Note that physical continuous acoustic pressure signals $p(t)$ are sampled at discrete time intervals $p(n\Delta t)$ and are interpreted digitally for the purpose of beamforming. However, the measured signals are prone to uncorrelated measurement noise at the array sensors. The measured cross-spectral power between any two microphone pairs, in the presence of measurement errors, is given by

$$C_{mn} = \mathbf{E}[(\tilde{p}_m(\omega) + \zeta_m(\omega))(\tilde{p}_m(\omega) + \zeta_m(\omega))^*], \tag{4}$$

where $\tilde{p}(\omega)$ is the frequency domain pressure obtained by Fourier-transforming the time domain measurement and $\zeta(\omega)$ is the measurement error. Assuming that these measurement errors have a zero mean and finite variance σ, and are statistically independent from the ambient acoustic signals, the cross-correlation between the errors as measured by any two microphones must be zero. Therefore, the above cross-power spectrum can be computed as

$$C_{mn} = \mathbf{E}[(\tilde{p}_m(f))(\tilde{p}_m(f))^*] + \sigma^2 \mathbf{I}, \tag{5}$$

where $\sigma^2 \mathbf{I}$ is the statistical variance of the measurement errors. As can be seen, the measurement errors only affect the diagonal elements of the cross-power spectrum matrix. To this end, we remove the auto-power from the beamforming power signal output by eliminating the diagonal of the cross-power spectrum matrix. Removing the main diagonal elements from the cross-spectral matrix reduces the effects of measurement errors and further thresholding against a noise floor suppresses ambient noise.

4. Neural Acoustic Beamforming

The diffraction limit of an acoustic camera is given by $0.5\lambda/\text{NA}$ where λ is the wavelength of the acoustic signal and NA is the numerical aperture of the system [33]. Therefore, a large aperture is desirable for achieving high-resolution beamforming that facilitates fusion with visual information from camera or lidar sensors. Fig. 4 shows the beamforming of traffic environment where the sound produced by the vehicle tires are clearly visualized. A small aperture acoustic camera results in larger PSFs, thereby corrupting the beamformed reconstruction. However, a large microphone array is challenging to integrate in automotive vehicles. In this work, we propose a learned method that synthesizes a virtual large aperture microphone array, thereby increasing the resolution of beamforming spatial maps. We experimentally show that these features from the beamforming maps of acoustic signals benefit downstream tasks when combined with other sensor modalities.

The proposed reconstruction network architecture broadly comprises of four stages: the beamforming stage f_{BF}, a synthetic aperture expander f_{AE}, a deconvolution stage f_{Deconv}, and task-specific applications f_{Task}. Our overall neural beamforming can be formally represented as

$$O_{\text{BF}} = f_{\text{Deconv}}\big(f_{\text{AE}}(f_{\text{BF}}(p, \mathbf{F})), f_{\text{BF}}(\delta, \mathbf{F}))\big), \tag{6}$$

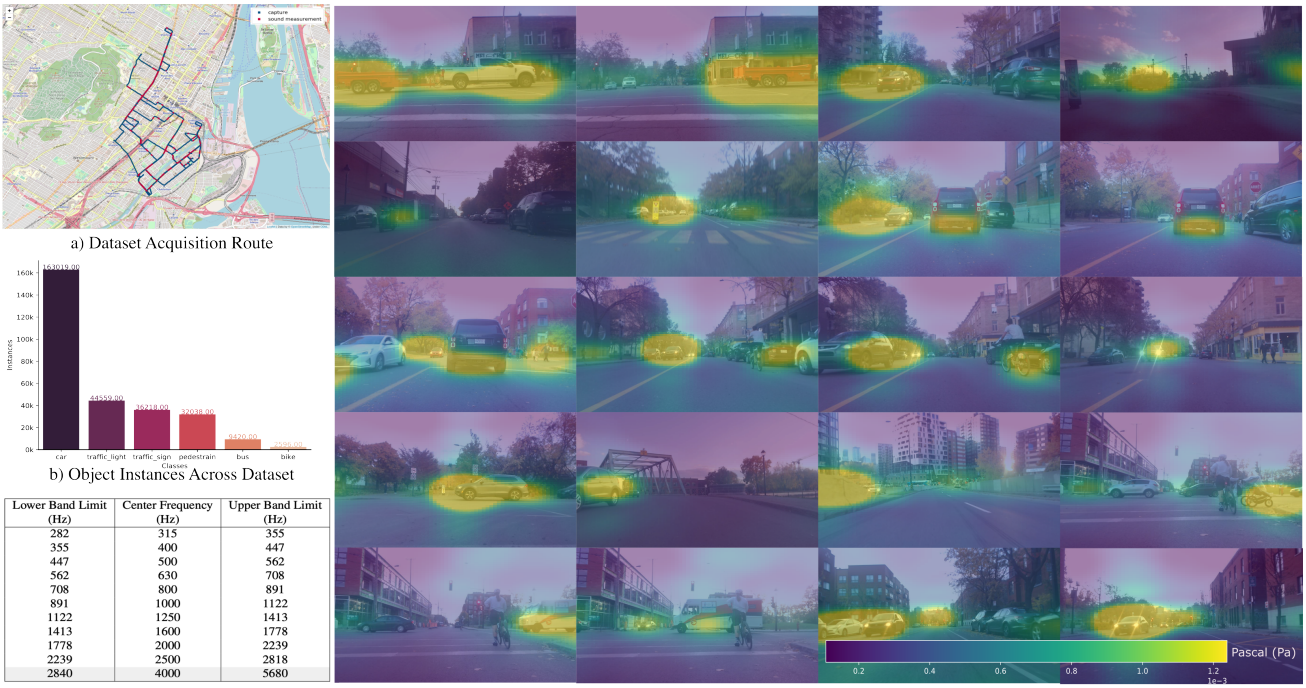
where δ is a synthetic audio point source, p is the raw microphone measurement of the pressure signals, $\mathbf{F} = [f_1, f_2, ..., f_n]$ are a set of acoustic frequencies used for beamforming. An illustration of our network architecture is also presented in Fig. 2.

The aperture expander is constructed as a fully convolutional neural network, whereas the beamforming stage is implemented as described in Section 3. The synthetic aperture expander network learns to scale the beamforming maps corresponding to a smaller aperture into that of a larger aperture, thereby effectively reducing the PSF of our acoustic sensor. The beamforming measurements are then deconvolved with the PSF of a synthetic point source δ in order to mitigate the PSF blur on final measurements, see Supplementary Material. Finally, the deconvolved features O_{BF} can be used directly for downstream tasks such as object detection and future frame interpolation. Specifically, the downstream task can be performed as

$$O_{\text{Task}} = f_{\text{Task}}(O_{\text{BF}}), \tag{7}$$

where f_{Task} is the function performing the downstream task and O_{Task} is the corresponding task-specific output. In the subsequent sections, we describe how these beamforming features can be used for object detection on unseen in-the-wild traffic scenarios ($f_{\text{Task}} = f_{\text{detect}}$) and future frame prediction ($f_{\text{Task}} = f_{\text{future}}$).

Aperture Expansion We define the beamforming map of a microphone array spanning $d \times d$ m^2 as $I_d = f_{\text{BF}}(p_d, \mathbf{F})$.



a) Dataset Acquisition Route

b) Object Instances Across Dataset

Lower Band Limit (Hz)	Center Frequency (Hz)	Upper Band Limit (Hz)
282	315	355
355	400	447
447	500	562
562	630	708
708	800	891
891	1000	1122
1122	1250	1413
1413	1600	1778
1778	2000	2239
2239	2500	2818
2840	4000	5680

c) Processed Frequency Bands

d) Beamforming Map Projections on Camera Images

Figure 3. (a) We collect a large dataset of acoustic pressure signals from road traffic, along with camera, lidar, GPS and IMU sensory streams. (b) Our data consists of a variety of traffic scenes with multiple object instances. (c) The raw sound signals are processed for a range of frequency bands in our dataset. (d) A snapshot visualization of our dataset with beamforming maps overlayed on RGB images.

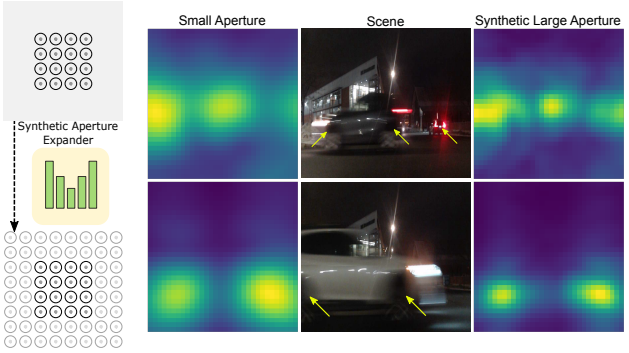


Figure 4. Neural aperture expansion: We train a network to synthetically expand the aperture of the microphone array to produce higher fidelity beamform maps with smaller PSF corruption. The above beamforming maps visualize sound from the vehicles (center, see arrows). As can be observed, a smaller aperture (left) results in blurry beamforming outputs compared to larger aperture (right).

We then train our aperture expansion stage by minimizing

$$\mathcal{L}_{\text{AE}} = (\mathcal{L}_2 + \mathcal{L}_s)(f_{\text{AE}}(I_d), I_{d'}), \tag{8}$$

where $\mathcal{L}_2$ is the mean-squared error, $\mathcal{L}_s$ is a spatial gradient loss, I_d is a smaller aperture beamforming input, and $I_{d'}$ is a larger aperture beamforming target. In our experiments, we collect groundtruth from a 32×32 microphone array and train our neural aperture expansion on smaller 24×24 sub-array.

Multimodal Downstream Tasks Next, we describe how we apply our complementary beamforming signals to specific vision tasks. We describe both multimodal tasks in detail in Section 6. The multimodal optimization loss, relying on visual and acoustic inputs for object detection ($f_{\text{Task}} = f_{\text{detect}}$) is given by

$$\mathcal{L}_{\text{detect}} = \mathcal{L}_{\text{IoU}}(f_{\text{detect}}(I_{\text{RGB}}, O_{\text{BF}}), B_{\text{gt}}), \tag{9}$$

where $\mathcal{L}_{\text{IoU}}$ is the intersection-over-union loss and B_{gt} is the ground truth bounding box.

For future frame prediction, we extrapolate from a previous RGB frame I^t_{RGB} at time t using signals $O^{t+k\tau}_{\text{BF}}$, where k is current BF sample modulo sampling rate, τ is sampling time of BF sensor, and $t + k\tau$ the current time. The corresponding loss for this task is given by

$$\mathcal{L}_{\text{future}} = \big(\mathcal{L}_{\text{perc}} + \mathcal{L}_{\text{adv}}\big)\big(O^{t+k\tau}_{\text{RGB}}, I^{t+k\tau}_{\text{RGB}}\big), \tag{10}$$

with $O^{t+k\tau}_{\text{RGB}} = f_{\text{future}}(O^{t-n+1,\ldots,t,t+k\tau}_{\text{BF}}, I^{t-n+1,\ldots,t}_{\text{RGB}})$, n is integer time steps, $\mathcal{L}_{\text{perc}}$ and $\mathcal{L}_{\text{adv}}$ are perceptual and adversarial losses respectively [45]. In this approach, we exploit the high framerate of acoustic measurements. Specifically, we feed n RGB and $n+1$ audio frames into the network and train it to predict the $n+1$-th RGB frame.

In the following Section 5, we describe our prototype test vehicle and dataset. We then discuss and validate our method on the aforementioned downstream tasks in Section 6.

5. Dataset

To assess long-range acoustic beamforming for automotive scene understanding, we have acquired a large dataset of roadside noise along with ambient scene information captured by an RGB camera, lidar, global positioning system (GPS) and an inertial measurement unit (IMU).

5.1. Sensors and Data Acquisition

In this section we present details on our dataset collection. To the best of our knowledge, we provide a *far-field multimodal acoustic beamforming dataset*.

Sensor Configuration We equipped a prototype vehicle with RGB cameras, lidar, IMU/GNSS, and a microphone array for beamforming, as shown in Fig. 1. Specifically, the prototype vehicle has the following sensor configuration.

- One Sorama CAM1K 1024 channel microphone array operating at 46875 Hz sampling rate, 1 Hz - 20 kHz frequency range, covering $640\,\mathrm{mm} \times 640\,\mathrm{mm}$ measurement area. Each microphone has 63 dB SNR (A-weighted, at 1 KHz), -26 dBFS sensitivity, and 116 dB acoustic overload point. The microphone channels are arranged in a 32×32 grid with a $20\,\mathrm{mm}$ grid spacing. For far field beamforming, the sensor's large surface area of $409\,600\,\mathrm{mm}^2$ enables operations on larger wavelengths and therefore measurement of lower frequency sources, whereas the microphone grid spacing dictates the upper frequency bounds. Our capture system is configured to sense frequencies as low as 250 Hz and as high as 10kHz emanating from ambient sources.
- Logitech C920 RGB camera operating at 1280×720 resolution, 25 Hz frame rate, 70.42°HFoV and 43.3°VFoV.
- Four Leopard Imaging LI-AR0231 GMSL serial cameras with 1920×1200 resolution, 30 Hz frame rate, 1/2.7" OnSemi AR0231 CMOS, rolling shutter and 60°HFoV
- Hesai Pandar 64 channel lidar operating at 20 Hz, 360°HFoV, 40°VFoV, covering a $200\,\mathrm{m}$ range and 0.4°angular resolution.
- Novatel PwrPak7-ED1 GNSS 20 Hz dual antenna navigation system, GPS/GLONASS/Galileo/BeiDou.

The microphone array is mounted on a rail attached to the front bumper, while the C920 camera is co-planar with, and mounted $36\,\mathrm{cm}$ below the array center on the same frame. This minimizes the projection errors of beamformed maps on the image caused by mount vibrations. The four AR0231 cameras are mounted on the roof and face, along with the lidar and the dual antennas for the GNSS navigation system. The four AR0231 cameras are mounted on roof rails in a dual stereo camera configuration of two different baselines. The PwrPak7 receiver unit which houses the IMU and the GNSS module serves as the car coordinate frame's origin, and is mounted in the trunk above the rear-axle mid-point. For a description of calibration and synchronization of sensors, please see Supplementary Material.

Acquisition Focused on urban scenes, the data acquisition took place in an urban northern American city, as shown in Fig. 3(a). The dataset spans $66\,\mathrm{km}$ of urban roads, amounting to 14 TB of storage. 2.8 million images were collected at 30 Hz by the four AR0231 cameras. The C920 camera was enabled for capturing only during the sound measurements by the microphone array and totalled 42250 images at 25 Hz. 480240 64-channel lidar point clouds were recorded at 20 Hz. All measurements were time-stamped and synchronized with GNSS as time reference.

The microphone array signals were recorded at 10 second intervals at a sampling rate of 46.875 kHz. All acoustic captures are highlighted in Fig. 3(a) in red. To the extent possible, the vehicle speed was kept constant between 30 km/h to 40 km/h, to minimize the effect of high winds on the sound readings. 79.2 million samples were collected from each of the 1024 microphones, resulting in more than 81.1 billion sound pressure samples in the dataset. Please see Supplementary Material for additional details on the diversity and distribution of our dataset.

5.2. Ground Truth Annotations

Manual annotations were done for visual and sound classes on data sampled at 5 Hz, for a total of 16324 keyframes, 11 sound classes and 6 vision classes. In addition to image class labels, each sampled image was also annotated with sound labels in two domains: *dominant* (distinct and in foreground) and *secondary* (in the background). All labels were created by highly experienced human annotators using a custom toolset. Their work passed through subsequent phases of verification and quality assurance to ensure high-quality labels. All object instances were annotated using tightly fitted 2D bounding boxes aligned to image axis, and encoded as top left and bottom right coordinates in the image frame. Please see Supplementary Material for details on annotations.

6. Applications

We demonstrate that acoustic beamforming, when combined with RGB data, can allow for multimodal scene understanding tasks and future frame prediction better than using existing RGB-only or acoustic-only methods. Note that our training and test video frames come from *entirely different* video sequences. Instead of holding out frames from the same sequences for splitting train and test sets, e.g., as in Stereo-sound by Gan et al. [16], our evaluation is conducted on *completely unseen* sequence frames. Multimodal inputs to the network consisted of concatenated vision and audio beamforming signals.

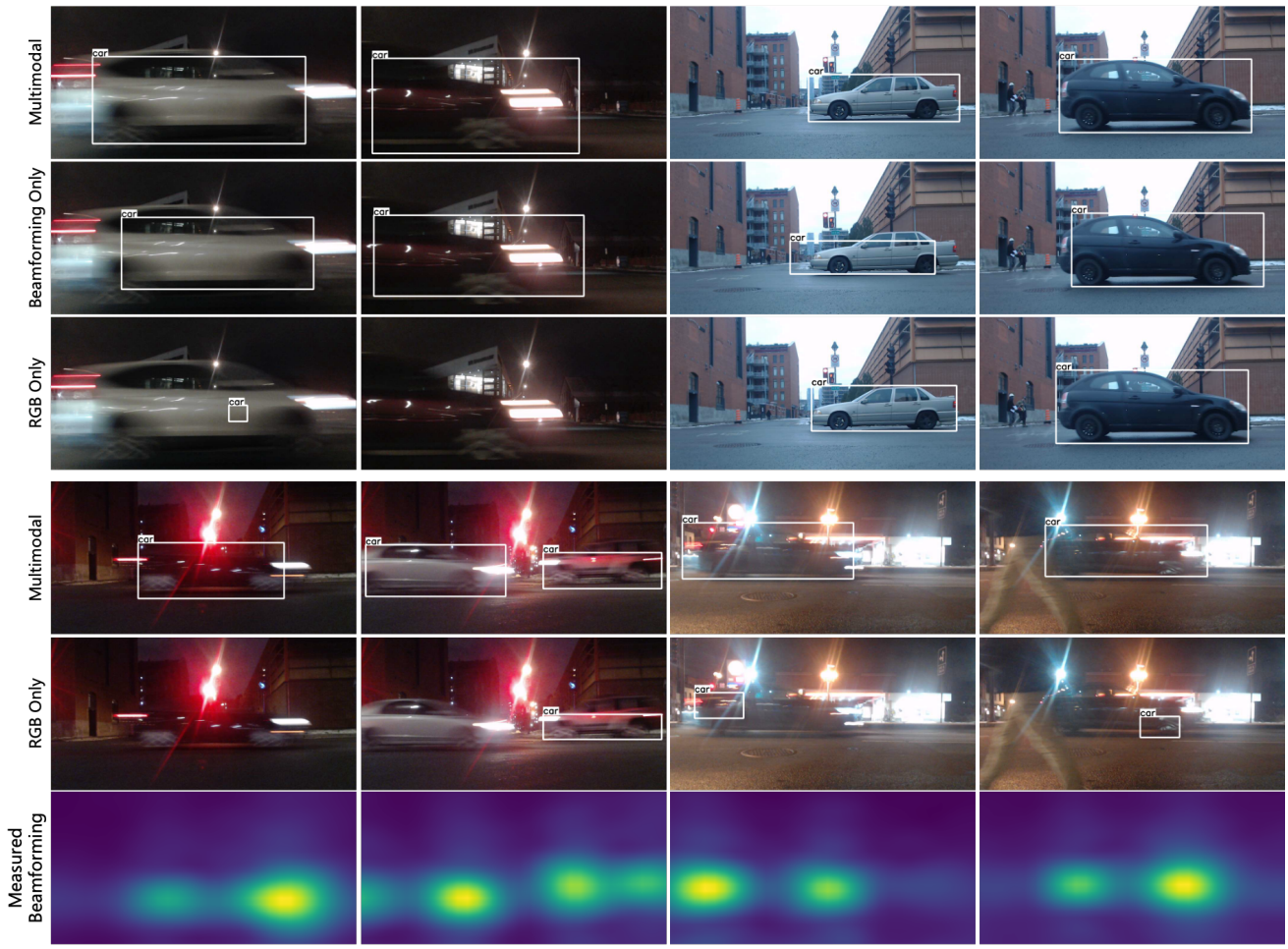


Figure 5. Complementary nature of RGB and acoustic signals. Traffic sound provides complementary signals that improve detections significantly on challenging scenarios where low-light, glare, and motion blur reduce the accuracy of RGB detections. Furthermore, beamforming allows us to reliably detect multiple sound source in the scene.

Table 2. The proposed detection method that utilizes beamform features in addition to RGB features outperform detections using audio spectrograms and RGB only. All evaluations were performed using YOLOv5 [39] with fine-tuning, see Supplementary Material for details. BF denotes Beamforming and NAE denotes Neural Aperture Expansion. $AP_{50}(D)$ and $AP_{50}(N)$ is the Average Precision scores for Day and Night scenes respectively.

Method	Detector	$AP_{50}(D)$	$AP_{50}(N)$	AP_{Ave}
RGB + BF (NAE)	Fine-tuned	**81.2**	**64.3**	**29.5**
RGB + BF	Fine-tuned	80.7	61.8	28.3
RGB-only	Fine-tuned	79.4	37.2	18.1
BF-only	Fine-tuned	62.5	61.1	21.3
Stereo-sound [16]	Fine-tuned	0	0	0

6.1. Object Detection on Sound and RGB Streams

We validate the proposed method for multimodal automotive object detection using RGB and acoustic signals, that is $f_{\text{Task}} = f_{\text{detection}}$. We employ a YOLOv5 [39] detection network for all experiments shown in Table 2. Fine-tuning the detector on concatenated image and beamforming maps allowed us to achieve high performance on challenging scenes, where low-light, motion blur, and glare confound the RGB detector. In contrast to small aperture measurements, beamforming signals via neural aperture expansion (NAE) showed improved performance as shown in Table 2. Compared to Gan et al. [16], we observe that the stereo-sound signals provided by spectrograms are insufficient for accurate detection with vanilla detector networks. We attribute this to the spatial cues provided by the beamforming signal maps. Please see the Supplementary Material for additional qualitative object detection results and comparisons.

We present object detection on unseen scenes using the proposed multimodal approach in Fig. 1 and 5 and demonstrate significantly improved detections compared to RGB-only methods. To demonstrate the complementary nature of the acoustic signals, we also present the measured beam-

Table 3. Frame Interpolation and Future Frame Prediction. We evaluate the utility of beamform maps for predicting $t = 1, 2, 3$ RGB frames into the future. We observe that the audio maps provide temporal context cues that enable more accurate future predictions than RGB frames or direct extrapolation of optical flow from previous frames.

	PSNR (dB)		
Future frame prediction	$t+1$	$t+2$	$t+3$
Beamforming + RGB	**28.56**	**27.47**	**22.94**
RGB only	27.62	25.57	21.50
Optical Flow Extrapolation	23.18	21.45	18.90
Last RGB Frame	23.06	21.31	18.85

forming signal maps in the last row of Fig. 5. Note that these video sequences were not used for training. Whereas detection on RGB-only frames need to contend with variable environmental factors such as glare and low-lighting, the accompanying beamforming maps demonstrate consistent complementary signals for automotive detection whether in night or day, allowing for superior detection performance.

6.2. Multimodal Future Frame Prediction

We also demonstrate that beamforming maps provide useful context cues for predicting future RGB frames from RGB streams with low temporal resolution, $f_{\text{Task}} = f_{\text{future}}$. Given the 46 kHz sampling rate of our acoustic capture system, we are able to extrapolate previous RGB frames at the same ultra-fast update rate using a temporal sliding window of beamforming, despite the RGB camera operating only at 30 Hz. Note that we do not access future beamforming maps. For this task, we train a modified Pix2PixHD network [45] to take temporal information of both RGB images and beamforming maps, essentially implementing Eq. (10). We show in Table 3 that incorporating audio cues improves future frame prediction over RGB-only extrapolation. We also compare against predictions using extrapolated optical flow and we demonstrate significant improvement. In order to predict several frames into the future, we *cascade* predictions by using previously predicted RGB frames and the corresponding measured audio inputs, following Eq. (10).

6.3. Multimodal Edge Cases

Non-line-of-sight and Partial Occlusion Scenarios Next, we further validate the complementary nature of acoustic sensors to photon-based sensors for the detection of an oncoming object that is not directly visible to an RGB or lidar sensor, e.g., hidden behind an opaque wall. Fig. 6(a)(top) shows an example of such a non-line-of-sight detection in our test dataset where a car not visible in the RGB camera view except for a thin roof region is detected robustly by the acoustic sensor. Similarly, Fig. 6(a)(bottom) reports an edge case of a low light scenario where a vehicle is partially occluded by a pedestrian but is robustly detected with acoustic features present. This validates that multi-modal acoustic sensing can allow for new redundancies, where acoustic sensors provide information that the other sensors cannot.

Figure 6. (a) Scenes with oncoming vehicles (arrows) being fully (top) or partially (bottom) occluded in the RGB frame but detected by the acoustic sensor (center column). (b) Edge case with loud vehicle exceeding the dynamic range, corrupting beamforming.

Acoustic Edge Cases In the following, we discuss acoustic edge cases where pressure signals can saturate the microphones. As discussed in Section 3, we suppress ambient noise by thresholding against a noise floor. The employed microphones have a high dynamic range of -26 dBFS +/- 1.5dB (94 dB SPL at 1kHz), which covers roadside sound from quiet electrical vehicles to large passenger trucks. Similarly, as the capture setup is mounted in front of the vehicle, beamforming also removes the ego-vehicle components. However, very large construction vehicles in close proximity can exceed 100 dB SPL in close proximity and saturate the microphones resulting in errors in detection and tracking, as shown in Fig. 6(b), where sound exceeds the acoustic overload point.

7. Conclusion

We introduce a method for learning from acoustic microphone arrays and interpret roadway traffic noise as a complementary sensing modality for automotive imaging and scene understanding. When combined with camera data, we validate that this sensing modality provides complementary information that facilitates detection within challenging environmental conditions. To train and evaluate the proposed method, we capture an automotive acoustic dataset. We envision researchers developing and evaluating multimodal methods incorporating acoustic information from roadside noise, in addition to existing sensory data. Motivated by the complementary nature and multi-modal experiments in this paper, areas for future work include extending our approach to arbitrary and optimizable camera and microphone array system geometries.

Acknowledgments Felix Heide was supported by an NSF CAREER Award (2047359), a Packard Foundation Fellowship, a Sloan Research Fellowship, a Sony Young Faculty Award, a Project X Innovation Award, and an Amazon Science Research Award.

References

[1] Unsupervised monocular depth estimation with left-right consistency. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, July 2017. 3

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018. 2

[3] Yutong Ban, Xavier Alameda-Pineda, Laurent Girin, and Radu Horaud. Variational bayesian inference for audio-visual tracking of multiple speakers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2

[4] Zohar Barzelay and Yoav Y Schechner. Harmony in motion. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 2

[5] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016. 3

[6] Michael Brandstein and Darren Ward. *Microphone arrays: signal processing techniques and applications*. Springer Science & Business Media, 2013. 2, 3

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 1

[8] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing*, 25(11):5187–5198, 2016. 3

[9] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 3

[10] Ziyang Chen, Xixi Hu, and Andrew Owens. Structure from silence: Learning scene structure from ambient sound. In *5th Annual Conference on Robot Learning*, 2021. 3

[11] M. Cobos, M. García-Pineda, and M. Arevalillo-Herráez. Steered response power localization of acoustic passband signals. *IEEE Signal Processing Letters*, 24:717–721, 2017. 1, 2

[12] Joseph H. DiBiase, H. Silverman, and M. Brandstein. Robust localization in reverberant rooms. In *Microphone Arrays*, 2001. 1, 2

[13] Jacek P. Dmochowski and J. Benesty. Steered beamforming approaches for acoustic source localization. 2010. 1, 2

[14] Dmitri Dolgov, Sebastian Thrun, Michael Montemerlo, and James Diebel. Path planning for autonomous vehicles in unknown semi-structured environments. *The International Journal of Robotics Research*, 29(5):485–501, 2010. 3

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 3

[16] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 3, 6, 7

[17] Ruohan Gao and Kristen Grauman. 2.5 d visual sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2019. 3

[18] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[19] Yoav Grauer. Active gated imaging in driver assistance system. *Advanced Optical Technologies*, 3(2):151–160, 2014. 3

[20] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time dense lidar from gated images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1506–1516, 2019. 3

[21] Sandro Guidati. Advanced beamforming techniques in vehicle acoustic. In *Berlin Beamforming Conference (BeBeC)*, 2010. 2, 3

[22] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in neural information processing systems*, pages 813–819, 2000. 2

[23] Aharon Bar Hillel, Ronen Lerner, Dan Levi, and Guy Raz. Recent progress in road and lane detection: a survey. *Machine vision and applications*, 25(3):727–745, 2014. 3

[24] Hamid Izadinia, Imran Saleemi, and Mubarak Shah. Multimodal analysis for identification and segmentation of moving-sounding objects. *IEEE Transactions on Multimedia*, 15(2):378–390, 2012. 2

[25] Morten Bornø Jensen, Mark Philip Philipsen, Andreas Møgelmose, Thomas Baltzer Moeslund, and Mohan Manubhai Trivedi. Vision for looking at traffic lights: Issues, survey, and perspectives. *IEEE Transactions on Intelligent Transportation Systems*, 17(7):1800–1815, 2016. 3

[26] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018. 3

[27] John J Leonard and Hugh F Durrant-Whyte. *Directed sonar sensing for mobile robot navigation*, volume 175. Springer Science & Business Media, 2012. 2

[28] Xiaofei Li, Yutong Ban, Laurent Girin, Xavier Alameda-Pineda, and Radu Horaud. Online localization and tracking of multiple moving speakers in reverberant environments. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):88–103, 2019. 2

[29] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2015. 3

[30] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg.

Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3

[31] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3569–3577, 2018. 3

[32] A. Marti, M. Cobos, J. López, and José Escolano. A steered response power iterative method for high-accuracy acoustic source localization. *The Journal of the Acoustical Society of America*, 134 4:2627–30, 2013. 1, 2

[33] AA Maznev and OB Wright. Upholding the diffraction limit in the focusing of light and sound. *Wave Motion*, 68:182–189, 2017. 4

[34] Oier Mees, Andreas Eitel, and Wolfram Burgard. Choosing smartly: Adaptive multimodal fusion for object detection in changing environments. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 151–156. IEEE, 2016. 3

[35] Ulf Michel, Bernd Barsikow, Peer Böhning, and Michael Hellmig. Localisation of sound sources on moving vehicles with phased microphone arrays. *Proceedings of the Inter-Noise 2004*, pages 22–25, 2004. 2, 3

[36] Danijela M. Miloradović, Jasna Glišović, and Jovanka Lukić. Regulations on road vehicle noise – trends and future activities. 2017. 1

[37] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent. Multichannel audio source separation with deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1652–1664, 2016. 2

[38] Cristiano Premebida, Oswaldo Ludwig, and Urbano Nunes. Lidar and vision-based pedestrian detection system. *Journal of Field Robotics*, 26(9):696–711, 2009. 3

[39] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 7

[40] Tara N Sainath, Ron J Weiss, Kevin W Wilson, Bo Li, Arun Narayanan, Ehsan Variani, Michiel Bacchiani, Izhak Shafran, Andrew Senior, Kean Chin, et al. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(5):965–979, 2017. 2

[41] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 3

[42] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 1

[43] Santani Teng, Verena R Sommer, Dimitrios Pantazis, and Aude Oliva. Hearing scenes: a neuromagnetic signature of auditory source and reverberant space separation. *Eneuro*, 4(1), 2017. 2

[44] Juan Manuel Vera-Diaz, Daniel Pizarro-Perez, and Javier Macias-Guarasa. Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates. *Sensors (Basel, Switzerland)*, 18, 2018. 1, 2

[45] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018. 5, 8

[46] Alan A Winder. Ii. sonar system technology. *IEEE Transactions on Sonics and Ultrasonics*, 22(5):291–332, 1975. 2

[47] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2018. 3

[48] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2174–2182, 2017. 3

[49] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 3

[50] Julius Ziegler, Philipp Bender, Markus Schreiber, Henning Lategahn, Tobias Strauss, Christoph Stiller, Thao Dang, Uwe Franke, Nils Appenrodt, Christoph G Keller, et al. Making bertha drive—an autonomous journey on a historic route. *IEEE Intelligent transportation systems magazine*, 6(2):8–20, 2014. 1

[51] Andrea Zunino, Marco Crocco, Samuele Martelli, Andrea Trucco, Alessio Del Bue, and Vittorio Murino. Seeing the sound: A new multimodal imaging device for computer vision. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 6–14, 2015. 2, 3