Semantic Object Prediction and Spatial Sound Super-Resolution with Binaural Sounds

Arun Balajee Vasudevan[1], Dengxin Dai[1], and Luc Van Gool[1,2]

[1] Computer Vision Lab, ETH Zurich
[2] KU Leuven
{arunv,dai,vangool}@vision.ee.ethz.ch

Abstract. Humans can robustly recognize and localize objects by integrating visual and auditory cues. While machines are able to do the same now with images, less work has been done with sounds. This work develops an approach for dense semantic labelling of sound-making objects, purely based on binaural sounds. We propose a novel sensor setup and record a new audio-visual dataset of street scenes with eight professional binaural microphones and a 360°camera. The co-existence of visual and audio cues is leveraged for supervision transfer. In particular, we employ a cross-modal distillation framework that consists of a vision 'teacher' method and a sound 'student' method – the student method is trained to generate the same results as the teacher method. This way, the auditory system can be trained without using human annotations. We also propose two auxiliary tasks namely, a) a novel task on Spatial Sound Super-resolution to increase the spatial resolution of sounds, and b) dense depth prediction of the scene. We then formulate the three tasks into one end-to-end trainable multi-tasking network aiming to boost the overall performance. Experimental results on the dataset show that 1) our method achieves promising results for semantic prediction and the two auxiliary tasks; and 2) the three tasks are mutually beneficial – training them together achieves the best performance and 3) the number and orientations of microphones are both important. The data and code will be released to facilitate the research in this new direction. Please refer to our project page

1 Introduction

Autonomous vehicles and other intelligent robots will have a substantial impact on people's daily life, both personally and professionally. While great progress has been made in the past years with visual perception systems [14, 26, 47], we argue that auditory perception and sound processing also play a crucial role in this context [32]. As known, animals such as bats, dolphins, and some birds have specialized on "hearing" their environment. To some extent, humans are able to do the same – to "hear" the shape, distance, and density of objects around us [39]. In fact, humans surely need to use this capability for many daily activities such as for driving – certain alerting stimuli, such as horns of cars and sirens of ambulances, police cars and fire trucks, are meant to be heard, i.e. are

(a) binaural sounds (b) super-resolved ones (c) semantic labels (d) depth values

Fig. 1. An illustration of our three tasks: (a) input binaural sounds, (b) super-resolved binaural sounds – from azimuth angle 0° to 90°, (c) auditory semantic perception for three sound-making object classes, and (d) auditory depth perception.

primarily acoustic [32]. Using multi-sensory information is a fundamental capability that allows humans to interact with the physical environment efficiently and robustly [2, 16, 18]. Future intelligent robots are expected to have the same perception capability to be robust and to be able to interact with humans naturally. Furthermore, auditory perception can be used to localize common objects like a running car, which is especially useful when visual perception fails due to adverse visual conditions or occlusions.

Numerous interesting tasks have been defined at the intersection of visual and auditory sensing like sound source localization in images [52], scene-aware audio generation for VR [31], geometry estimation for rooms using sound echos [4], sound source separation using videos [20], and scene [8] and object [19] recognition using audio cues. There are also works to learn the correlation of visual objects and sounds [51, 52]. While great achievements have been made, previous methods mostly focus on specific objects, e.g. musical instruments or noise-free rooms, or on individual tasks only, e.g. sound localization or geometry estimation. This work aims to learn auditory semantic prediction, for general, sound-making objects like cars, trains, and motorcycles in unconstrained, noisy environments.

This work primarily focuses on semantic object prediction based on binaural sounds aiming to replicate human auditory capabilities. To enhance the semantic prediction, this work proposes two auxiliary tasks: depth prediction from binaural sounds and spatial sound super-resolution (S^3R). Studies [34, 49] have shown that depth estimation and semantic prediction are correlated and are mutually beneficial. S^3R is a novel task aiming to increase the directional resolution of audio signals, e.g. from Stereo Audio to Surround Audio. S^3R, as an auxiliary task, is motivated from the studies [33, 45] showing that humans are able to better localize the sounding sources by changing their head orientations. S^3R is also a standalone contribution and has its own applications. For instance, spatially resolved sounds improve the spatial hearing effects in AR/VR applications. It offers better environmental perception for users and reduces the ambiguities of sound source localization [27].

In particular, we propose a sensor setup containing eight professional binaural microphones and a 360° camera. We used it to record the new 'Omni Auditory Perception Dataset' on public streets. The semantic labels of the sound-making objects and the depth of the scene are inferred from the video frames using

well-established vision systems as shown in Fig. 1. The co-existence of visual and audio cues is leveraged for supervision transfer to train the auditory perception systems without using human annotations. In particular, we employ a cross-modal framework that consists of vision teacher methods and sound student methods to let the students imitate the performance of the teachers. For evaluation, we have manually annotated a test set. The task of S^3R has accurate ground truth to train, thanks to our multi-microphone rig. Finally, we formulate the semantic prediction task and the two auxiliary tasks into a multi-tasking network which can be trained in an end-to-end fashion.

We evaluate our method on our new Omni Auditory Perception Dataset. Extensive experimental results reveal that 1) our method achieves good results for auditory semantic perception, auditory depth prediction and spatial sound super-resolution; 2) the three tasks are mutually beneficial – training them together achieves the best results; and 3) both the number and orientations of microphones are important for auditory perception.

This work makes multiple contributions: 1) a novel approach to dense semantic label prediction for sound-making objects of multiple classes and an approach for dense depth prediction; 2) a method for spatial sound super-resolution, which is novel both as a standalone task and as an auxiliary task for auditory semantic prediction; and 3) a new Omni Auditory Perception Dataset with four pairs of binaural sounds (360° coverage) accompanied by synchronized 360° videos which will be made publicly available.

2 Related Works

Auditory Scene Analysis. Sound segregation is a well-established research field aiming to organize sound into perceptually meaningful elements [1]. Notable applications include background sounds suppression and speech recognition. Recent research has found that the motion from videos can be useful for the task [20, 51]. Sound localization has been well studied with applications such as localizing sniper fire on the battle field, cataloging wildlife in rural areas, and localizing noise pollution sources in an urban environment. It also enriches human–robot interaction by complementing the robot's perceptual capabilities [7, 38]. The task is often tackled by the beamforming technique with a microphone array [46], with a notable exception that relies on a single microphone only [42]. The recent advance of deep learning enables acoustic camera systems for real-time reconstruction of acoustic camera spherical maps [44].

Auditory scene analysis has been widely applied to automotive applications as well. For instance, auditory cues are used to determine the occurrence of abnormal events in driving scenarios [32]. An acoustic safety emergency system has also been proposed [17]. Salamon et al. have presented a taxonomy of urban sounds and a new dataset, UrbanSound, for automatic urban sound classification [41]. The closest to our work is the very recent work of car detection with stereo sounds [19] in which a 2D bounding-box is proposed for the sound-making object in an image frame. While being similar in spirit, our work differs signifi-

cantly from theirs. Our method is designed for dense label prediction for multiple classes instead of a 2D bounding box prediction for a single class. Our method also includes dense depth prediction and spatial sound super-resolution. Binaural sounds are different from general stereo sounds and our method works with panoramic images to have omni-view perception. Another similar work to ours is [28]. We differ significantly in: 1) having multiple semantic classes, 2) working in unconstrained real outdoor environment, and 3) a multi-task learning setup.

Audio-Visual Learning. There is a body of work to localize the sources of sounds in visual scenes [5, 6, 9, 20, 22, 43, 51, 52]. The localization is mostly done by analyzing the consistency between visual motion cues and audio motion cues over a large collection of video data. These methods generally learn to locate image regions which produce sounds and separate the input sounds into a set of components that represent the sound for each pixel. Our work uses binaural sounds rather than monaural sounds. Our work localizes and recognizes sounds at the pixel level in image frames, but performs the task with sounds as the only inputs. Audio has also been used for estimating the geometric layout of indoor scenes [4, 15, 29, 50]. The general idea is that the temporal relationships between the arrival times of echoes allows us to estimate the geometry of the environment. There are also works to add scene-aware spatial audio to 360° videos in typical indoor scenes [31] by using a conventional mono-channel microphone and a speaker and by analyzing the structure of the scene. The notable work by Owens et al. [36] have shown that the sounds of striking an object can be learned based on the visual appearance of the objects. By training a neural network to predict whether video frames and audio are temporally aligned [35], audio representations can be learned. Our S^3R can be used as another self-learning method for audio representations.

Cross-domain Distillation. The interplay among senses is basic to the sensory organization in human brains [18] and is the key to understand the complex interaction of the physical world. Fortunately, most videos like those available in the Internet contain both visual information and audio information, which provide a bridge linking those two domains and enable many interesting learning approaches. Aytar et al. propose an audio-based scene recognition method by cross-domain distillation to transfer supervision from the visual domain to the audio domain [8]. A similar system was proposed for emotion recognition [3]. Ambient sounds can provide supervision for visual learning as well [37].

3 Approach

3.1 Omni Auditory Perception Dataset

Training our method requires a large collection of omni-directional binaural audios and accompanying 360° videos. Since no public video dataset fulfills this requirement, we collect a new dataset with a custom rig. As shown in Fig. 2, we assembled a rig consisting of a 3Dio Omni Binaural Microphone, a 360° GoPro Fusion camera and a Zoom F8 MultiTrack Field Recorder, all attached to a tripod. We mounted the GoPro camera on top of the 3Dio Omni Binaural

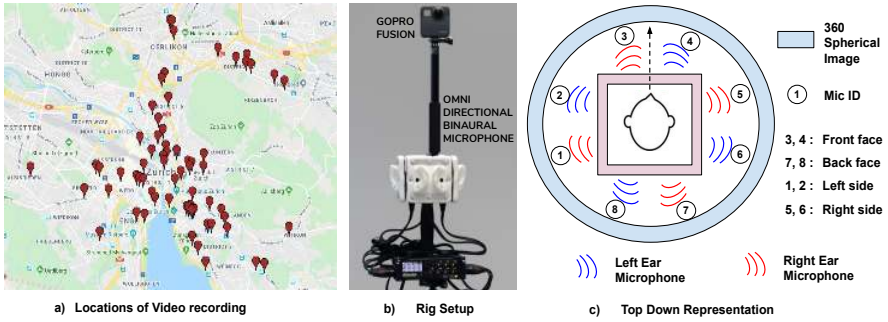


Fig. 2. Sensor and dataset: a) data capture locations, b) our custom rig and c) abstract depiction of our recording setup with sensor orientations and microphone ids.

Microphone by a telescopic pole to capture all of the sights with minimum occlusion from the devices underneath. This custom rig enables omni seeing and omni binaural hearing with 4 pairs of human ears. The 8 microphones are connected to 8 tracks of the MultiTrack Recorder. The recorder has 8 high-quality, super low-noise mic preamps to amplify the sounds, provides accurate control of microphone sensitivity and allows for accurate synchronization among all microphones. The GoPro Fusion camera captures 4K videos using a pair of 180° cameras housed on the front and back faces. The two cameras perform synchronized image capture and are later fused together to render 360° videos using the GoPro Fusion App. We further use the clap sound to synchronize the camera and 3Dio microphones. After the sensors are turned on, we do hand clapping near them. This clapping sounds recorded by both the binaural mics and the built-in mic of the camera are used to synchronize the binaural sounds and the video. The video and audio signals recorded by the camera are synchronized by default. Videos are recorded at 30 fps. Audios are recorded at 96 kHz.

We recorded videos on the streets of a big European City covering 165 locations within an area of 5km × 5km as shown in Fig. 2(a). We choose the locations next to road junctions, where we kept the rig stationary for the data recording of the traffic scenes. For each location, we recorded data for around 5-7 minutes. Our dataset consists of 165 city traffic videos and audios with an average length of 6.49 minutes, totalling 15 hours. We post-process the raw video-audio data into 2 second segments, resulting in 64,250 video clips. The videos contain numerous sound-making objects such as cars, trams, motorcycles, pedestrians, buses and trucks.

It is worth noticing that the professional 3Dio binaural mics simulate how human ears receive sound, which is different from the general stereo mics or monaural mics. Humans localize sound sources by using three primary cues [38]: interaural time difference (ITD), interaural level difference (ILD), and head-related transfer function (HRTF). ITD is caused by the difference between the times sounds reach the two ears. ILD is caused by the difference in sound pressure level reaching the two ears due to the acoustic shadow casted by the listener's

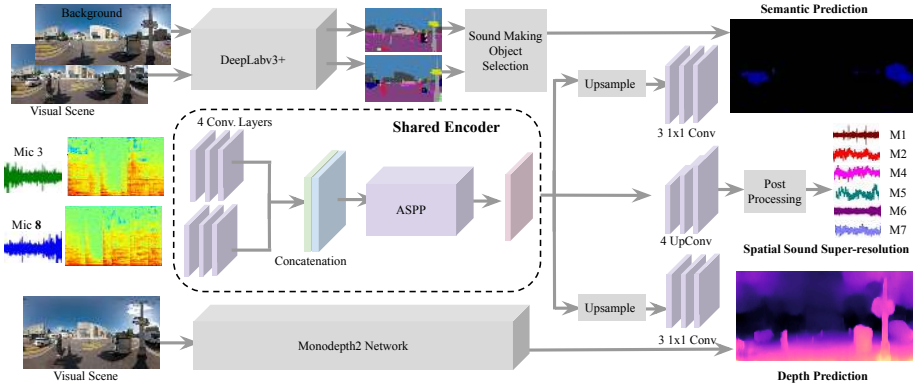


Fig. 3. The diagram of our method for the three considered tasks. The encoder is shared by all tasks and each task has its own decoder.

head. HRTF is caused because the pinna and head affect the intensities of sound frequencies. All these cues are missing in monaural audio, thus in this work, we focus on learning to localize semantic objects with binaural audios. This difference makes our dataset more suitable for cognitive applications.

3.2 Auditory Semantic Prediction

Since it is costly to create a large collection of human annotations, we follow a teacher-student learning strategy to transfer knowledge from vision to audio [8, 19]. Thus, our auditory semantic object prediction system is composed of two key components: a teacher vision network and a student audio network. The difference to previous methods is that we learn to transfer precise semantic segmentation results instead of scene labels [8] or bounding boxes [19].

Vision network. We employ the DeepLabv3+ model [12] as it is the current state-of-the-art (s-o-t-a) semantic segmentation model. We pick the middle frame of our 2-second video clip as the target frame and feed it to the teacher network to generate the semantic map. During training, each target frame is fed into a Cityscapes [13] pre-trained DeepLabv3+ to assign a semantic label to each pixel. Since objects in many classes such as *sky*, *road* and *parked cars* are not sound-making, it is very challenging to predict their semantic masks. Therefore, an object selection policy needs to be designed in order to collect the semantic masks of major sound-making objects.

Sound-making Object Collection. Our dataset contains numerous sound making objects such as cars, trams, motorcycles, pedestrians, buses and bicycles. The target objects must be constrained to make sure that the task is challenging but still achievable by current sensing systems and learning methods. In this work, we focus on *car*, *tram*, *motorcycle* due to their high occurrences in the datasets and because of them producing sufficient noise, as long as they move.

As to the motion status, we employ background subtraction to remove the background classes, such as road, building and sky, and the stationary foreground

classes such as parked cars. In the end, only the semantic masks of moving trams, moving cars and moving motorcycles are taken as the prediction targets. This selection guides the attention of the learning method to major sound-making objects and avoids localizing 'rare' sound-making objects and sound-irrelevant objects such as *parked car* and *sky*.

There is a rich body of studies for background estimation [10], in this work we employ a simple method based on majority voting. The method works surprisingly well. Specifically, the background image is computed as

$$I_{bg}(h,w) = \text{Mode}\{I_1(h,w), I_2(h,w), ..., I_T(h,w)\}, \tag{1}$$

where T is the length of the complete video sequence, (h,w) are pixel indexes, and Mode$\{.\}$ computes the number which appears most often in a set of numbers. Since the complete video sequence is quite long (about 5-7 mins), the background estimation is accurate and reliable.

The sound-making objects are detected by the following procedure: given an video frame I_t and its corresponding background image I_{bg}, we use DeepLabv3+ to get their semantic segmentation results Y_t, and Y_{bg}. Fig. 3 gives an illustration. The detection is done as

$$S(h,w) = \begin{cases} 1 \text{ if } Y_t(h,w) \in \{car, train, motorcycle\} \\ \quad \text{and } Y_t(h,w) \neq Y_{bg}(h,w), \\ 0 \text{ otherwise,} \end{cases} \tag{2}$$

where 1 indicates pixel locations of sound-making objects and 0 otherwise. Fig. 5 shows examples of the detected background and the detected sound-making target objects (i.e. ground truth).

Audio network. We treat auditory semantic prediction from the binaural sounds as a dense label prediction task. We take the semantic labels produced by the teacher vision network and filter by Eq. 2 as pseudo-labels, and then train a student audio network (BinauralSematicNet) to predict the pseudo semantic labels directly from the audio signals. A cross-entropy loss is used. The description of the network architecture can be found in Sec. 3.5 and in the supple. material.

3.3 Auditory Depth Perception

Similar to auditory semantic object localization, our auditory depth prediction method is composed of a teacher vision network and a student audio network. It provides auxiliary supervision for semantic perception task.

Vision network. We employ the MonoDepth2 model [24] given its good performance. We again pick the middle frame of our 2-second video clip as the target frame and feed it to the teacher network to generate the depth map. The model is pre-trained on KITTI [23]. We estimate the depth for the whole scene, similar to previous methods for holistic room layout estimation with sounds [4, 29].

Audio network. We treat depth prediction from the binaural sounds as a dense regression task. We take the depth values produced by the teacher vision network as pseudo-labels, and then train a student audio network (BinauralDepthNet) to regress the pseudo depth labels directly from the audio signals. The L2 loss is used. The description of the network architecture can be found in Sec. 3.5.

3.4 Spatial Sound Super-resolution (S^3R)

We leverage our omni-directional binaural microphones to design a novel task of spatial sound super-resolution (S^3R), to provide auxiliary supervision for semantic perception task. The S^3R task is motivated by the well-established studies [33, 48] about the effects of head movement in improving the accuracy of sound localization. Previous studies also found that rotational movements of the head occur most frequently during source localization [45]. Our omni-directional binaural microphone set is an ideal device to simulate head rotations at different discrete angles. Inspired by these findings, we study the effect of head rotation on auditory semantic and depth perception. It is very challenging to make a rotating sensor rig during signal presentation. Hence, we leverage the fact that our device has four pairs of binaural microphones resembling sparsely sampled head rotations, i.e. to an azimuth angle of 0° 90°, 180° and 270°, respectively.

Specifically, the S^3R task is to train a neural network to predict the binaural audio signals at other azimuth angles given the signals at the azimuth angle of 0°. We denote the received signal by the left and right ears at azimuth 0° by $x^{L_0}(t)$ and $x^{R_0}(t)$, respectively. We then feed those two signals into a deep network to predict the binaural audio signals $x^{L_\alpha}(t)$ and $x^{R_\alpha}(t)$ at azimuth $\alpha°$. Inspired by [21], we predict the difference of the target signals to the input signals, instead of directly predicting the absolute values of the targets. This way, the network is forced to learn the subtle difference. Specifically, we predict the difference signals:

$$\begin{aligned} x^{DL_\alpha}(t) &= x^{L_0}(t) - x^{L_\alpha}(t) \\ x^{DR_\alpha}(t) &= x^{R_0}(t) - x^{R_\alpha}(t), \end{aligned} \tag{3}$$

where $\alpha \in \{90°, 180°, 270°\}$. In order to leverage the image processing power of convolutional neural network, we follow the literature and choose to work with the spectrogram representation. Following [21], real and imaginary components of complex masks are predicted. The masks are multiplied with input spectrograms to get the spectrograms of the difference signals; the raw waveforms of the difference signals are then produced by applying Inverse Short-time Fourier Transform (ISTFT) [25]; and finally the target signals are reconstructed by adding back the reference raw waveform.

3.5 Network Architecture

Here, we present our multi-tasking audio network for all the three tasks. The network is composed of one shared encoder and three task-specific decoders. The pipeline of the method is shown in Fig. 3. As to the encoder, we convert the two channels of binaural sounds to log-spectrogram representations. Each spectrogram is passed through 4 strided convolutional (conv) layers with shared weights before they are concatenated. Each conv layer performs a 4×4 convolution with a stride of 2. Each conv layer is followed by a BN layer and a ReLU activation. The concatenated feature map is further passed to a Atrous Spatial Pyramid Pooling (ASPP) module [11]. ASPP has one 1×1 convolution and three 3×3

convolutions with dilation rates of 6, 12, and 18. Each of the convolutions has 64 filters and a BN layer. ASPP concatenates all the features and passes them through a 1×1 conv layer to generate binaural sound features. This feature map is taken as the input to our decoders.

Below, we present the three task-specific decoders. For the semantic prediction task, we employ a decoder to predict the dense semantic labels from the above feature map given by the shared encoder. The decoder comprises of an upsampling layer and three 1×1 conv layers. For the first two conv layers, each is followed by a BN and a ReLU activation; for the last one, it is followed by a softmax activation. We use the same decoder architecture for the depth prediction task, except that we use ReLU activation for the final conv layer of the decoder. For the S^3R task, we perform a series of 5 up-convolutions for the binaural feature map, each convolution layer is followed by a BN and a ReLU activation. The last layer is followed by a sigmoid layer which predicts a complex valued mask. We perform a few post processing steps to convert this mask to binaural sounds at other azimuth angles as mentioned in Sec. 3.4.

Loss function. We train the complete model shown in Fig. 3 in an end-to-end fashion. We use a) cross-entropy loss for the semantic prediction task which is formulated as dense pixel labelling to 3 classes, b) L2 loss for the depth prediction task to minimize the distance between the predicted depth values and the ground-truth depth values, and c) L2 loss for the S^3R task to minimize the distance between the predicted complex spectrogram and the ground truths. Hence, the total loss L for our multi-tasking learning is

$$L = L_{semantic} + \lambda_1 L_{depth} + \lambda_2 L_{s^3r} \tag{4}$$

where λ_1 and λ_2 are weights to balance the losses. The detailed network architecture will be provided in the supple. material.

4 Experiments

Data Preparation. Our dataset comprises of 64,250 video segments, each of 2 seconds long. We split the samples into three parts: 51,400 for training, 6,208 for validation and 6,492 for testing. We use 2-seconds segments following [19], which shows that performances are stable for $\geq$1 second segments. For each scene, a background image is also precomputed according to Eqn. 1. For the middle frame of each segment, we generate the ground truth for semantic segmentation task by using the Deeplabv3+ [12] pretrained on Cityscapes dataset [13] and the depth map by using the Monodepth2 [24] pretrained on KITTI dataset [23]. We use *AuditoryTestPseudo* to refer the test set generated this way. In order to more reliably evaluate the method, we manually annotate the middle frame of 80 test video segments for the three considered classes, namely car, train and motorcycle. We carefully select the video segments such that they cover diverse scenarios such as daylight, night, foggy and rainy. We use LabelMeToolbox [40] for the annotation and follow the annotation procedure of Cityscapes [13]. We call this test set *AuditoryTestManual*.

Methods	Microphone		Auxiliary Tasks		AuditoryTestPseudo				AuditoryTestManual			
	Mono	Binaural	S^3R	Depth	Car	MC	Train	All	Car	MC	Train	All
BG					8.79	4.17	24.33	12.61	-	-	-	-
Visual					-	-	-	-	79.01	39.07	77.34	65.35
Mono	✓				33.53	7.86	24.99	22.12	30.13	9.21	24.1	21.14
Ours(B)		✓			35.80	19.51	40.71	32.01	35.30	13.28	35.48	28.02
Ours(B:D)		✓		✓	33.53	28.01	55.32	38.95	32.42	25.8	50.12	36.11
Ours(B:S)		✓	✓		35.62	36.81	**56.49**	42.64	**38.12**	26.5	49.02	37.80
Ours(B:SD)		✓	✓	✓	**35.81**	**38.14**	56.25	**43.40**	35.51	**28.51**	**50.32**	**38.01**

Table 1. Results of auditory semantic prediction. The results of DeepLabv3+ on the background image (BG) and on the target middle frame (Visual) are reported for reference purpose. mIoU (%) is used. MC denotes Motorcycle.

For all the experiments, a training or a testing sample consists of a 2-second video segment and eight 2-second audio channels. We preprocess audio samples following techniques from [21,52]. We keep the audio samples at 96kHz and their amplitude is normalized to a desired RMS level, which we set as 0.1 for all the audio channels. For normalization, we compute mean RMS values of amplitude over the entire dataset separately for each channel. An STFT is applied to the normalized waveform, with a window size of 512 (5.3ms), hop length of 160 (1.6ms) resulting a Time-Frequency representation of size of 257×601 pixels. Video frames are resized to 960×1920 pixels to fit to the GPU.

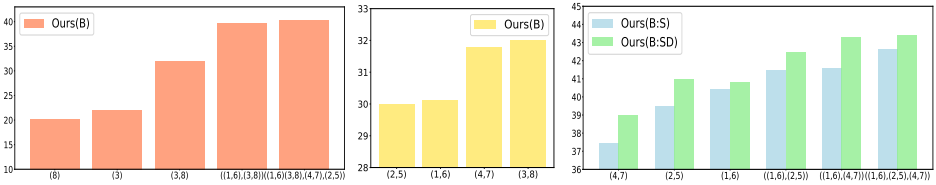
Implementation Details. We train our complete model using Adam solver [30] with a learning rate of 0.00001 and we set a batch size of 2. We train our models on GeForce GTX 1080 Ti GPUs for 20 epochs. For joint training of all the three tasks, we keep $\lambda_1 = 0.2$ and $\lambda_2 = 0.2$ in Eq. 4.

Evaluation metrics. We use the standard mean IoU for the semantic prediction task. For audio super resolution, we use MSE error for the spectrograms and the envelope error for the waveforms as used in [21]. For depth prediction, we employ RMSE, MSE, Abs Rel and Sq Rel by following [24].

4.1 Auditory Semantic Prediction

We compare the performance of different methods and report the results in Tab. 1. The table shows our method learning with sounds can generate promising results for dense semantic object prediction. We also find that using binaural sounds *Ours(B)* generates significant better results than using *Mono* sound. This is mainly because the major cues for sound localization such as ILD, ITD, and HRTF are missing in *Mono* sound. Another very exciting observation is the joint training with the depth prediction task and the S^3R task are beneficial to the semantic prediction task. We speculate that this is because all the three tasks benefit from a same common goal – reconstructing 3D surround sounds from Binaural sounds. Below, we present our ablation studies.

S^3R and depth prediction both give a boost. As can be seen in Tab. 1, by adding S^3R or depth prediction as an auxiliary task, indicated by *Ours(B:S)* and *Ours(B:D)* respectively, improves the performance of our baseline *Ours(B)*



(a) #(input mics) (b) input mic pairs (c) Combination of output mics

Fig. 4. Semantic prediction ablation study results (mIoU) with different set of microphone used as inputs under *Ours(B)* in (a) and (b) and ablation on output microphones for S^3R in *Ours(B:S)* and *Ours(B:SD)* in (c).

clearly. We also observe that using both of the auxiliary tasks together, indicated by *Ours(B:SD)*, yields the best performance. This shows that both S^3R and depth helps. This is because the three tasks share the same goal – extracting spatial information from binaural sounds.

Adding input channels increases the performance. We compare the auditory semantic prediction (without using auxiliary tasks) accuracies in Fig. 4(a) for different set of input microphones. Here, we experiment with a) Mono sound from 3 (front) or 8 (back) microphone, b) binaural sounds from the pair (3,8), c) 2 pairs of binaural sound channels ((1,6),(3,8)) which faces in four orthogonal directions, and d) 4 pairs of binaural sound channels. We see that semantic prediction accuracy increases from 22.12% when using *Mono* sound to 40.32% when using all 8 channels. This shows that semantic prediction improves with the access to more sound channels.

Orientation of microphones matter. Fig. 4(b) shows the auditory semantic prediction results (without using auxiliary tasks) from different orientations of the input binaural pairs for the same scene. The pair (3,8) is aligned in parallel with the front facing direction of the camera as can be seen in Fig. 2. We define this as orientation of 0°. Then, we have other pairs (1,6), (4,7) and (2,5) orientating at azimuth angles of 90°, 180°and 270°respectively. We observe that (3,8) outperforms all other pairs.

Removing output channels degrades the performance. We vary the number of output microphone pairs for S^3R under the two multi-tasking models *Ours(B:S)* and *Ours(B:SD)*. We fix the input to pair (3,8) and experiment with different number of output pairs, ranging from 1 to 3. The results are presented in Fig. 4(c) under these settings. We see that spatial sound resolution to 3 binaural pairs performs better than to 1 or 2 binaural pairs. The more output channels we have, the better the semantic prediction results are.

ASPP is a powerful audio encoder. We have found in our experiments that ASPP is a powerful encoder for audio as well. We compare our audio encoder with and without the ASPP module. For instance, Mono sound with ASPP clearly outperforms itself without ASPP – adding ASPP improves the performance from 13.21 to 22.12 for mean IoU. The same trend is observed for other cases.

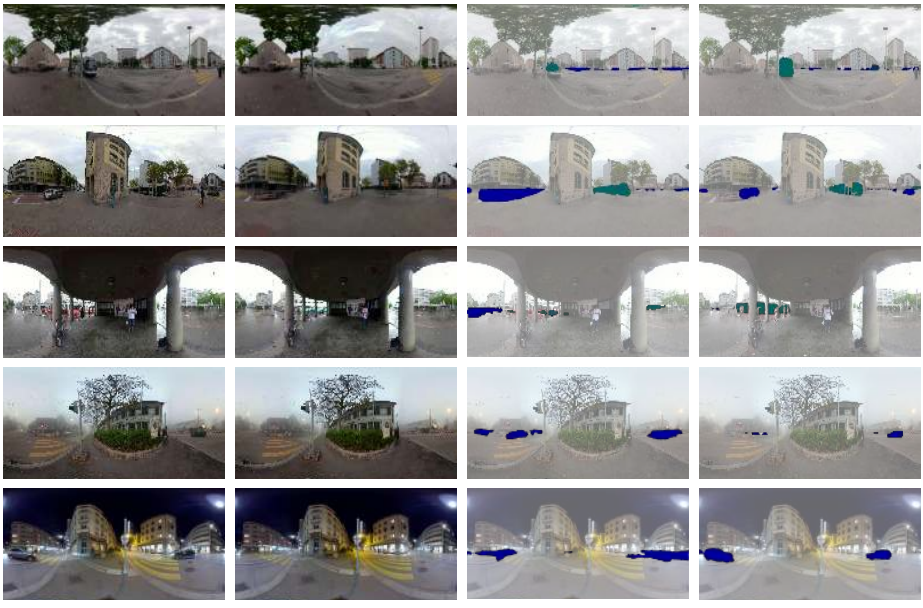


Fig. 5. Qualitative results of auditory semantic prediction by our approach. The first column shows the visual scene, the second for the computed background image, the third for the semantic object masks predicted by our approach, and the fourth for the ground truth.

Microphone		Joint Tasks		Metrics			
Mono	Binaural	Semantic	S^3R	Abs Rel	Sq Rel	RMSE	MSE
✓				118.88	622.58	5.413	0.365
	✓			108.59	459.69	5.263	0.331
	✓		✓	90.43	400.53	5.193	0.318
	✓	✓		87.96	290.06	5.136	0.315
	✓	✓	✓	**84.24**	**222.41**	**5.117**	**0.310**

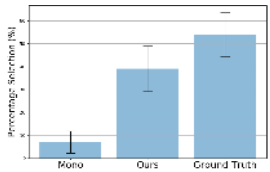
Table 2. Depth prediction results with Mono and binaural sounds under different multi-task settings. For all the metrics, lower score is better.

4.2 Auditory depth prediction

We report the depth prediction results in Tab. 2. The first row represents the depth prediction from Mono sounds alone while the second row represents our baseline method where the depth map is predicted from binaural sounds. As a simple baseline, we also compute the mean depth over the training dataset and evaluate over the test set. The RMSE and MSE scores are 15.024 and 0.864, respectively, which are 2.5 times worse than our binaural sound baseline. The multi-task learning with S^3R and semantic prediction under shared audio encoder also provides pronounced improvements for depth prediction. Jointly training all the three tasks yields the best performance for depth prediction as well.

Joint Tasks		Mic Ids		Metrics			
Semantic	Depth	In	Out	MSE-1	ENV-1	MSE-2	ENV-2
		(3,8)	(1,6)	0.1228	0.0298	0.1591	0.0324
	✓	(3,8)	(1,6)	0.0984	0.0221	0.1044	0.0267
✓	✓	(3,8)	(1,6)	0.0978	0.0218	0.1040	0.0264
✓		(3,8)	(1,6)	**0.0956**	**0.0214**	**0.1001**	**0.0243**

(a) S^3R results in multi-tasking



(b) Subjective evaluation

Table 3. a) S^3R results. MSE1 and MSE2 represent mean squared error while ENV1 and ENV represent envelope error for the 2 output channels of binaural sounds. For all metrics, lower score is better. b) Subjective assessment of the generated binaural sounds.

4.3 Spatial Sound Super-resolution

Tab. 3 shows the results of S^3R as a stand-alone task (first row) and under the multi-task setting. To keep it simple, we estimate the sound signals of microphone pair $(1, 6)$ alone from the microphone pair $(3, 8)$. We can see from Fig. 2 that these two pairs are perpendicular in orientation, so the prediction is quite challenging. We can observe from Tab. 3 that the multi-task learning with semantic prediction task and depth prediction task outperforms the accuracy of the stand-alone S^3R model. Hence, the multi-task learning also helps S^3R – the same trend as for semantic perception and depth perception. We also conduct a user study for the subjective assessment of the generated binaural sounds. The participants listen to ground-truth binaural sounds, binaural sounds generated from *Mono* approach (Tab. 1) and from our approach. We present two (out of three) randomly picked sounds and ask the user to select a preferred one in temrs of binaural sound quality. Tab. 3(b) shows the percentage of times each method is chosen as the preferred one. We see that *Ours* is close to the ground truth selection implying that our predicted binaural sounds are of high quality.

4.4 Qualitative results

We show qualitative results in Fig. 5 for the task of auditory semantic prediction. We also show the detected background image and the ground truth segmentation mask. The last three rows are devoted to the results in rainy, foggy and night conditions respectively. We observe that our model remains robust to adverse visual condition. Of course, if the rain is too big, it will become an adverse auditory condition. In Fig. 6, we show two results by the multi-task setting. We show the predictions and the ground truths for all the three tasks. It can be seen that the major sound-making objects can be properly detected. We see that the depth results reflect the general layout of the scene, though they are still coarser than the results of a vision system. This is valuable given the fact that binaural sounds are of very low-resolution – two channels in total. More qualitative results are provided in the supplemental material.

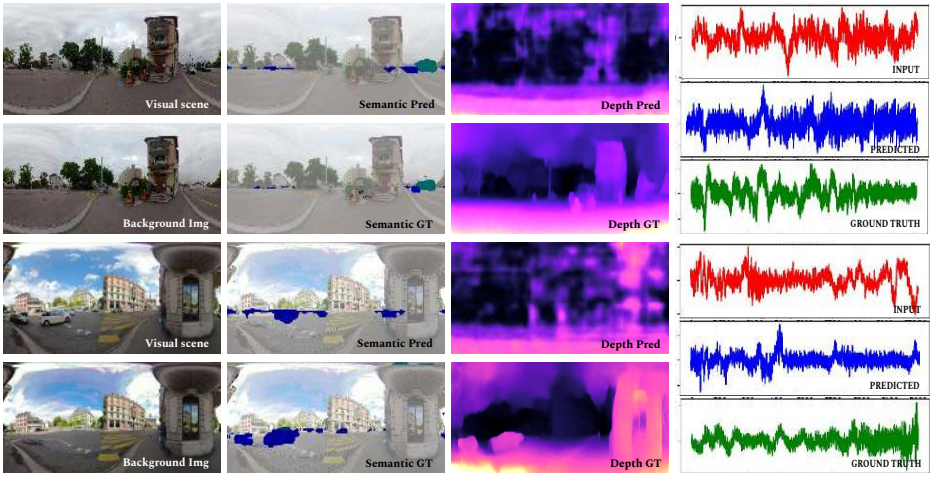


Fig. 6. Qualitative results of all three tasks. Better view in color.

4.5 Limitations and future work

We obtain the ground truth of semantic segmentation and depth prediction by using the s-o-t-a pretrained vision models. These pretrained models remain as the upper bound in our evaluations. Moreover, the vision models are pretrained on perspective images and we apply them to panoramic images. This is due to the limited amount of datasets and annotations for panoramic images. We would like to note that most of the interesting objects appear in the equatorial region. For that region, the distortion of the panoramic images is less severe, and hence the results are less affected. In future, we plan to incorporate a 3D LiDAR to our sensor setup to get accurate ground-truth for depth prediction.

We work with 3 object classes for semantic prediction task. This is because some classes are very rare in the middle-sized dataset, which already takes a great deal of effort to create. In comparison, the recent works [19,28] only deals with one class in real world environment. To our best knowledge, we are the first to work with multiple classes in an unconstrained real environment.

5 Conclusion

The work develops an approach to predict the semantic labels of sound-making objects in a panoramic image frame, given binaural sounds of the scene alone. To enhance this task, two auxiliary tasks are proposed – dense depth prediction of the scene and a novel task of spatial sound super-resolution. All the three tasks are also formulated as multi-task learning and is trained in an end-to-end fashion. This work has also proposed a novel dataset Omni Auditory Perception dataset. Extensive experiments have shown that 1) the proposed method achieves promising results for all the three tasks; 2) the three tasks are mutually beneficial and 3) the number and orientations of microphones are both important.

Acknowledgement

This work is funded by Toyota Motor Europe via the research project TRACE-Zurich. We would like to thank Danda Pani Paudel and Vaishakh Patil for helpful discussions.

References

1. Computational auditory scene analysis. Computer Speech and Language **8**(4), 297 – 336 (1994)
2. Seeing and Hearing Egocentric Actions: How Much Can We Learn? (2019)
3. Albanie, S., Nagrani, A., Vedaldi, A., Zisserman, A.: Emotion recognition in speech using cross-modal transfer in the wild. In: ACM Multimedia (2018)
4. Antonacci, F., Filos, J., Thomas, M.R.P., Habets, E.A.P., Sarti, A., Naylor, P.A., Tubaro, S.: Inference of room geometry from acoustic impulse responses. IEEE Transactions on Audio, Speech, and Language Processing **20**(10), 2683–2695 (2012)
5. Arandjelovic, R., Zisserman, A.: Look, listen and learn. In: The IEEE International Conference on Computer Vision (ICCV) (2017)
6. Arandjelović, R., Zisserman, A.: Objects that sound. In: ECCV (2018)
7. Argentieri, S., Danès, P., Souères, P.: A survey on sound source localization in robotics: From binaural to array processing methods. Computer Speech and Language **34**(1), 87 – 112 (2015)
8. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: Advances in Neural Information Processing Systems (2016)
9. Barzelay, Z., Schechner, Y.Y.: Harmony in motion. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
10. Brutzer, S., Höferlin, B., Heidemann, G.: Evaluation of background subtraction techniques for video surveillance. In: CVPR (2011)
11. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence **40**(4), 834–848 (2017)
12. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
13. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
14. Delmerico, J., Mintchev, S., Giusti, A., Gromov, B., Melo, K., Horvat, T., Cadena, C., Hutter, M., Ijspeert, A., Floreano, D., Gambardella, L., Siegwart, R., Scaramuzza, D.: The current state and future outlook of rescue robotics. Journal of Field Robotics **36**(7), 1171–1191 (2019)
15. Dokmanic, I., Parhizkar, R., Walther, A., Lu, Y.M., Vetterli, M.: Acoustic echoes reveal room shape. Proceedings of the National Academy of Sciences **110**(30), 6. 12186–12191 (2013)
16. Ernst, M.O., Bülthoff, H.H.: Merging the senses into a robust percept. Trends in Cognitive Sciences **8**, 162–169 (2004)

17. Fazenda, B., Hidajat Atmoko, Fengshou Gu, Luyang Guan, Ball, A.: Acoustic based safety emergency vehicle detection for intelligent transport systems. In: ICCAS-SICE (2009)
18. Fendrich, R.: The merging of the senses. Journal of Cognitive Neuroscience **5**(3), 373–374 (July 1993). https://doi.org/10.1162/jocn.1993.5.3.373
19. Gan, C., Zhao, H., Chen, P., Cox, D., Torralba, A.: Self-supervised moving vehicle tracking with stereo sound. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
20. Gao, R., Feris, R., Grauman, K.: Learning to separate object sounds by watching unlabeled video. In: ECCV (2018)
21. Gao, R., Grauman, K.: 2.5 d visual sound. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 324–333 (2019)
22. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
23. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)
24. Godard, C., Aodha, O.M., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3828–3838 (2019)
25. Griffin, D., Jae Lim: Signal estimation from modified short-time fourier transform. In: ICASSP '83. IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 8, pp. 804–807 (April 1983). https://doi.org/10.1109/ICASSP.1983.1172092
26. Hecker, S., Dai, D., Van Gool, L.: End-to-end learning of driving models with surround-view cameras and route planners. In: The European Conference on Computer Vision (ECCV) (September 2018)
27. Huang, W., Alem, L., Livingston, M.A.: Human factors in augmented reality environments. Springer Science and Business Media (2012)
28. Irie, G., Ostrek, M., Wang, H., Kameoka, H., Kimura, A., Kawanishi, T., Kashino, K.: Seeing through sounds: Predicting visual semantic segmentation results from multichannel audio signals. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3961–3964. IEEE (2019)
29. Kim, H., Remaggi, L., Jackson, P.J., Fazi, F.M., Hilton, A.: 3d room geometry reconstruction using audio-visual sensors. In: International Conference on 3D Vision (3DV). pp. 621–629 (2017)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
31. Li, D., Langlois, T.R., Zheng, C.: Scene-aware audio for 360° videos. ACM Trans. Graph. **37**(4) (2018)
32. Marchegiani, L., Posner, I.: Leveraging the urban soundscape: Auditory perception for smart vehicles. In: IEEE International Conference on Robotics and Automation (ICRA) (2017)
33. McAnally, K.I., Martin, R.L.: Sound localization with head movement: implications for 3-d audio displays. Frontiers in Neuroscience **210**(8) (2014)
34. Mousavian, A., Pirsiavash, H., Košecká, J.: Joint semantic segmentation and depth estimation with deep convolutional networks. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 611–619. IEEE (2016)
35. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: The European Conference on Computer Vision (ECCV) (September 2018)

36. Owens, A., Isola, P., McDermott, J., Torralba, A., Adelson, E.H., Freeman, W.T.: Visually indicated sounds. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
37. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV (2016)
38. Rascon, C., Meza, I.: Localization of sound sources in robotics: A review. Robotics and Autonomous Systems **96**, 184 – 210 (2017)
39. Rosenblum, L.D., Gordon, M.S., Jarquin, L.: Echolocating distance by moving and stationary listeners. Ecological Psychology **12**(3), 181–206 (2000)
40. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: a database and web-based tool for image annotation. International journal of computer vision **77**(1-3), 157–173 (2008)
41. Salamon, J., Jacoby, C., Bello, J.P.: A dataset and taxonomy for urban sound research. In: ACM International Conference on Multimedia (2014)
42. Saxena, A., Ng, A.Y.: Learning sound location from a single microphone. In: IEEE International Conference on Robotics and Automation (2009)
43. Senocak, A., Oh, T.H., Kim, J., Yang, M.H., So Kweon, I.: Learning to localize sound source in visual scenes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
44. Simeoni, M.M.J.A., Kashani, S., Hurley, P., Vetterli, M.: Deepwave: A recurrent neural-network for real-time acoustic imaging p. 38 (2019)
45. Thurlow, W.R., Mangels, J.W., Runge, P.S.: Head movements during sound localizationtd. The Journal of the Acoustical Society of America **489**(42) (1967)
46. Tiete, J., Domínguez, F., da Silva, B., Segers, L., Steenhaut, K., Touhafi, A.: Soundcompass: A distributed mems microphone array-based sensor for sound source localization. In: Sensors (2014)
47. Urmson, C., Anhalt, J., Bae, H., Bagnell, J.A.D., Baker, C.R., Bittner, R.E., Brown, T., Clark, M.N., Darms, M., Demitrish, D., Dolan, J.M., Duggins, D., Ferguson, D., Galatali, T., Geyer, C.M., Gittleman, M., Harbaugh, S., Hebert, M., Howard, T., Kolski, S., Likhachev, M., Litkouhi, B., Kelly, A., McNaughton, M., Miller, N., Nickolaou, J., Peterson, K., Pilnick, B., Rajkumar, R., Rybski, P., Sadekar, V., Salesky, B., Seo, Y.W., Singh, S., Snider, J.M., Struble, J.C., Stentz, A.T., Taylor, M., Whittaker, W.R.L., Wolkowicki, Z., Zhang, W., Ziglar, J.: Autonomous driving in urban environments: Boss and the urban challenge. Journal of Field Robotics Special Issue on the 2007 DARPA Urban Challenge, Part I **25**(8), 425–466 (June 2008)
48. Wallach, H.: The role of head movements and vestibular and visual cues in sound localization. Journal of Experimental Psychology **27**(4), 339 (1940)
49. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2800–2809 (2015)
50. Ye, M., Yu Zhang, Yang, R., Manocha, D.: 3d reconstruction in the presence of glasses by acoustic and stereo fusion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
51. Zhao, H., Gan, C., Ma, W.C., Torralba, A.: The sound of motions. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
52. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: The European Conference on Computer Vision (ECCV) (September 2018)