Supplementary Material: 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] CVL, ETH Zurich
[2] PSI, KU Leuven
{arunv,dai,vangool}@vision.ee.ethz.ch

Abstract. In the supplementary material, we organize as follows: a) computed background images, b) learning details, c) more dataset statistics, d) evaluation of auditory semantic prediction for the following cases: i) ablation of different input and output microphone combinations, ii) semantic prediction for the last frame of each segment instead of the middle one, e) ablation study on different weights used in the loss function, f) more qualitative results with attached audio samples.

1 Background Images

We compute background image for each scene in our dataset as explained in Sec. 3 of the main paper. In Figure 1, we provide some examples of the computed background images from scenes in daylight, twilight, nighttime, and foggy conditions. We can observe that these background images are quite clean without any foreground objects like car, bus, tram or motorcycles. Nevertheless, the bottom row in Figure 1 shows few noisy background images due to the heavy wind or direct sunlight during the shoot of the scene with our setup.

2 Learning Details

Here, our multi-tasking audio network for all the three tasks is composed of one shared encoder and three task-specific decoders.

Network. Our shared encoder network has 4 Conv blocks, each comprising of a 4×4 convolution, a ReLU, and a BatchNorm. Each of the 3 decoder branches of our network upsamples the encoder's output to the spatial size for the corresponding tasks. We keep the spatial size to 480×960 for semantic prediction and depth estimation while 257×601 for sound spatial super-resolution (S^3R) which is same as the input audio spectrograms. We set the number of output channels to 3 and 1 for the semantic prediction and depth estimation branches respectively and 2 for S^3R task. The upsampling is done using bilinear interpolation. We train the network from scratch for all the three tasks in a multi-tasking

Tasks			Parameters		
Sem Seg	S^3R	Depth	# Parameters	Training time	Inference time
✓		✓	20.1M	0.7	0.5s
✓	✓		20.7M	0.7	0.5s
	✓	✓	20.9M	0.7	0.5s
✓	✓	✓	22.4M	1.5s	1s

Table 1. Comparison of parameters of our models. Training and inference time are for an iteration.

Fig. 1. Selected images of background images from our dataset.

framework. We initialize the weights of the network by following He *et al.* [1] which is motivated for networks with ReLU-like activations.

Learning. The loss function is explained in the main paper. We use the Adam optimizer with a learning rate of 0.00001 and a batch size of 2. Our network is trained for a total of 20 epochs on 51.4k audio segments from our training dataset. In Table 1, we show the comparison of number of parameters of our model in different cases and the training time of an iteration with batch size 2, when a NVIDIA GeForce GTX 1080 Ti GPU is used.

3 Dataset Stats

We perform a careful selection of training samples such that they are not in silent phase of the traffic scene. A sample is used only if its audio energy is

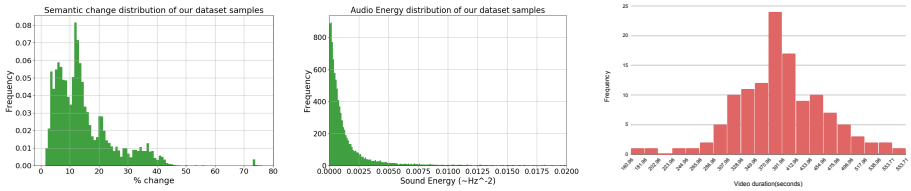


Fig. 2. Distribution of foreground semantic changes on the left and sound energy distribution over the samples from our dataset in the middle. On the right, we show the distribution of video lengths in our dataset.

#	Mics	Car	MC	Train	All
1	3	33.53	7.86	24.99	22.12
	8	32.66	6.55	21.58	20.26
2	3,8	35.8	19.51	40.71	32.01
4	1,6,3,8	**49.01**	18.05	51.98	39.68
8	1-8	40.61	**22.23**	**58.13**	**40.32**

(a) Ablation on # input microphones

Mics	Head∠	Car	MC	Train	All
1,6	90°	40.74	10.18	39.44	30.12
2,5	270°	40.31	10.72	38.91	29.98
4,7	180°	**42.37**	18.2	34.78	31.78
3,8	0°	35.8	**19.51**	**40.71**	**32.01**

(b) Ablation on input mic orientations

Table 2. Auditory semantic prediction results (mIoU (%)) with different choices of microphones and choices of microphone pairs. These are detailed numbers of Figure 4(a) and (b) of the main paper.

beyond a chosen threshold. We plot the audio energy distribution over all the audio segments in Figure 2. Also, we only use samples of which the images have at least 5% different semantic labels than the corresponding background images. Figure 2 also shows the distribution of lengths of the captured videos in our dataset. The average length is 6.49 minutes long.

4 Evaluation

4.1 Different number of microphones

We compare the auditory semantic prediction (without using auxiliary tasks) accuracies in Figure 4(a) and Figure 4(b) of the main paper for a) different set of input microphones and b) different orientations of the input binaural pairs for the same scene, respectively. Here, we provide detailed table of scores for the same in Table 2(a) and 2(b). We perform the ablation on the number of output microphone pairs for S^{3}R under the two multi-tasking models *Ours(B:S)* and *Ours(B:SD)* in Figure 4(c) of the main paper and we provide its detailed numbers in Table 3.

Methods	Mic IDs	Car	Motorcycle	Train	All
Ours(B:S)	4,7	40.90	24.38	46.98	37.42
	2,5	37.93	30.99	49.50	39.47
	1,6	**44.18**	26.83	50.27	40.42
	1,6,2,5	34.81	34.78	54.84	41.47
	1,6,4,7	33.29	32.33	**59.16**	41.59
	1,6,2,5,4,7	35.62	**36.81**	56.49	**42.64**
Ours(B:SD)	4,7	37.26	28.57	51.13	38.98
	2,5	35.09	34.09	53.78	40.98
	1,6	**38.54**	29.90	54.02	40.81
	1,6,2,5	34.16	37.39	56.76	42.47
	1,6,4,7	34.28	37.21	**58.42**	43.30
	1,6,2,5,4,7	35.81	**38.14**	56.25	**43.40**

Table 3. Auditory semantic prediction results (mIoU (%)) with ablation on the combination of output microphones for S^3R in *Ours(B:S)* and *Ours(B:SD)*.

4.2 Semantic prediction for the last frame

We also train and evaluate our method for the last frame of the 2-second audio-video segment. We train and test the baselines and our final model for the last frame and report the results in Table 4. The results show similar trend as shown in Table 1 of the main paper where evaluation is done on the middle frame.

Models	Car	MC	Train	All
Mono	33.87	8.14	25.35	22.45
Ours	**39.24**	**30.05**	**54.57**	**41.28**

Table 4. Auditory semantic prediction results (mIoU(%)) for the last frame of the segment with binaural sounds of microphones (3,8).

5 Ablation study

Figure 3 shows the ablation studies on varying weights for individual task loss functions. We experiment on 4 different weights for semantic prediction loss and depth estimation loss. Let w_1 and w_2 be weights for semantic prediction and depth estimation losses respectively. We experiment with $0.5, 1, 5$ and 10 for w_1 and w_2. These values can be mapped to λ_1 and λ_2 of Equation 4 of the main paper as,

$$\lambda_1 = w_2/w_1 \text{ and } \lambda_2 = 1/w_1. \quad (1)$$

The plots show that the performance of the primary task is better if larger weight is given to it compared to the weights of the auxiliary tasks. We note

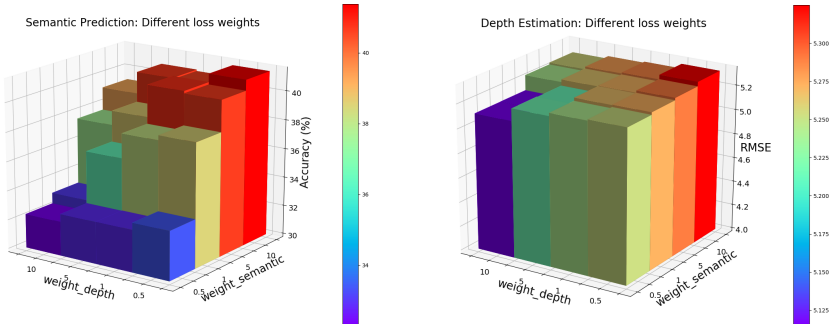


Fig. 3. Grid search on the weights of loss functions of auditory semantic prediction and depth estimation. On the left, we show the results on semantic prediction and on the right, the results are of depth estimation. Higher accuracy in semantic prediction is better while lower RMSE is better for depth estimation.

the same in Figure 3 that semantic prediction works best at $w_1 = 10$ and depth estimation at $w_2 = 10$.

6 Qualitative results

6.1 Differing sound volumes

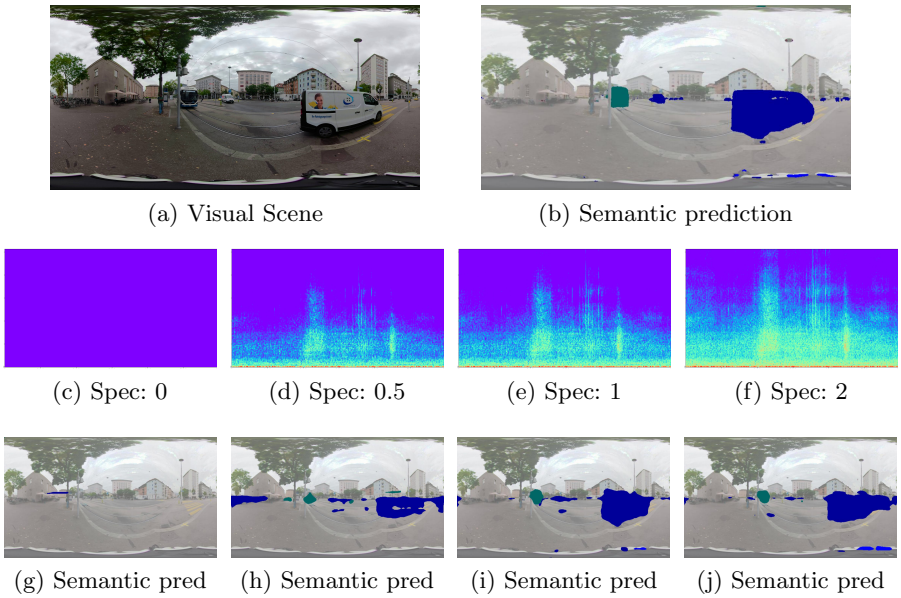
We examine the invariance of our model to the volume of the audio signal. We experiment by scaling the input binaural sounds with different multipliers: 0, 0.5, 1 and 2 as shown in the Figure 4. We can observe that the semantic prediction remains robust with different scales of the input sounds.

6.2 More Qualitative Results

We present more qualitative results in Figure 5 for the task of auditory semantic prediction. In Figure 6, we present the results of depth prediction and S^{3}R under multi-task setting and show their corresponding ground truths. We have also attached the sample audio files of S^{3}R task from our approach and the ground truth, as a part of the supplementary material.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)



(a) Visual Scene (b) Semantic prediction

(c) Spec: 0 (d) Spec: 0.5 (e) Spec: 1 (f) Spec: 2

(g) Semantic pred (h) Semantic pred (i) Semantic pred (j) Semantic pred

Fig. 4. Qualitative results of auditory semantic prediction by our approach on different amplification of input audio signal at inference stage. We show the input spectrograms and the multiplier of amplitude for the signal from (c) to (f) and corresponding semantic prediction in (g) to (j). We show the visual scene and ground truth semantic prediction in (a) and (b) respectively.

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. The object masks are depicted as highlighted colours in Car, Train and Motorcycle.

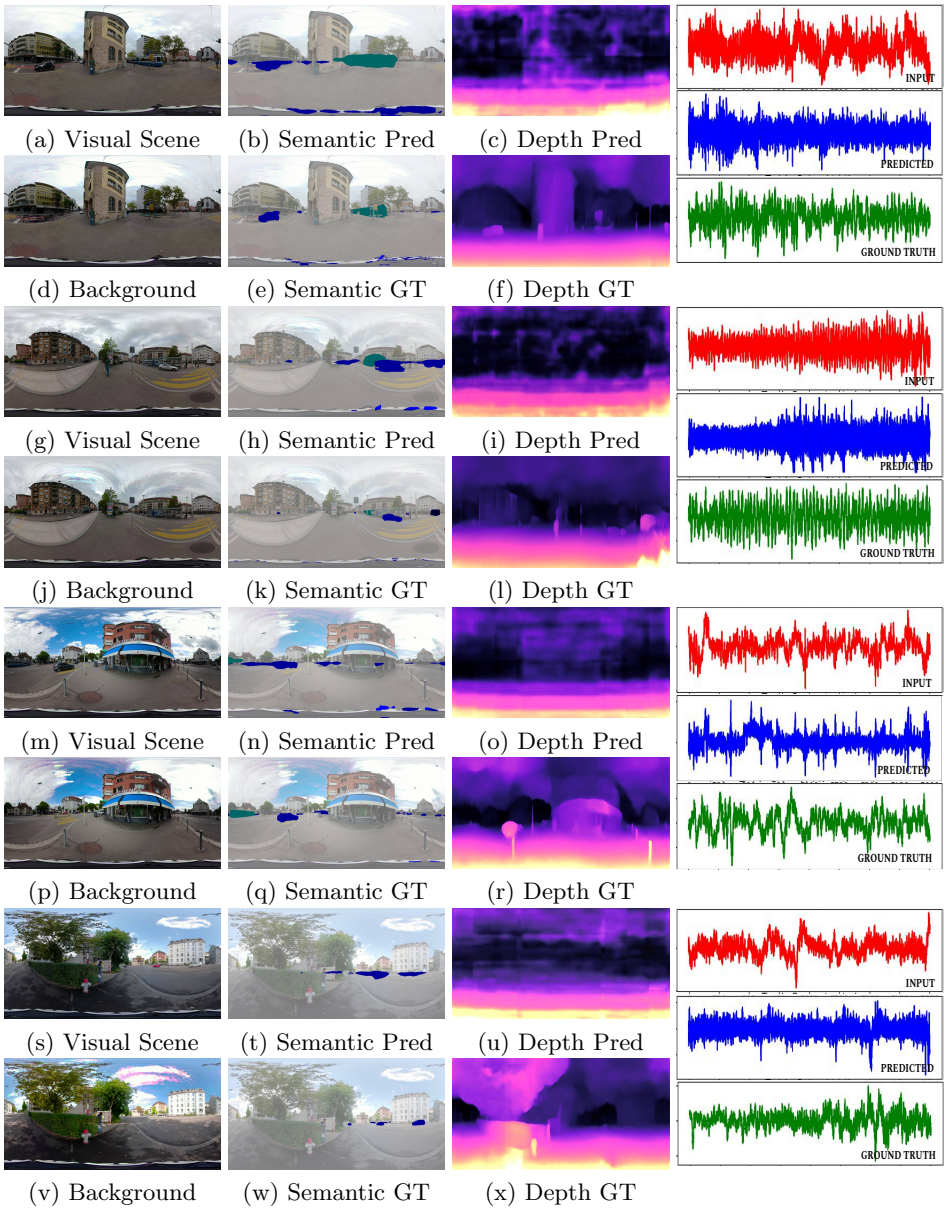


Fig. 6. Qualitative results of all three tasks. The object masks are depicted as highlighted colours in Car, Train and Motorcycle. Better view in color.