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

Supplementary Document

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

In this supplementary document, we present additional details, results and discussion in support of the main manuscript. Specifically, we present

- Additional details on the dataset (Section 1)
- Additional details on beamforming algorithm (Section 2)
- Additional results on object detection from acoustic signals (Section 3)
- Additional results on future frame prediction (Section 4)

The code and dataset described in this document are available on our https://light.princeton.edu/seeingwithsound/ .

1. Additional Dataset Details

We introduce a long-range automotive acoustic beamforming dataset along with additional sensing modalities such as RGB images, lidar, GPS and IMU sensor measurements. Our dataset covers automotive scenes from residential neighborhoods and urban downtown, and includes both day and night conditions. As described in the main manuscript, we use a Sorama CAM1K 1024 channel microphone array with over 45 kHz sampling rate and capable of measuring acoustic signals within a large frequency range of 1 Hz - 20 kHz.

We provide both processed beamforming maps as well as raw microphone array recordings in our dataset. See Figure 1 for a schematic of our microphone array. The code to process the microphone array measurements is also made available for exploring beamforming signals at a variery of frequencies. Each acoustic frequency range uncovers information about different sources. We visualized beamformed signals at a frequency of 4000 Hz in the main manuscript. We showcase additional examples of our acoustic beamforming dataset in Figure 2. Specifically, we show examples of beamformed maps at different frequencies (315 Hz, 400 Hz, 500 Hz, 630 Hz, 800 Hz, 1000 Hz, 1250 Hz, 1600 Hz, 2000 Hz, 2500 Hz, 4000 Hz). As can be seen in Figure 2, the resolution of spatial localization of vehicles improves with increasing frequency. However, we empirically notice that ambient noise is captured well at very high frequencies, thereby corrupting the beamformed signal maps. We use the beamforming maps corresponding to 4000 Hz for our experiments.

As discussed in the main manuscript, while most existing sensors leverage light (electromagnetic) waves which are undisturbed in good weather conditions but suffer in low light, our method utilizes the acoustic pressure signals whose measurements are not dramatically distorted in adverse conditions such as low light. As can be seen in Section 3, while the RGB camera measurements exhibit significant degradations due to low light and glare, the beamformed signals localize the sound emitting vehicles similar to that of daylight scenes. This results in improved detections in challenging environments.

1.1. Additional Details on Annotations

Our dataset consists of 16,324 images annotated manually by experts for various image and sound classes as described in Table 1.

Images 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. Each instance of an object belonging to any of the automotive

Image	Sound	Description
car	small_vehicle	sound from a small vehicle like car, van, suv
bus/truck/tram	ego-vehicle	sounds from the data collection platform eg. engine revving and tyre
pedestrian	trailer	sound from an accessory or an unpowered vehicle towed by another vehicle
traffic_sign	horn	warning noises emitted by vehicles
traffic_light	construction_noise	sounds relating to construction activity
	crosswalk_noise	pedestrian crosswalk alert sounds
	large_vehicle	sounds from heavy vehicles like semi-trucks, buses
	emergency_vehicles	sirens from emergency vehicles
	walking_sounds	sounds from a pedestrian or a large group of pedestrians
	cannot_distinguish	unidentifiable sound sources with less than 30% certainty
	custom	identified sounds that are not part of the list above

Table 1. List of image labels and sound labels used for annotating the 16,324 images in the dataset.

classes (Table 1) were annotated using tightly fitted 2D bounding boxes aligned to the image axis. Object labels are encoded as $[x_1, y_1, x_2, y_2]$, which are the top left (x_1, y_1) and bottom right (x_2, y_2) coordinates in the image frame. In addition to bounding boxes, additional attributes like occlusion, truncation, direction, parked status, motion details are also provided.

Sound 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). Sound labels based on typical sound sources in a road scene are shown in Table 1. Our expert annotators attached one of the 11 labels as described in Table 1 to each of the two domains in each image, based on whether the sound was distinct and in the foreground (*dominant*) or in the background (*secondary*). Eleven labels are attached to either domains in each image. A *cannot_distinguish* label is used when the annotator is less than 30% sure of the sound source, while a *custom* label is assigned when they are able to identify a source not in the label list.

2. Additional Beamforming Details

Measuring Environment Sounds Note that the 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

$$C_{mn} = \mathbf{E}[(\tilde{p}_m(\omega) + \zeta_m(\omega))(\tilde{p}_m(\omega) + \zeta_m(\omega))^*], \tag{1}$$

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

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

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. As such, removing the main diagonal elements from the cross-spectral matrix is reducing the effects of measurement errors.

Spatial Aliasing From Sparse Measurements As the microphones are sparsely placed, the measured pressure from an acoustic source is spatially aliased, corrupting the beamforming output map. An ideal infinitely large microphone array response to a single acoustic source is a dirac delta on the beamforming output. However, the real response is a *point spread function* (PSF). This causes the beamforming output map of acoustic sources to be convolved with the microphone array PSF, and the PSF shape depends on the noise source location, the acoustic camera aperture and the focus plane, analogous to a conventional imaging camera. The corresponding beamforming output map from a finite PSF acoustic camera, following Eq. 3 in the main manuscript, is therefore

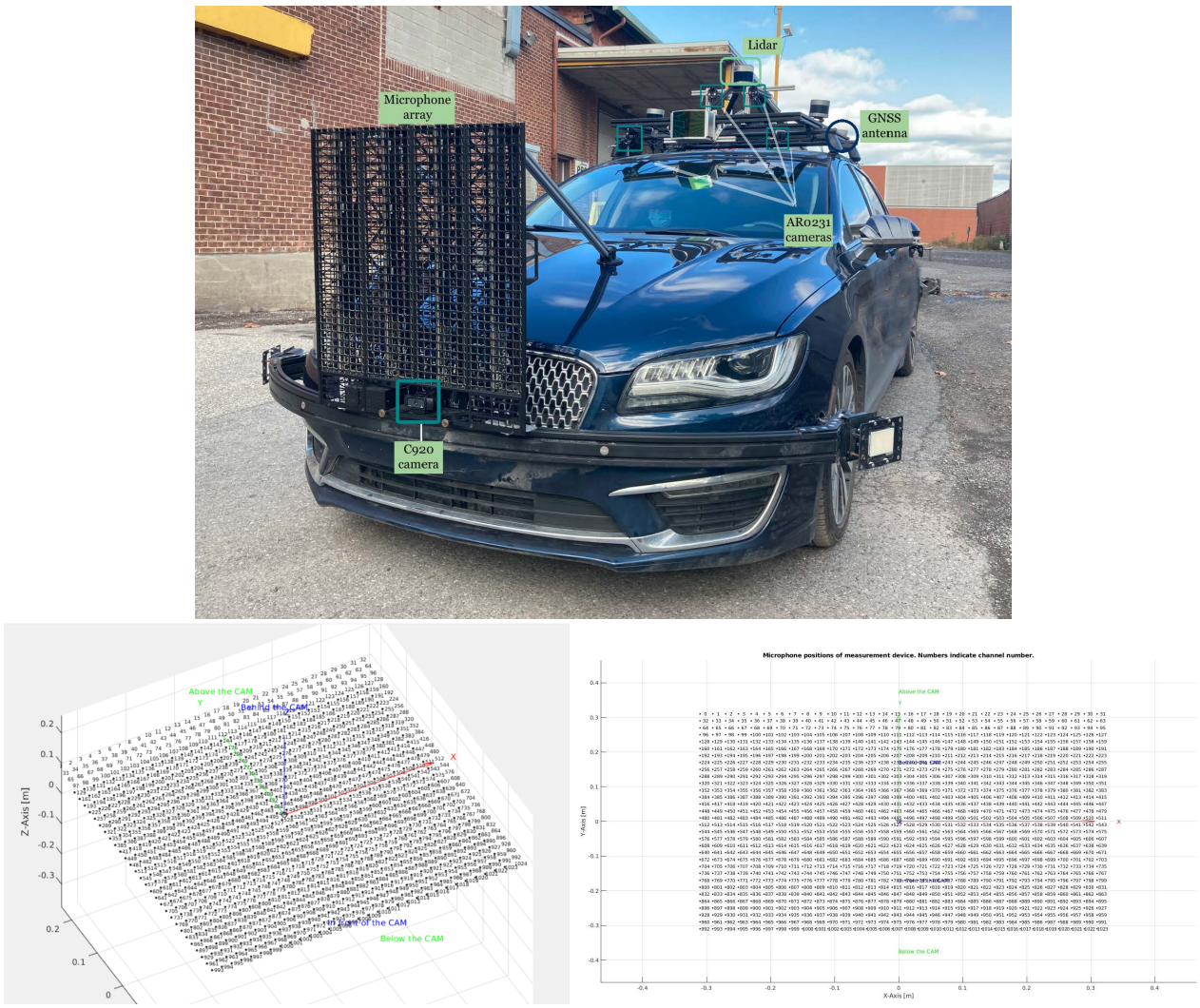

Figure 1. We built a prototype capture vehicle [*Top*] for acquiring our acoustic beamforming dataset. Our vehicle consisted of a 1024 channel microphone array, as shown in the schematic [*Bottom*], used for road traffic sound measurements.

$$\mathrm{BF}(t, \vec{x}_s) = \frac{4\pi}{M} \sum_{m=0}^{M} \Big[p_m(\vec{x}_s, t + \Delta t_m) |\vec{x}_m - \vec{x}_s| \Big] * h, \tag{3}$$

where $*$ is the convolution operator and $h = \mathrm{PSF}(\vec{x}_m, \vec{x}_s)$. We compensate for this spatial aliasing which corrupts the beamforming map as discussed next, and reliably perform automotive vision tasks using this sensing modality.

PSF Deconvolution As described in Eq. (3), the acoustic beamforming map is corrupted by a PSF kernel determined by the focus plane and the aperture of the microphone array. We simulate this PSF using a synthetic point source δ as

$$\mathrm{PSF}(x, y, z) = f_{\mathrm{BF}}(\delta(x, y, z), \mathbf{F}), \tag{4}$$

where we set x, y to be at the center of the field-of-view and we set z to be the focal plane as described in the main manuscript. The beamforming map produced by f_{AE} can then be deconvolved with the synthetic PSF by using a Wiener Filter layer to mitigate the artifacts due to PSF corruption.

Beamforming of Acoustic Signals In order to determine the unknown Δt_m, we first assume that the sound source resides within a plane that is co-planar to the microphone array and at a distance z away. This distance z can be thought of as the

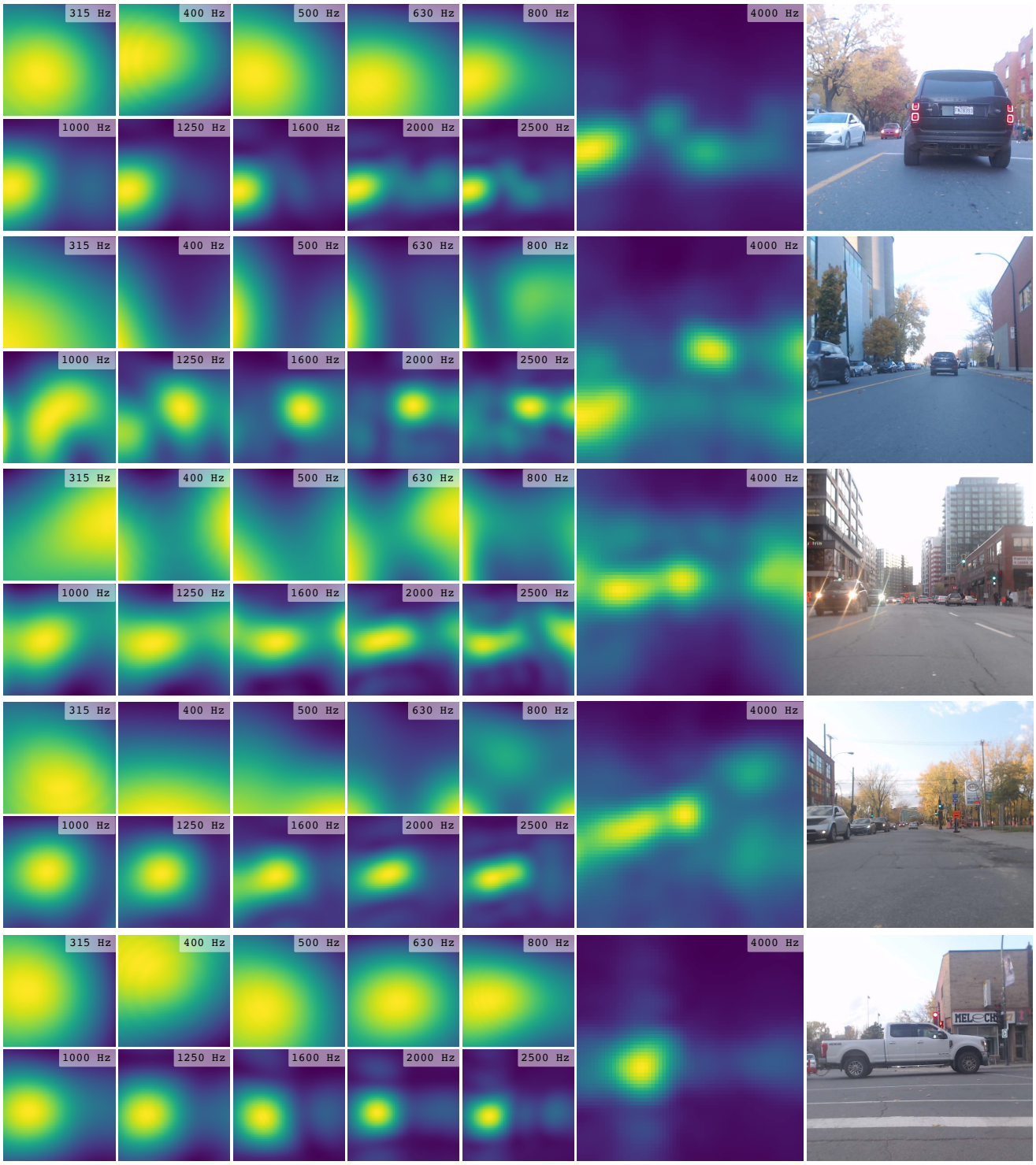

Figure 2. Qualitative examples of different RGB scenes and their corresponding beamforming maps across measured frequencies.

focal plane of the acoustic camera. We then construct the beamforming map BF by raster scanning and computing Eq. (3) at every point on the plane using the corresponding time delays. Note that the signals from a given sound source as captured by the microphones are similar in wave form, but show different time delays and hence phases. Therefore, if the point under consideration coincides with the noise source, the computed delay-and-sum signals are in phase and constructively

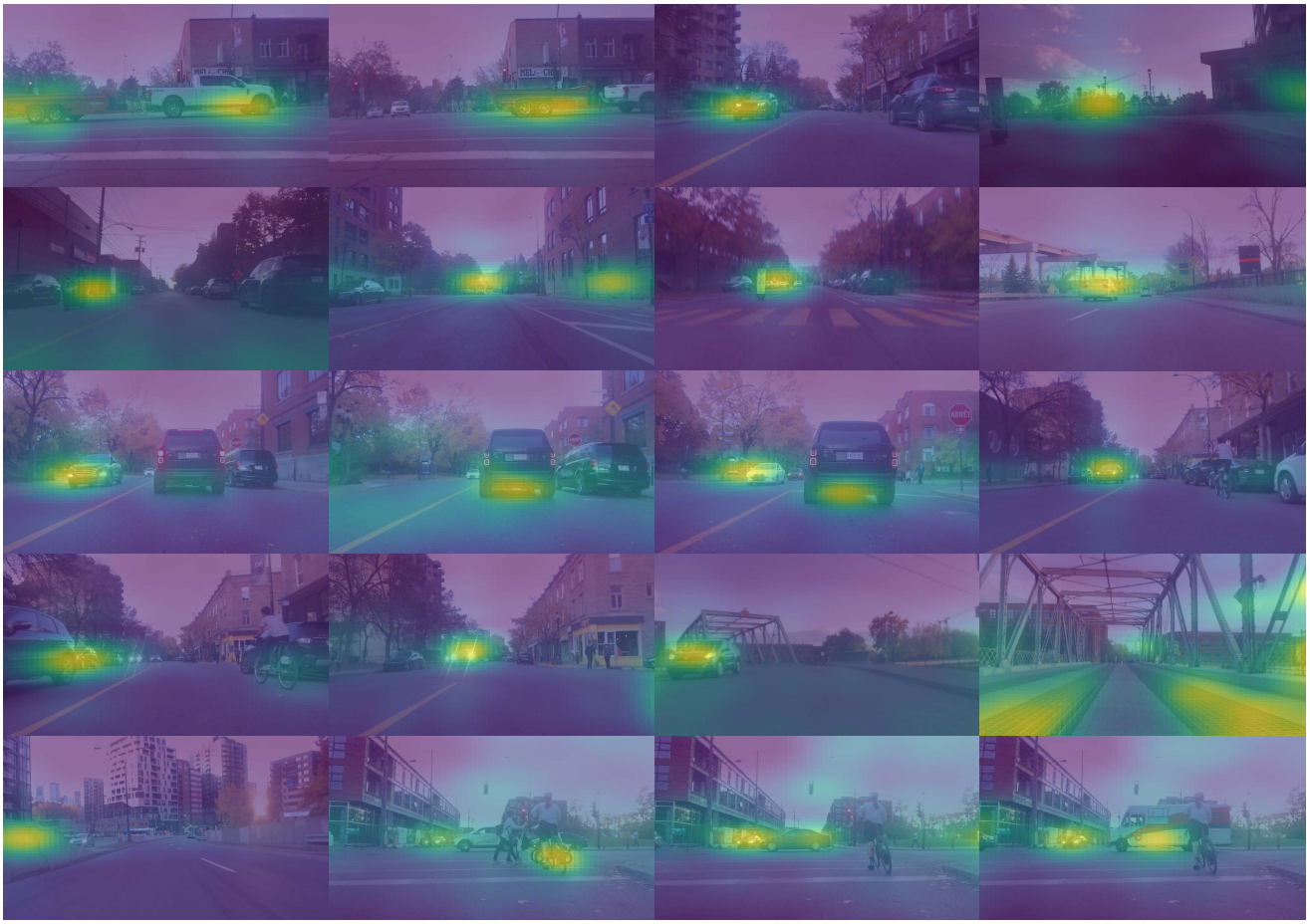

Figure 3. Examples of diverse scenes from the dataset, with the corresponding beamforming maps overlaid. Our dataset consists of diverse scenes from day and night, and residential neighborhoods and urban downtown.

interfere. Otherwise, in the absence of a sound source, a phase mismatch will occur causing destructive interference. The final beamforming map of multiple sound sources is the superposition of those corresponding to each individual sound source.

Processed Beamforming Projections in Dataset The raw sound pressure data was used for creating beamforming maps in the world coordinate frame and were then projected onto the image plane of the C920 camera located below the array. The beam formations were computed using delay-and-sum beamforming in the time domain with auto-power removal as described in Section 5 of the main paper. Beamforming is done on a grid of $5m \times 5m$ with a resolution of 64×64 at a distance of 10 m from the microphone array center, in front of the car. The maps were calculated at a frame rate of 25 Hz for the 11 frequency bands shown in Figure 4c of the main paper. 282 Hz is the lower band limit of the first one-third octave band after the 250 Hz lower bound specification of the microphone array, described earlier in the section. A range of 282 Hz to 5000 Hz was chosen for beamforming which enabled localization of low frequency noise sources that are more prevalent in the environment [1]. Since most road and tire noise frequency spectra display a pronounced peak in the range of 700 Hz - 1300 Hz [2], using one-third octave bands for 10 of the 11 ranges (non-highlighted rows in Figure 4c of the main paper) helped to map a larger focus on noises around this frequency range.

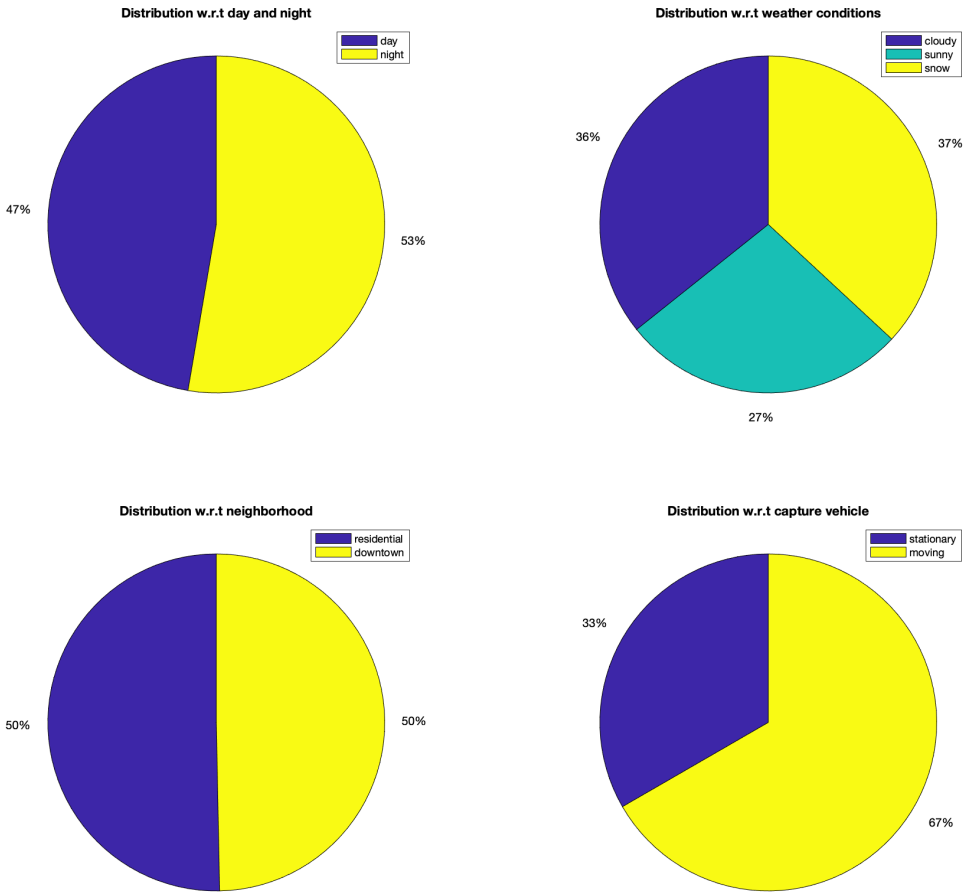

Figure 4. Diversity and distribution of data in our long-range acoustic beamforming dataset. Our dataset consists of high resolution microphone array recordings along with pre-processed beamforming maps at several frequencies, RGB camera captures, lidar, GPS and IMU measurements. The dataset covers both day and night scenes, variety of weather conditions, neighborhoods and scenery and includes capture sequences with our prototype capture vehicle moving as well as being stationary.

Table 2. Additional experiment validating utility of beam forming over RGB-only detection method. AP50 results are reported.

	low res RGB+Sound (Day)	RGB-only (Day)	low res RGB+Sound (Night)	RGB-only (Night)
AP50	78.1	79.4	61.4	37.2

3. Additional Multimodal Object Detection from Acoustic Signals

We provide additional object detection results in Figs. 6 and 7. Specifically, Fig. 7 shows multimodal object detection where RGB camera based detections fail completely, validating the proposed approach. Fig. 6 shows detections using only acoustic beamforming signals to validate the complementary information present in sound pressure where RGB signals fail.

We additionally investigate object detection using a combination of low-resolution RGB and acoustic signals, and report the AP50 scores in Table 2. We downsample the RGB images by $16\times$ and fine-tune the downstream detection network. We notice that while the AP50 scores for day scenes are comparable to full resolution RGB-only detections, we see a significant increase of about 64% in AP50 scores for the night scenes. This trend can also be observed with sound-only detections (i.e., without any image input) as reported in Table 2 of the main manuscript. These experiments again validate that the acoustic signals carry complementary information to that of RGB images for automotive driving tasks.

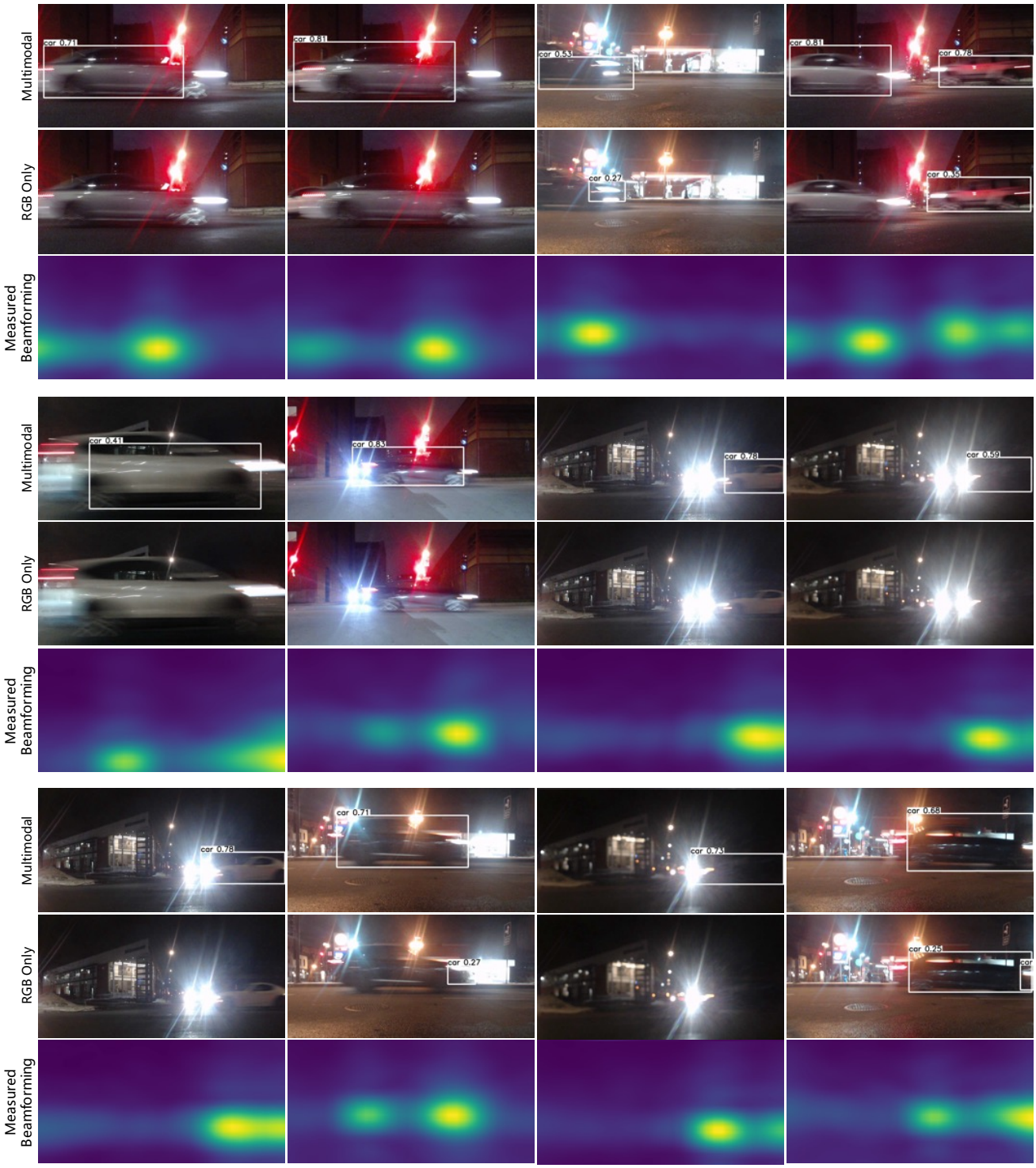

Figure 5. Additional object detection results on unseen scenes. We show the corresponding beamform maps (4000 Hz) for each RGB frame. The detrimental impact of glare and motion blur in low light conditions on RGB detection can be seen in the third row. Detection using beamforming maps circumvents these problems.

Figure 6. Additional object detection results on unseen scenes. We show the corresponding beamform maps (4000 Hz) for each RGB frame. While detection using beamforming maps in daytime is not as good as RGB detection, we are still able to achieve accurate detection.

Figure 7. Additional object detection results on unseen scenes. We show the detections on unseen scenes with RGB-only, Beamforming-only and Multimodal methods. As can be seen, sound pressure signals via beamforming carry complementary information that results in superior detections for scenes where RGB images fail. Overall, multimodal detection method outperforms RGB-only and beamforming-only detections.

4. Results for Multimodal Future Frame Prediction

We report additional results that show that beamform maps when combined with RGB image data can be used as a cue for multimodal future frame prediction. Figures 8, 9, and 10 show these multimodal future frame predictions on a variety of scenes using the Pix2PixHD network described in the main manuscript (labeled "Pressure2Pix" in the figures), compared with baselines. Our best-performing approach predicts the RGB frame at time t using the RGB frame at time $t-1$ stacked with audio maps at time steps $t-1$ and t; we compare this approach to a method predicting RGB at time t from just the RGB at time $t-1$ (labeled "Pressure2Pix (without audio)" in the figures), and an optical-flow-based approach which warps the RGB at time $t-1$ using the optical flow from $t-2$ to $t-1$, warped by itself, to produce the RGB at time t.

As explained in the main manuscript, these predictions are performed in a cascaded fashion, with the prediction for one frame acting as the input for predicting the next frame. In other words, the whole predicted sequence is based only on the seen frame at $t=0$ (except for optical flow, in which case two previous frames are needed to produce a starting flow). In the case of multimodal Pressure2Pix with audio, ground truth audio maps are included in future predictions, as this information is available at a much higher sampling rate than RGBs (motivating the use of beamforming data for video upsampling).

For all scenes presented here, we observe that our approach, utilizing audio information, confers three main qualitative advantages over other approaches. First, unlike optical-flow-based prediction, our approach does not produce warped backgrounds (see, e.g., column 1 of Figure 9, in which optical flow progressively warps the streetlights in the background). Rather, when trained on a static scene (i.e. in which the capture vehicle does not move), the Pressure2Pix network is able to learn which elements of the scene remain constant, and to avoid modifying those areas of the image during inference: see columns 2 and 3 in Figure 9, and the accurately synthesized background in columns 2 and 3, row 5, of Figure 8, in contrast to column 1, in which optical flow starts to distort the building on the left. Second, we observe that optical flow fails to adequately displace the vehicle of interest in the scene, again a consequence of a lack of information about the behavior of background pixels—the algorithm cannot predict that sections of the background will be obscured in future frames as the car moves forward. See column 1 of Figures 8, 9, and 10 for examples of this; when predicted using optical flow, the front of the car is not synthesized any further along its path. Finally, we note the ability of the full multimodal Pressure2Pix network to retain the approximate shape of the vehicle during inference, in contrast to both optical flow and Pressure2Pix without audio. As can be seen especially in Figures 8 and 10, our network synthesizes accurate wheel and roof shapes when provided audio maps as additional input; compare columns 2 and 3 in row 5 of Figure 8, or in row 3 of Figure 10.

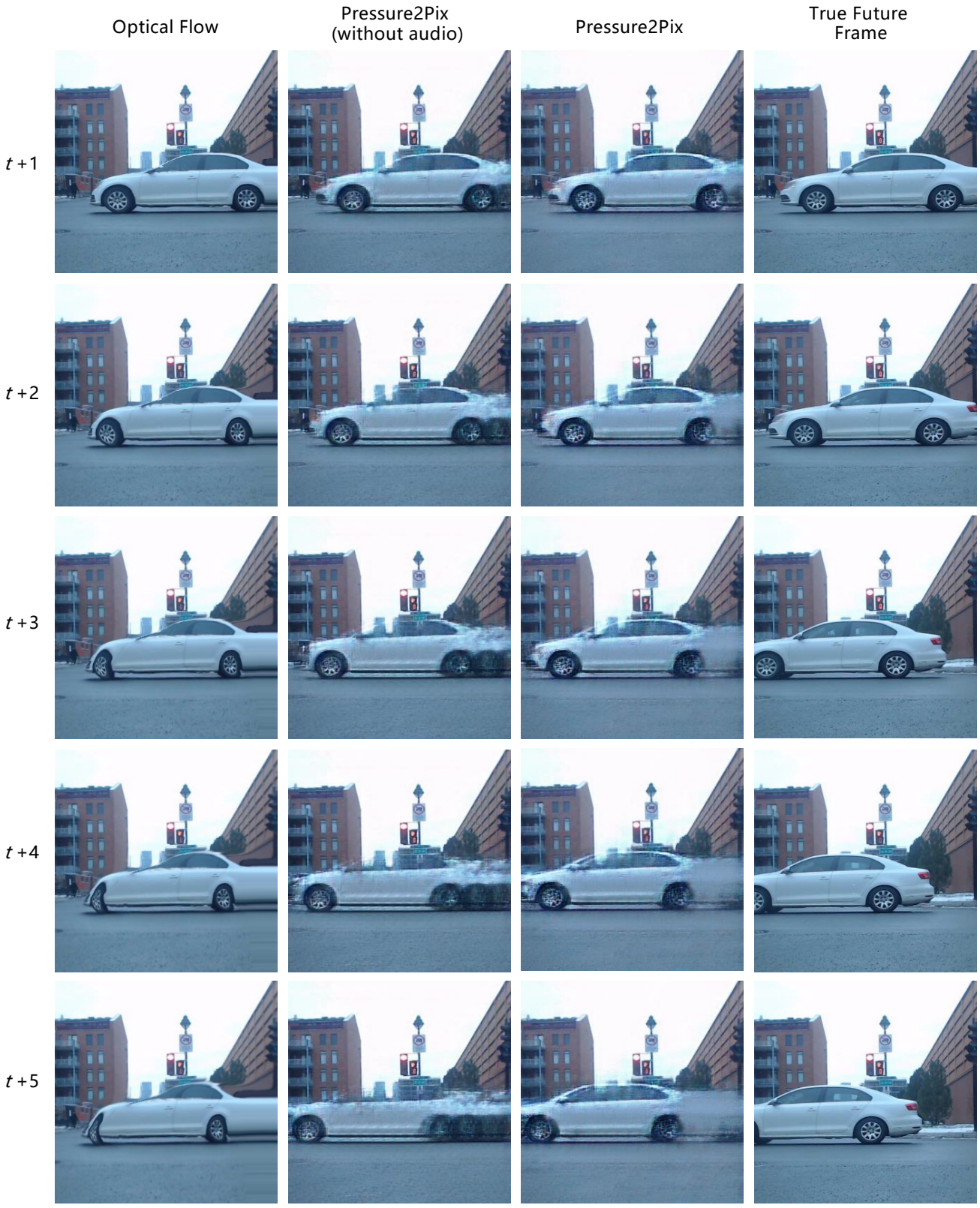

Figure 8. Additional future frame prediction results. In this scene, note that Pressure2Pix with audio is better able to synthesize a plausible car, preserving wheel shape and the top of the car.

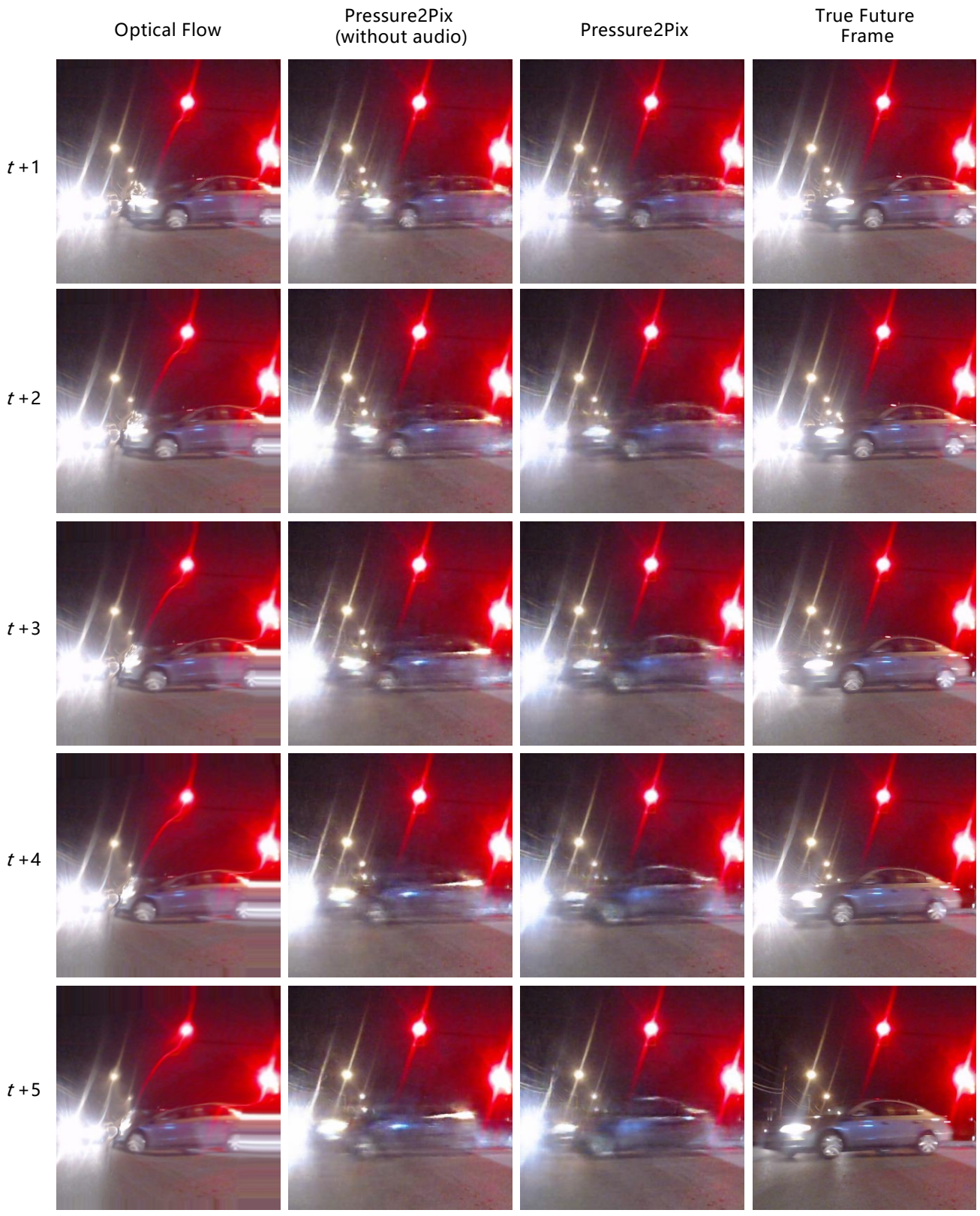

Figure 9. Additional future frame prediction results, on a challenging night scene. Note the warping of the background in optical flow predictions. Furthermore, we again see Pressure2Pix synthesize a more realistic car when provided audio information than when not, especially for predictions far in the future.

Figure 10. Additional future frame prediction results. Pressure2Pix with audio synthesizes a more accurate car than Pressure2Pix without audio, while displacing the vehicle in the right direction.

References

[1] R. Cousson, Q. Leclere, M.-A. Pallas, and M. Berengier. Identification of acoustic moving sources using a time-domain method. 2018. 5

[2] U. Sandberg. The multi-coincidence peak around 1000 hz in tyre/road noise spectra. 5