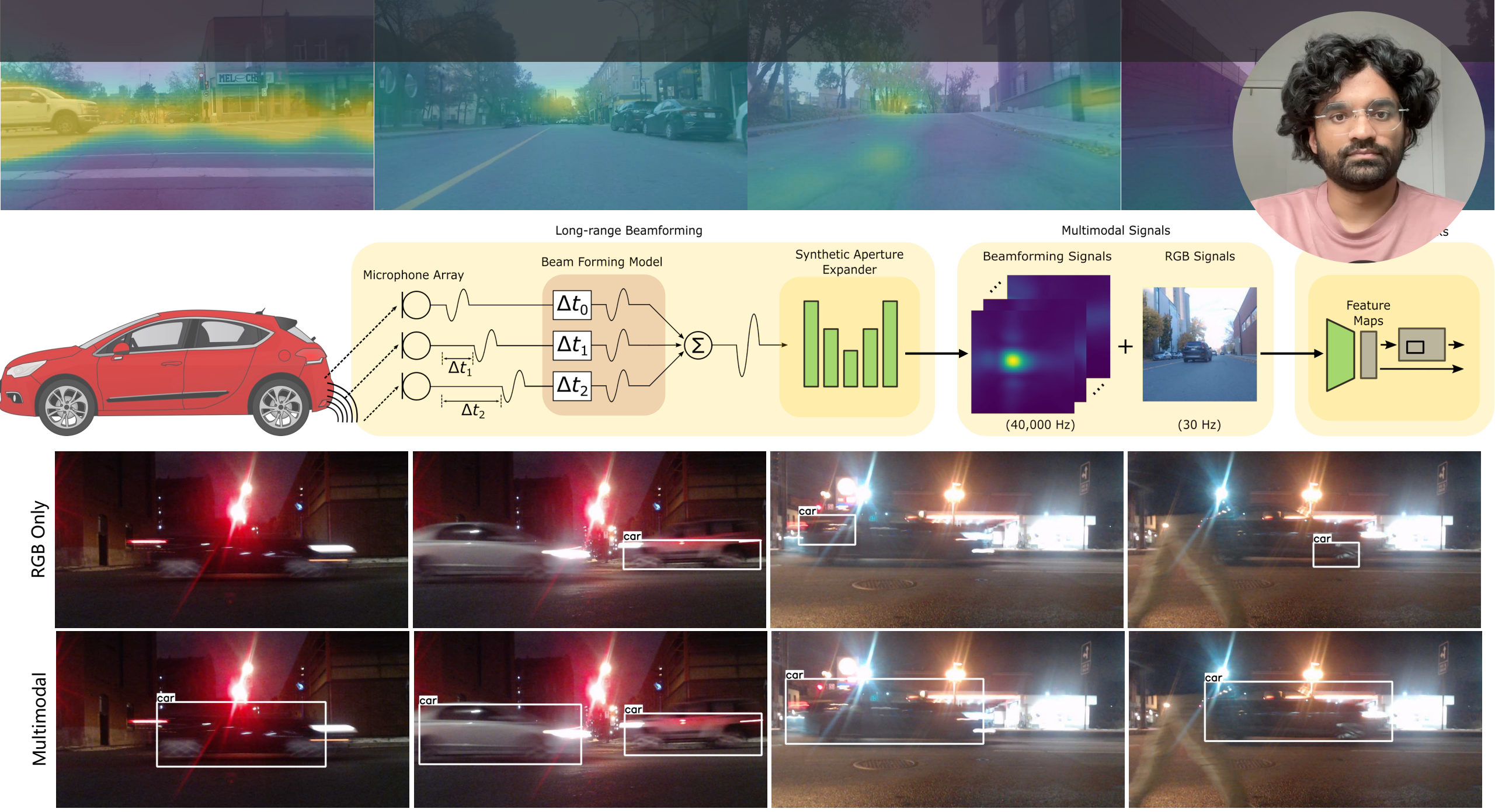
Seeing With Sound
Acoustic Beamforming for Multimodal Scene Understanding

Praneeth Chakravarthula, Jim Aldon D'Souza, Ethan Tseng, Joe Bartusek, Felix Heide

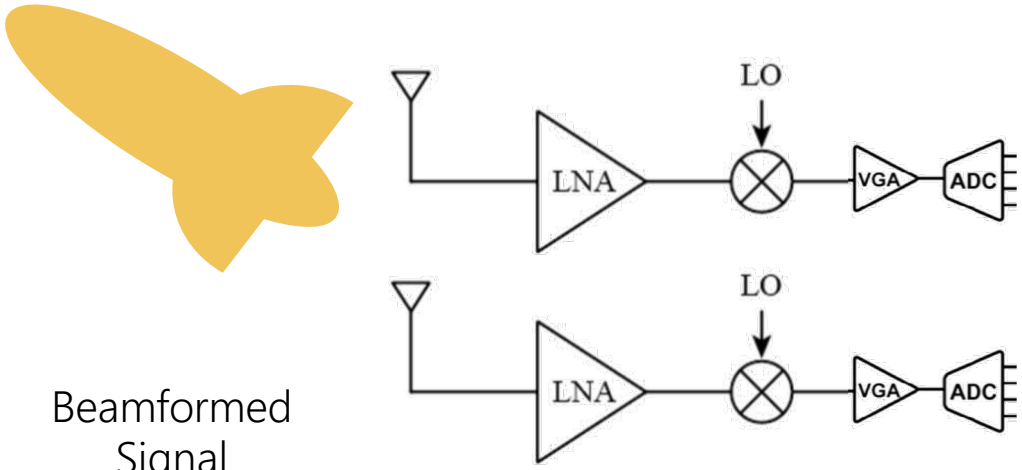
PRINCETON
COMPUTATIONAL IMAGING LAB

https://light.princeton.edu/publication/seeingwithsound/


Sound Source


Long-range Beamforming
Microphone Array
Beam Forming Model
Δt_0
Δt_1
Δt_2
Σ
Synthetic Aperture Expander
Multimodal Signals
Beamforming Signals
RGB Signals
(40,000 Hz)
(30 Hz)
Feature Maps
RGB Only
Multimodal
car

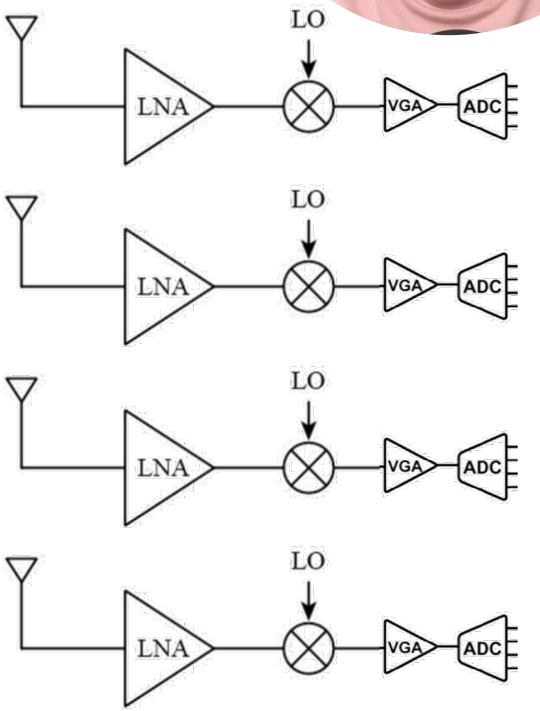
Acoustic Beamforming



Beamformed Signal

Microphones

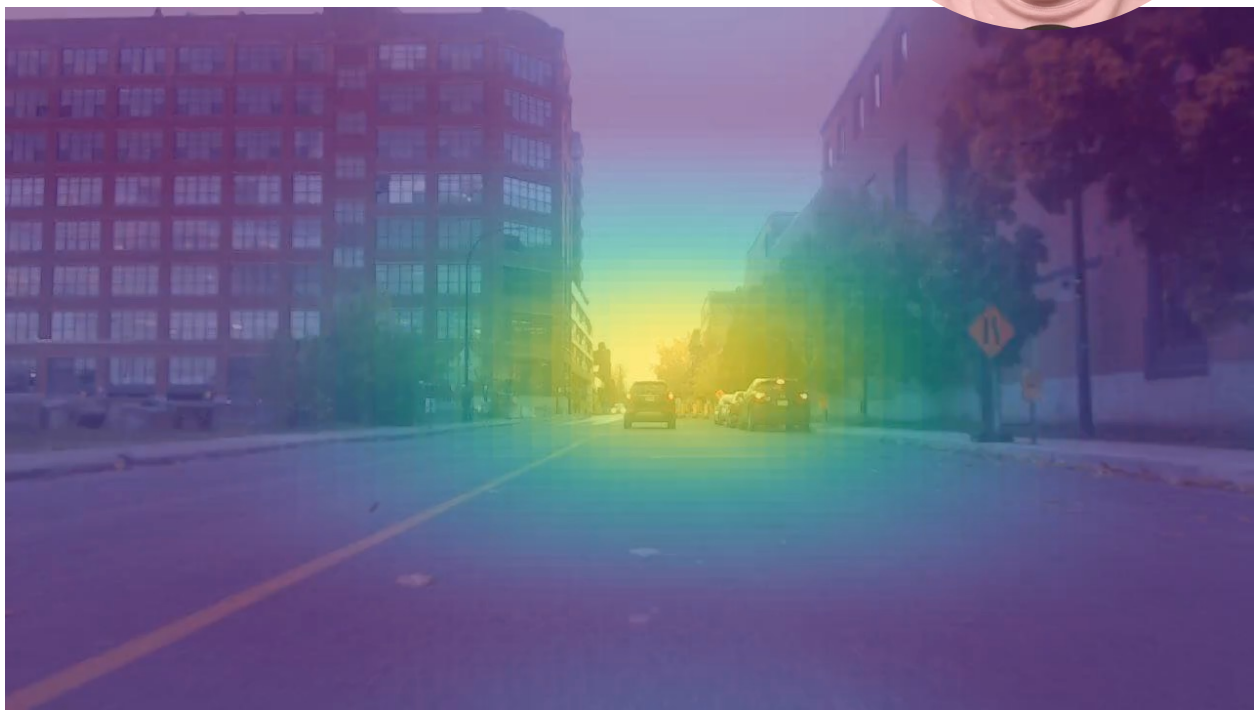
Beamformed Signal



Microphone Array

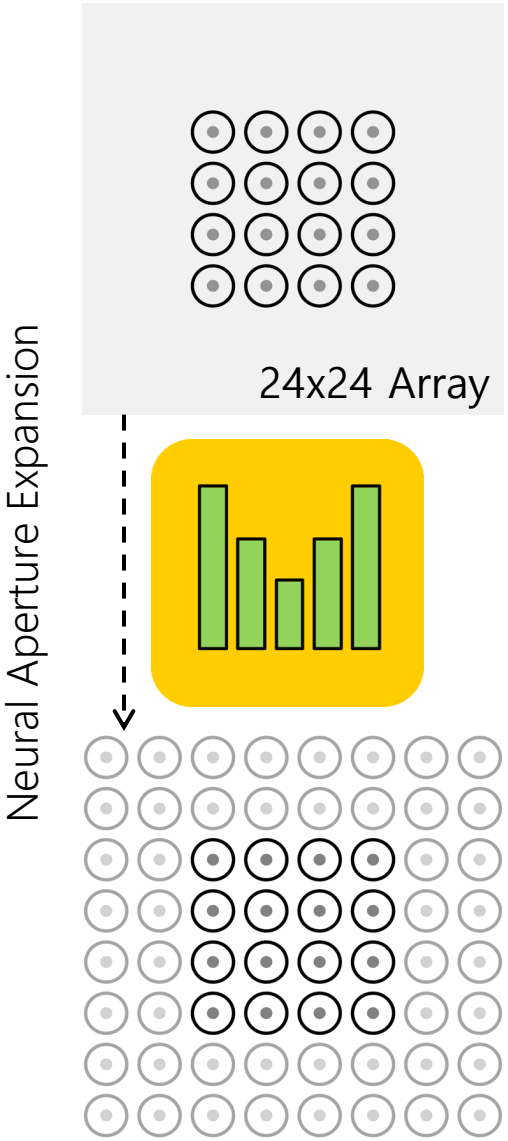
Prototype Vehicle

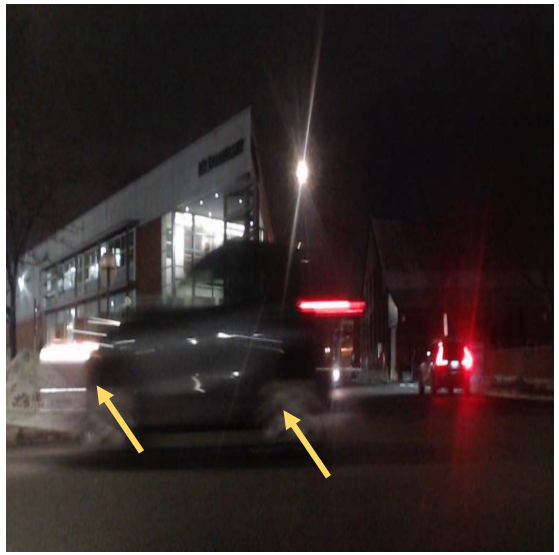
Beamforming In-the-wild

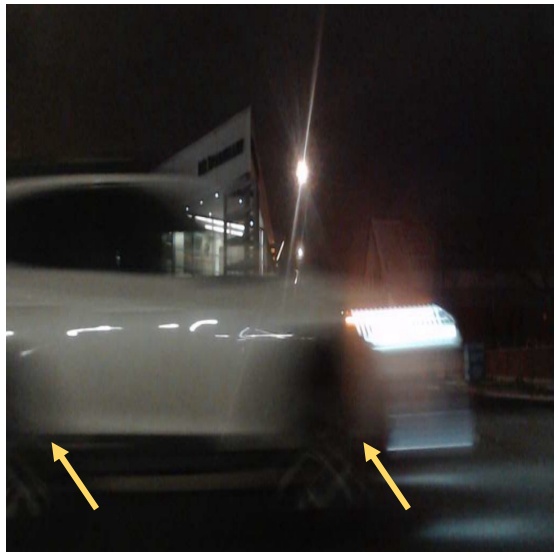


Beamformed Sound Pressure Signal Overlayed on Traffic Video

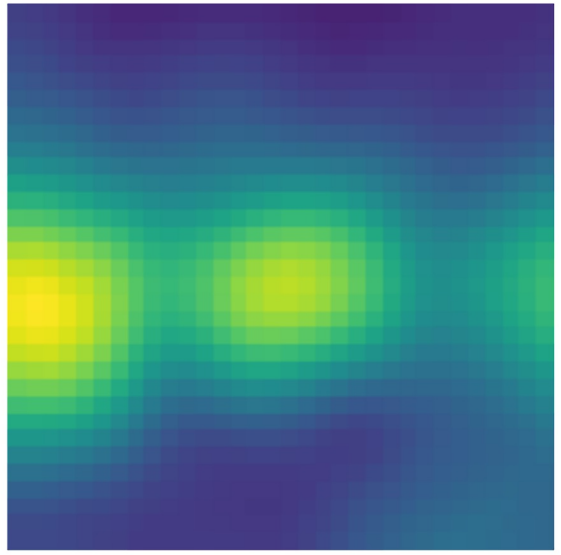
Neural Aperture Expansion

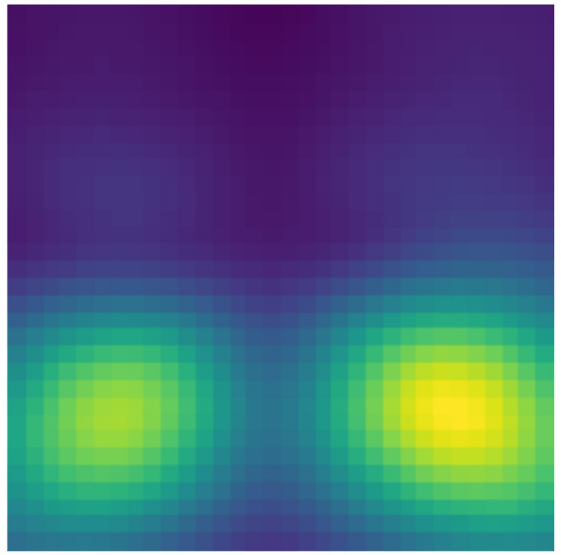




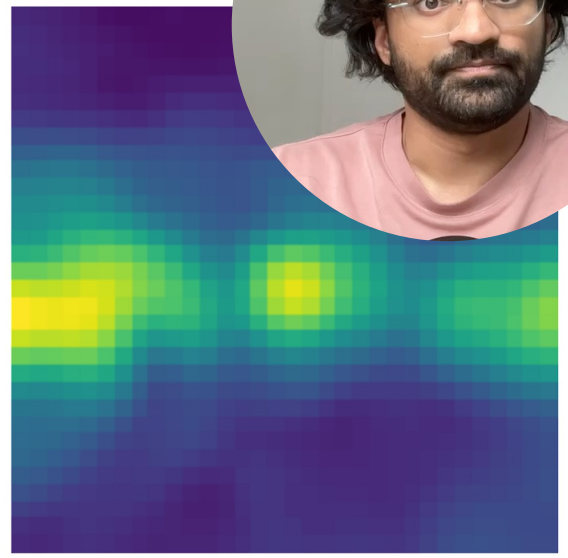


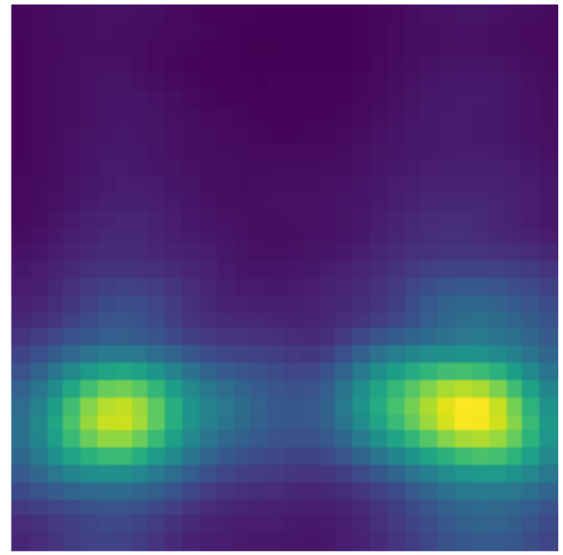
Scene



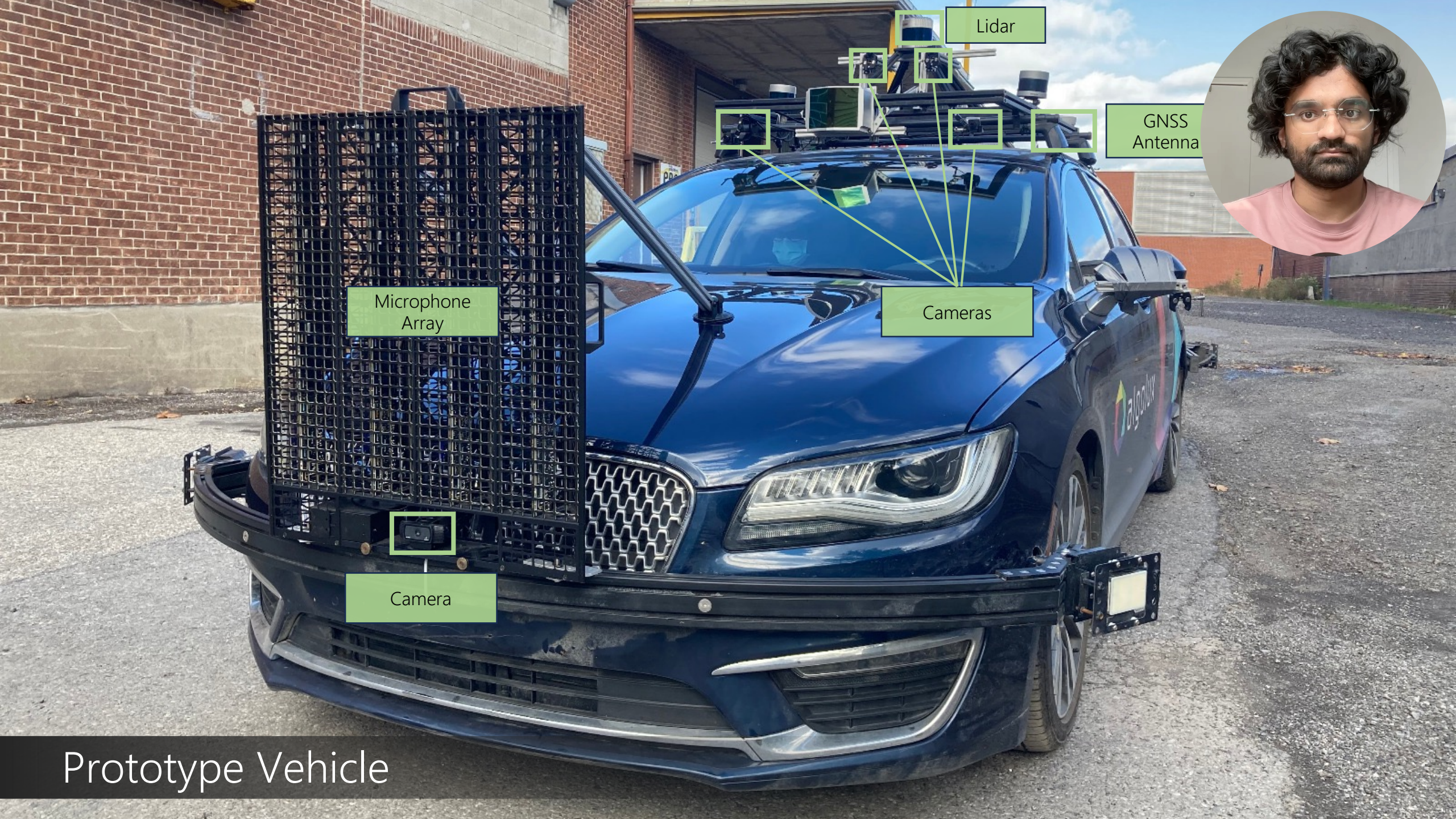


Small Aperture





Synthetic Large Aperture

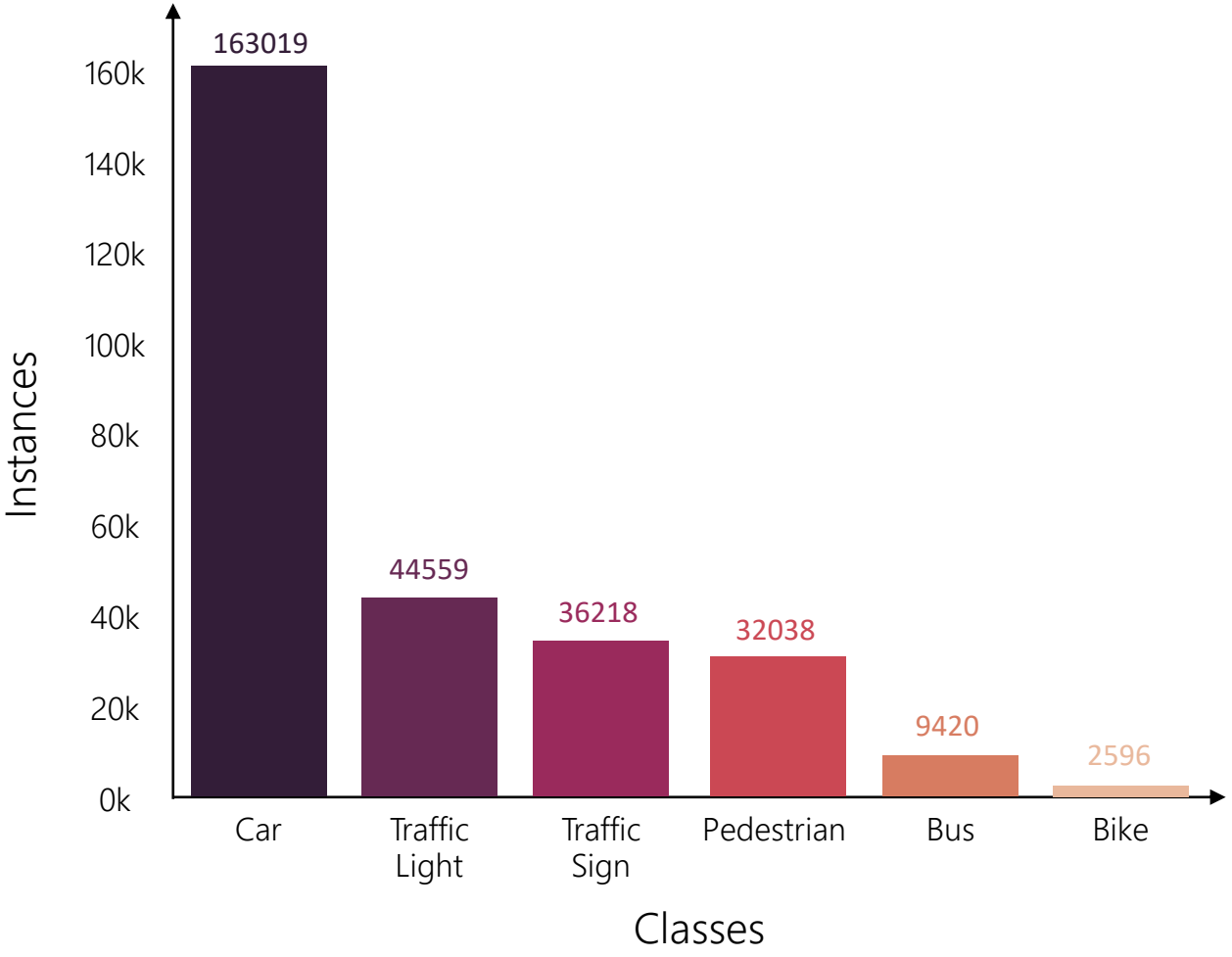


Prototype Vehicle

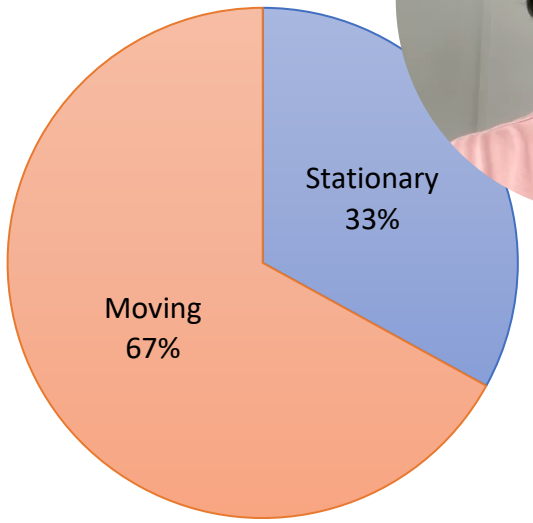
Dataset Acquisition

Total: 66 km

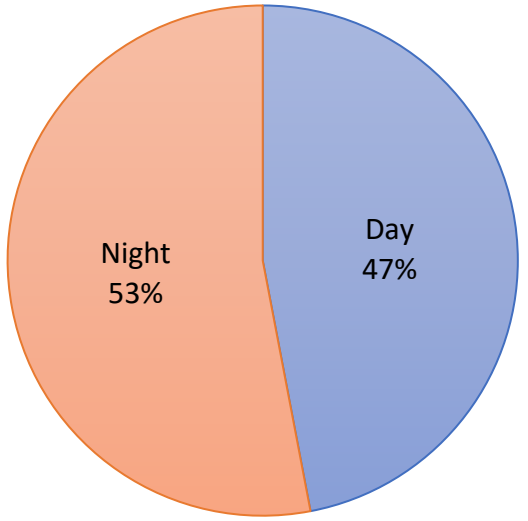
Multimodal Acoustic Beamforming Dataset



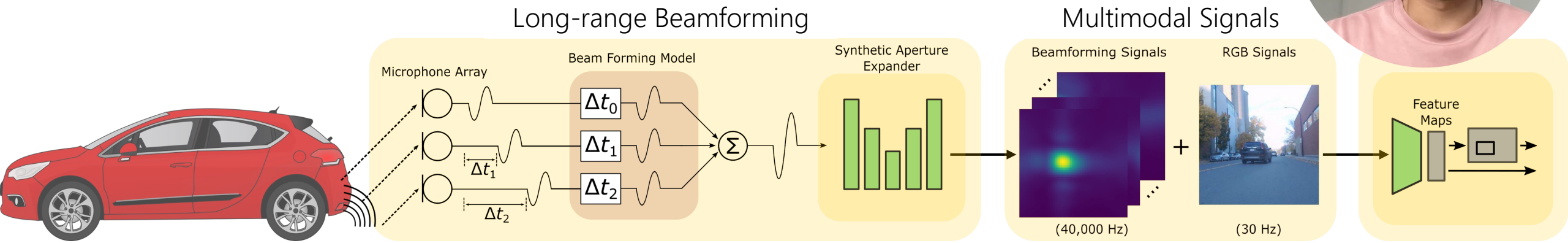
Distribution w.r.t. Capture



Distribution w.r.t. Day and Night



Multimodal Vision and Acoustic Detection



Multimodal Object Detection

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

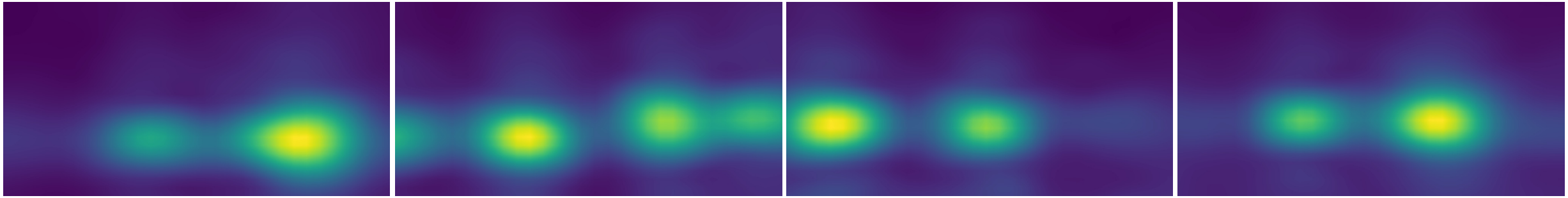
Frame Interpolation and Extrapolation

	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

Multimodal Sensing where Cameras Fail

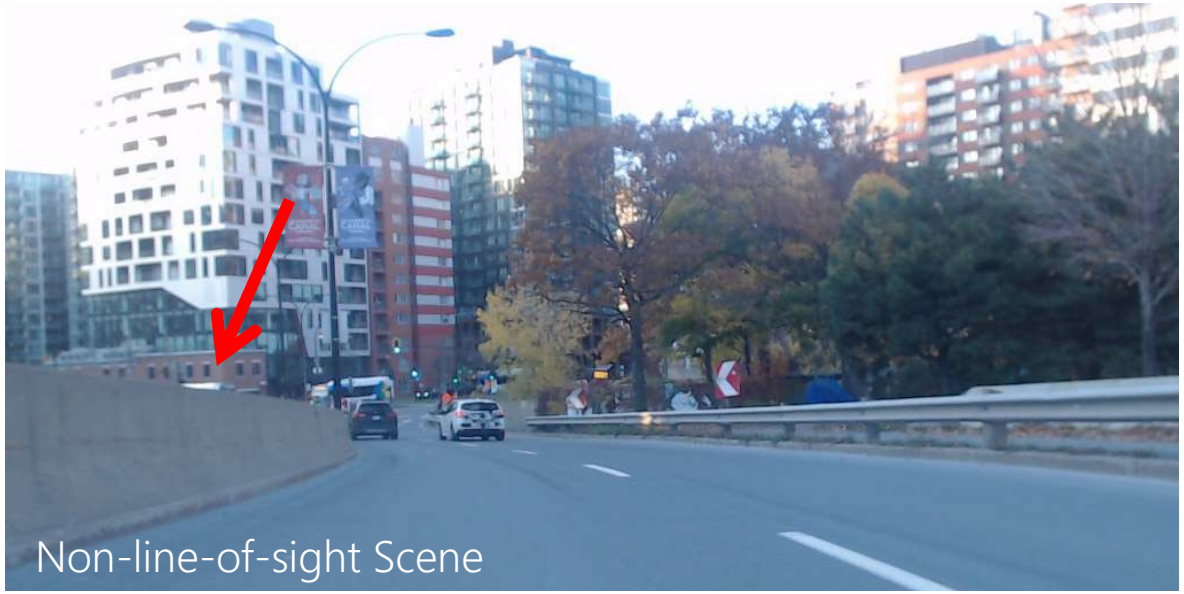
RGB Only

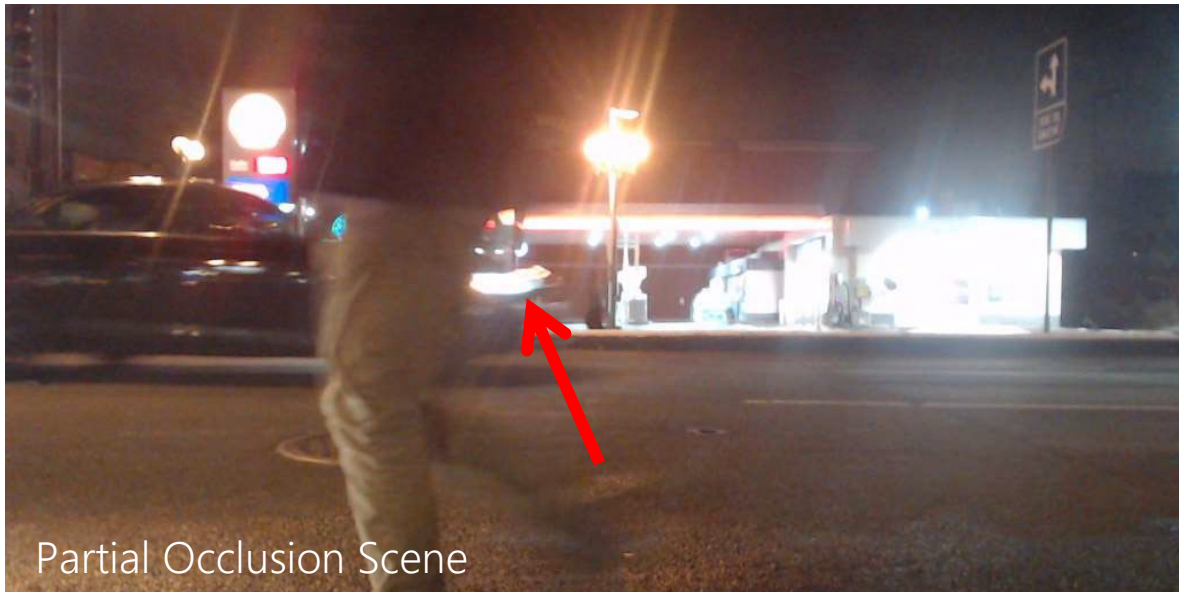
Measured Beamforming



Multimodal

Non-line-of-sight Detection Where RGB Cameras Fail





Find Dataset Here

Project Page