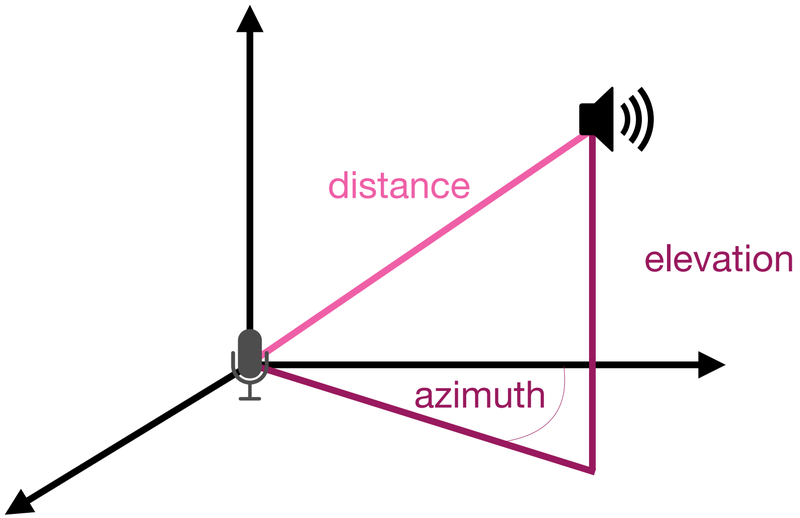
Urbansas: Urban Sound & Sight dataset and benchmark

Magdalena Fuentes[1], Bea Steers[1], Pablo Zinemanas[2], Martín Rocamora[3], Luca Bondi[4],
Julia Wilkins[1], Qianyi Shi[1], Yao Hou[1], Samarjit Das[4], Xavier Serra[2], Juan Pablo Bello[1]

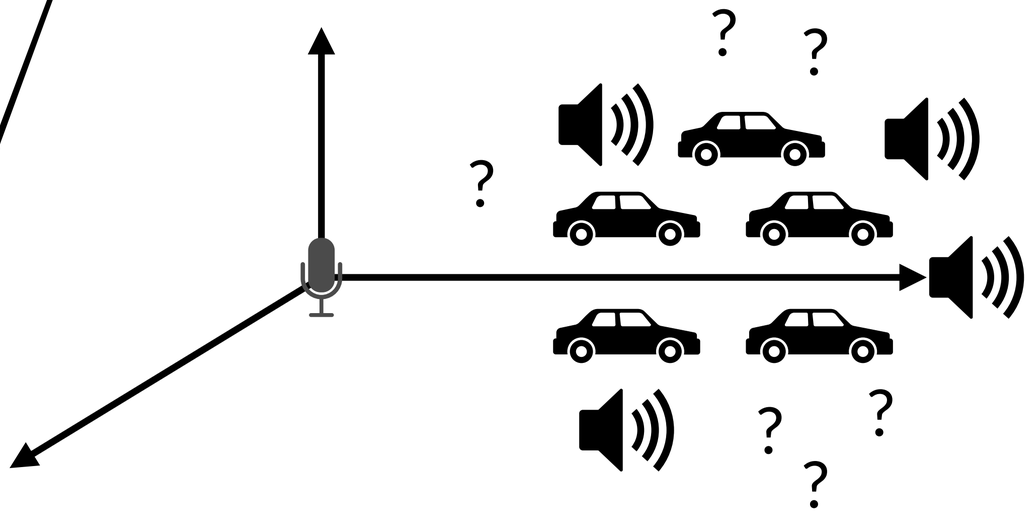
[1]New York University, USA
[2]Universitat Pompeu Fabra, Spain,
[3]Universidad de la República, Uruguay
[4]Bosch Research, USA

https://magdalenafuentes.github.io/
mf3734@nyu.edu

We use location of sound sources and their motion to navigate the world around us.

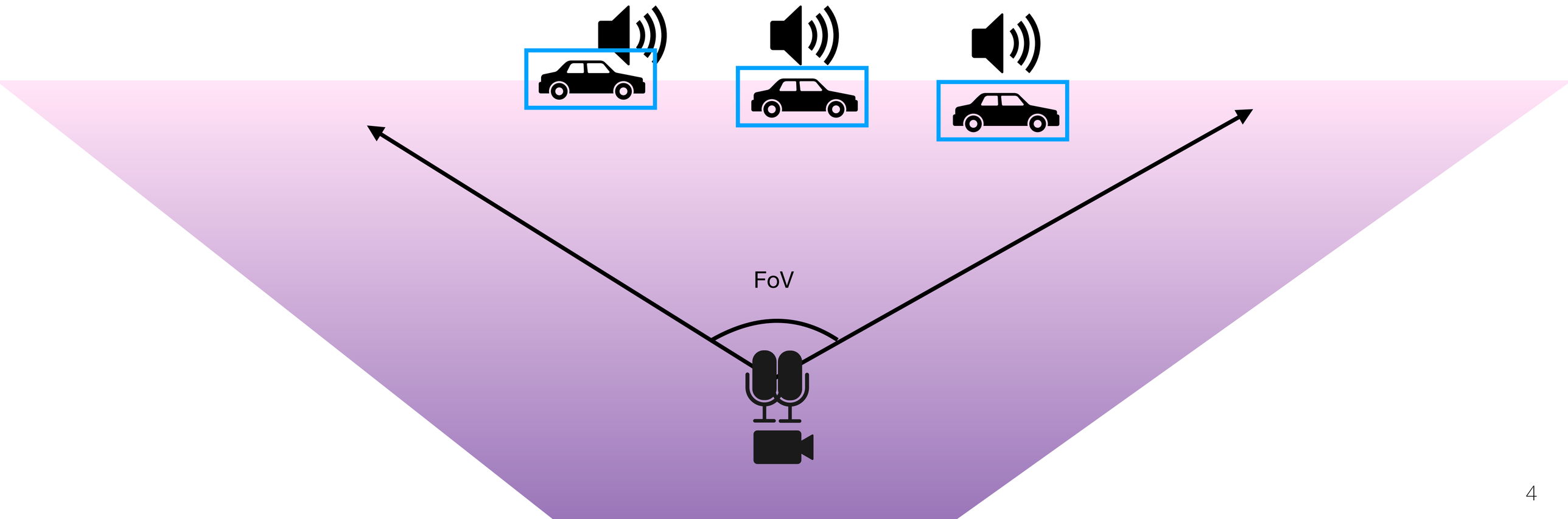


Synthetic unrealistic conditions

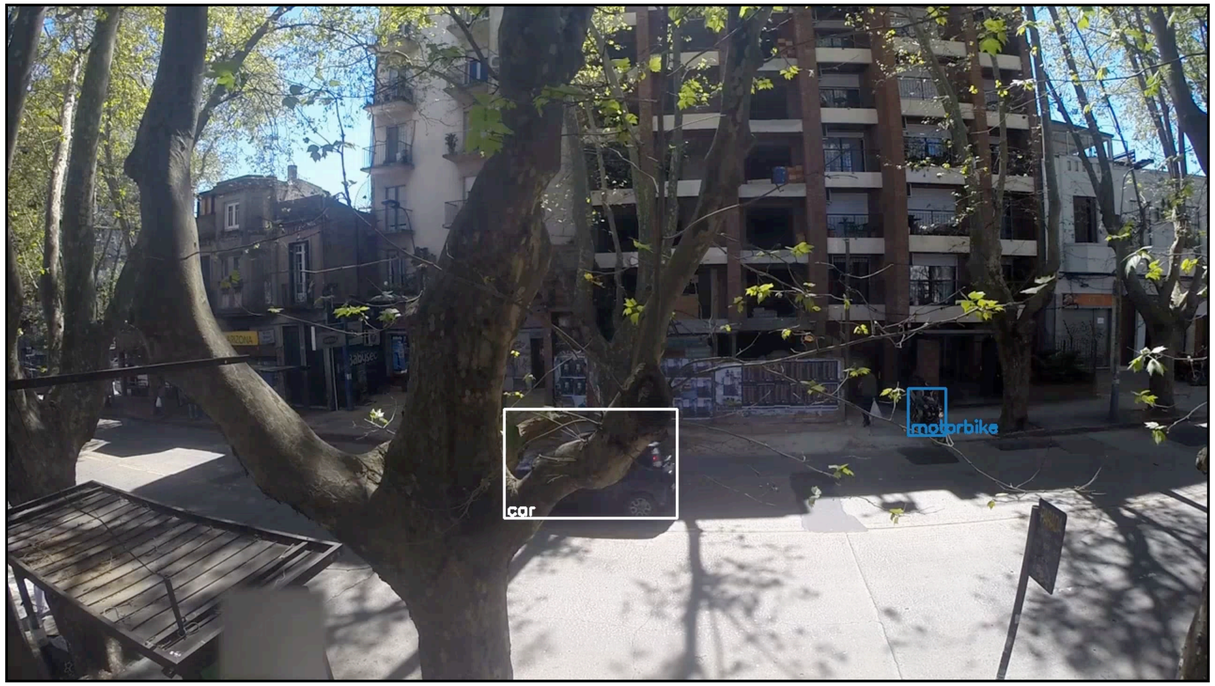


Real-world applications: how?

Inferring vehicles position from video



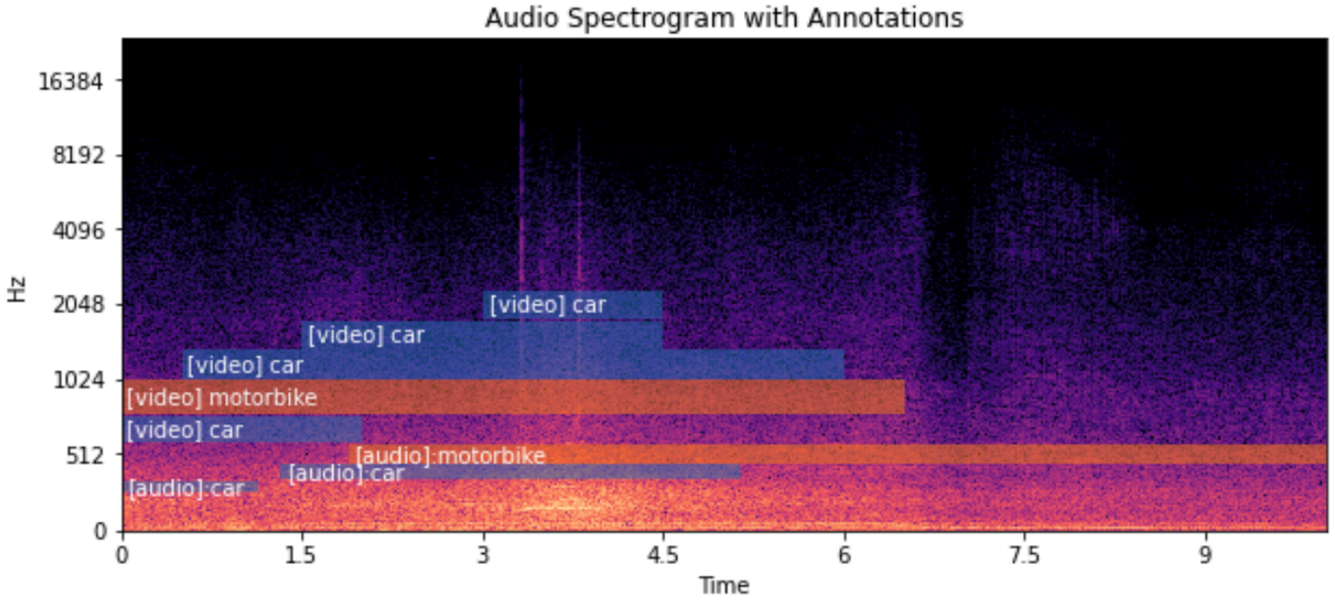
Urban Sound & Sight



Data

Video

Stereo Audio



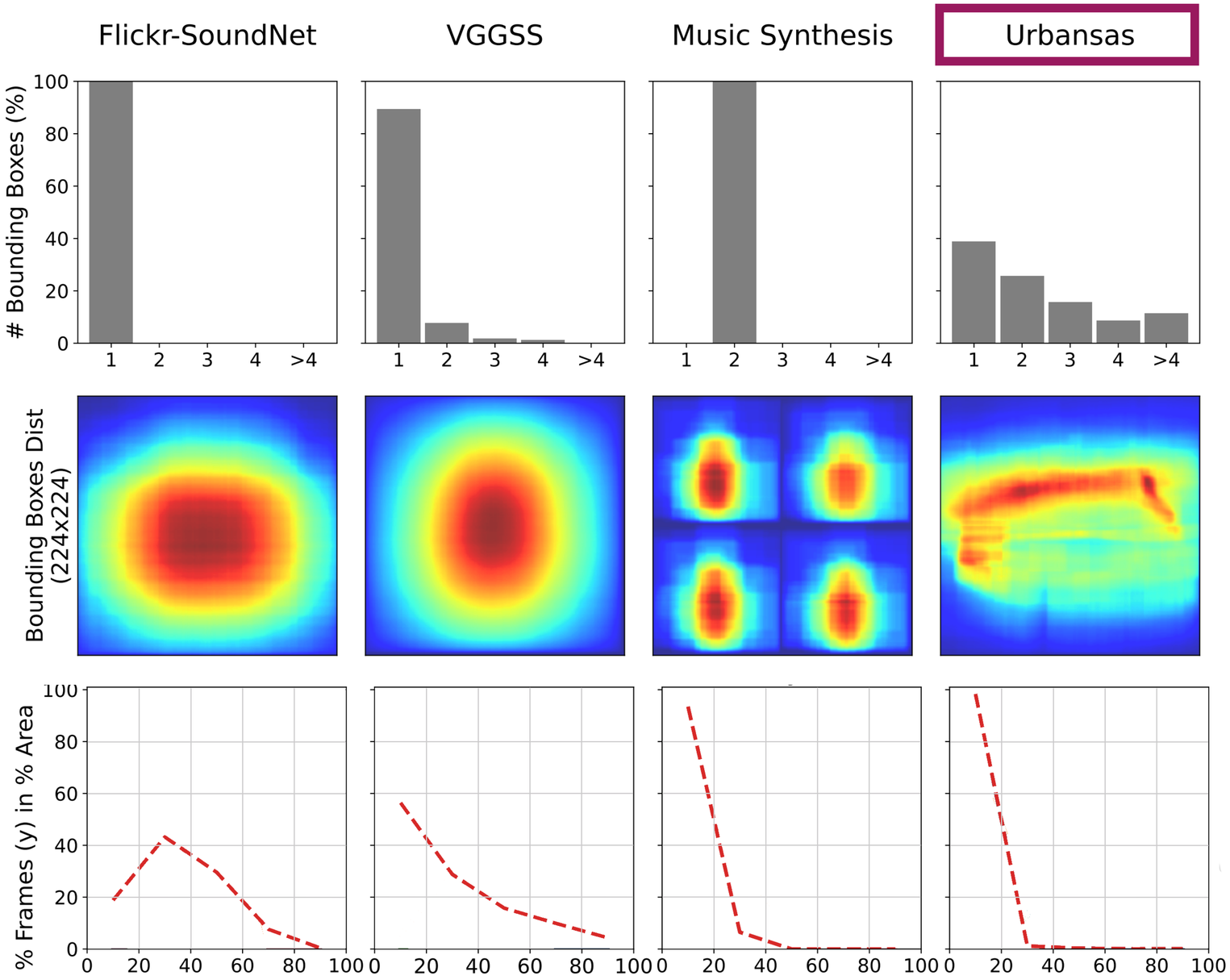
Annotations

Bounding boxes

Audio labels

Clip metadata

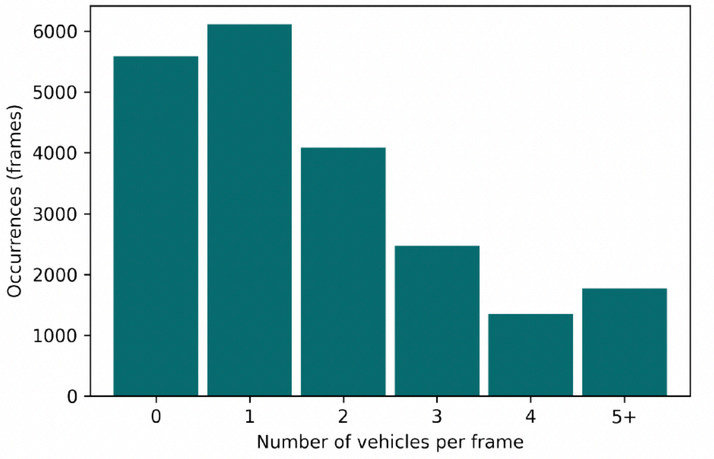
Urbansas has variety in bounding box count, center position, and area.



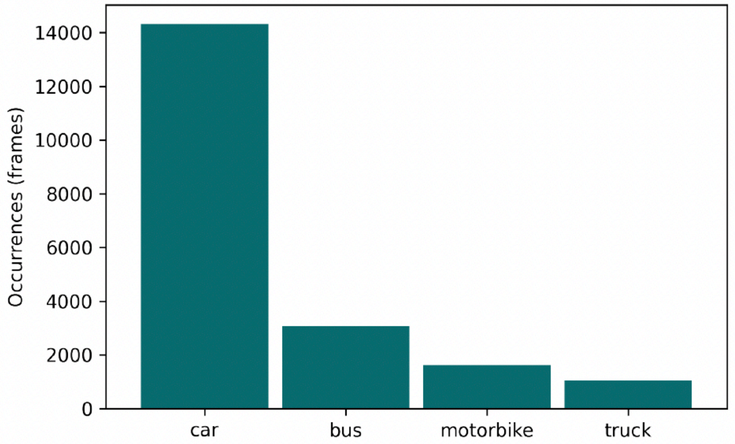
Wu, H. H., Fuentes, M., Seetharaman, P., & Bello, J. P. (2022). How to Listen? Rethinking Visual Sound Localization. arXiv preprint arXiv:2204.05156.

Diversity of conditions

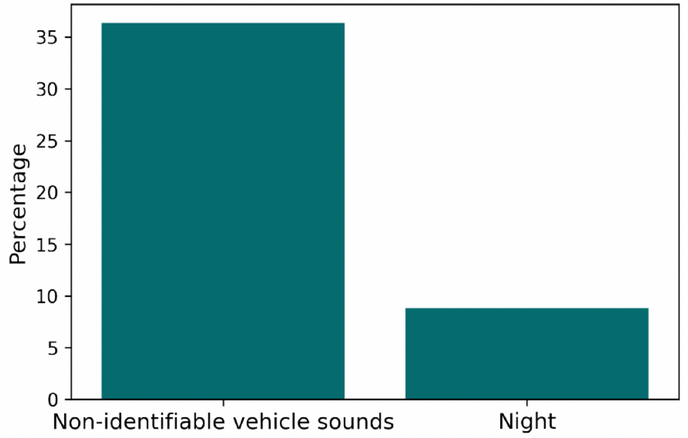
- Both active and inactive scenes
- Different amount of vehicles per frame



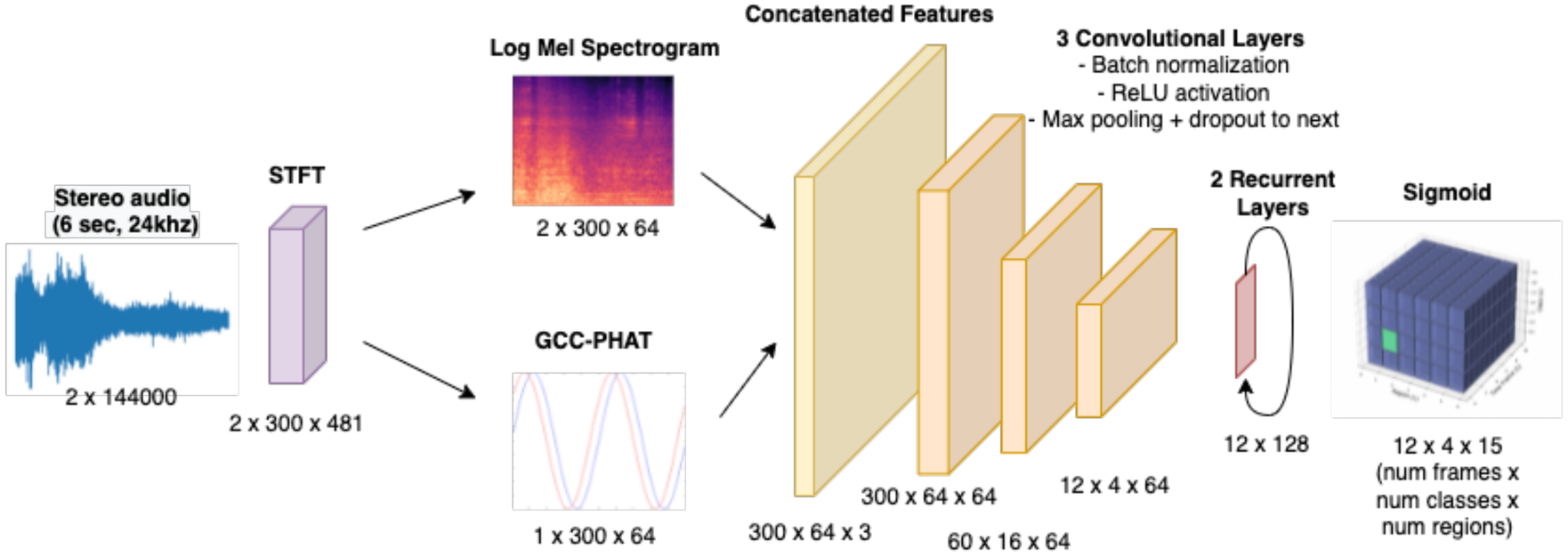
- Different class of vehicles
- Different cities, different locations

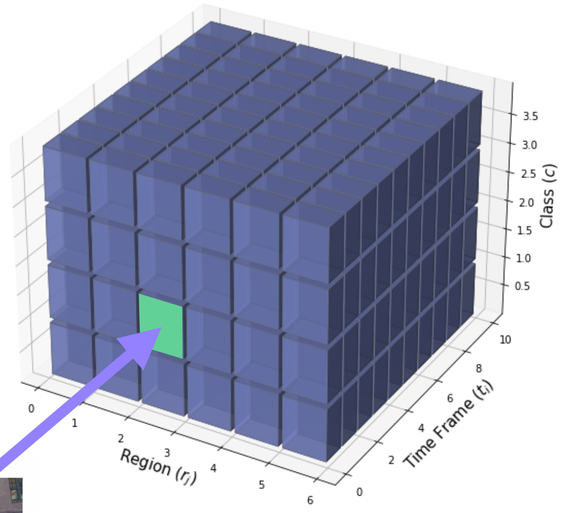


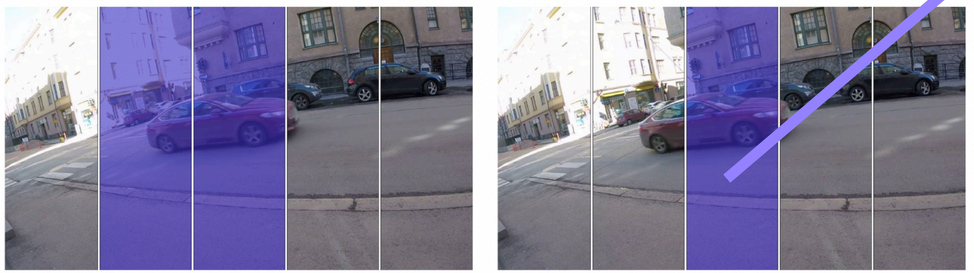
- Different scene compositions
- Different lighting conditions



Urbansas benchmark

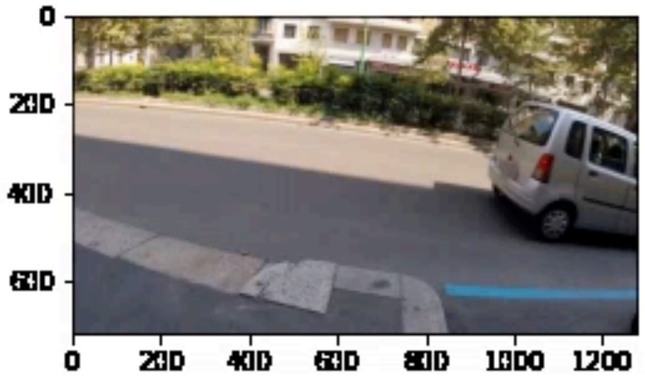




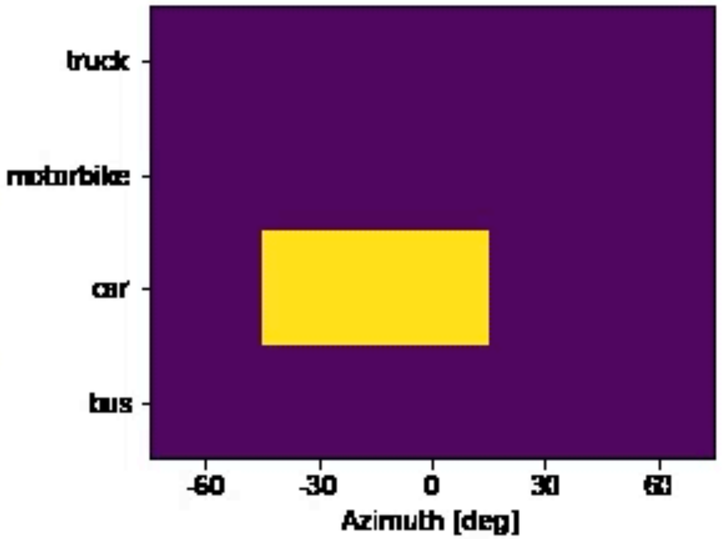


Based on Adavanne et al. 2019
Figures: Julia Wilkins

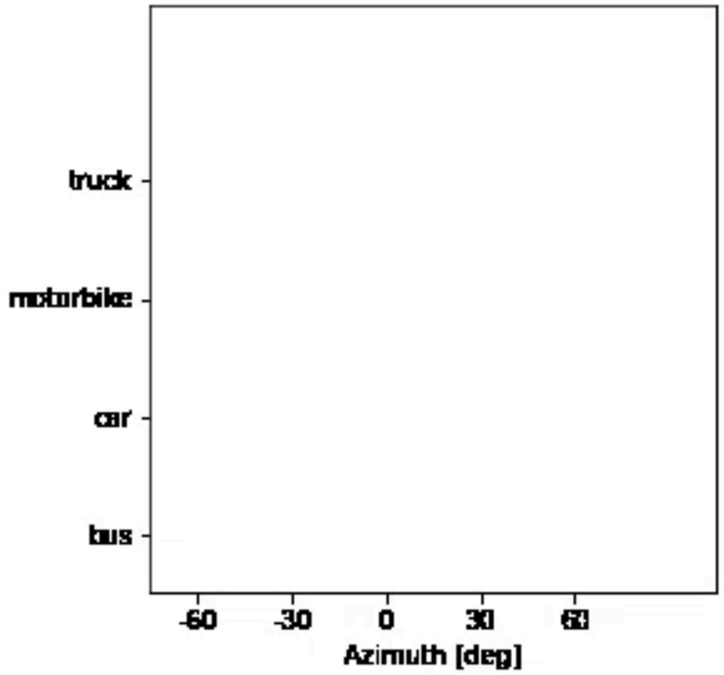
Data Output Sample



Model predictions



Ground truth



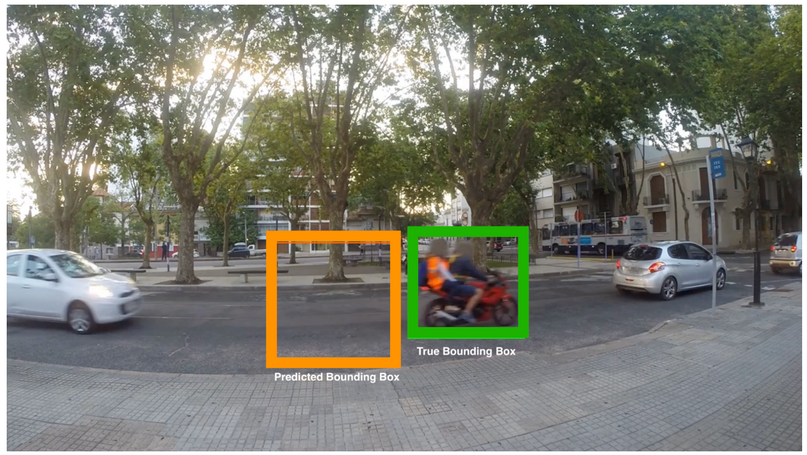
Exploring IoU for audio

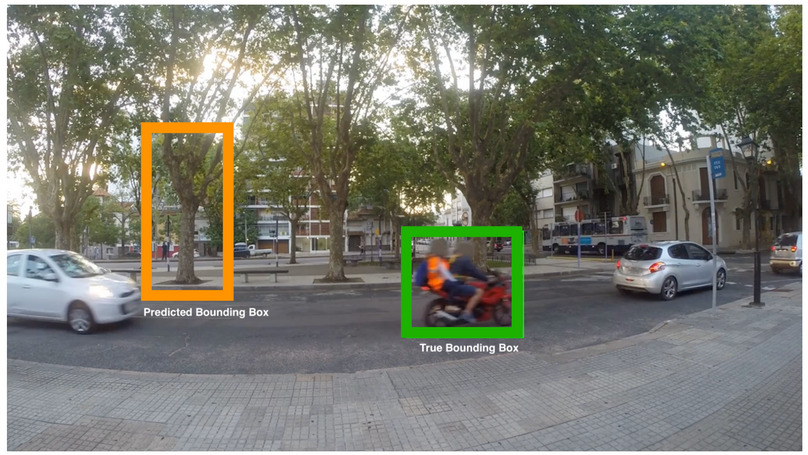
Per class IoU allows for multi-direction multi-class evaluations

$$IoU(\tau, c) = \frac{\sum_{i \in A_c(\tau)} g_{i,c}}{\sum_i g_{i,c} + \sum_{i \in A_c(\tau) - G_c} 1}$$

where i indicates the region in the image, c is the class index, τ is the threshold to determine if a prediction is positive or not so $A_c(\tau) = \{i | p_i > \tau\}_c$ and $G_c = \{i | g_i > 0\}_c$. IoU scores range

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$





Figures: Julia Wilkins

Results and challenges

model	*IoU* ($\tau = 0.05$)				
	bus	*car*	*motorbike*	*truck*	*all*
point-wise (pw)	0.332	0.344	0.231	0.143	0.260
box-wise (bw)	0.473	0.468	0.285	0.180	0.351
pw-random	0.045	0.045	0.048	0.037	0.044
bw-random	0.102	0.100	0.089	0.115	0.102

Table 2. IoU per-class of baseline models on non-empty frames.

- Urbansas is a diverse and challenging dataset
- Performance on bus and car is better - due to larger area and more instances, respectively
- Due to the number of empty frames in (between vehicle instances), the model was conditioned to under-predict
- May perform better if classification and localization tasks are separated
- IoU is unable to handle angular distance, we need to explore more to make connections to the computer vision metrics

Conclusions and future work

- Urbansas opens up the path to new research on audio and audio-visual sound source localization, vehicle tracking, self-supervised audio-visual representation for real world applications
- We present first experiments on vehicle localization and detection, including a baseline and evaluation metric exploration for the task.
- Future (and ongoing) work:
    - Two-stage or multi-task approach to disentangle detection and classification
    - Move away from binary detection (regions)
    - Investigate the different annotations of the dataset (there are plenty!)
    - Develop models for audio-visual sound source localization in urban scenes
- The data and code are open to the research community.

https://magdalenafuentes.github.io/
mf3734@nyu.edu

https://github.com/beasteers
bsteers@nyu.edu