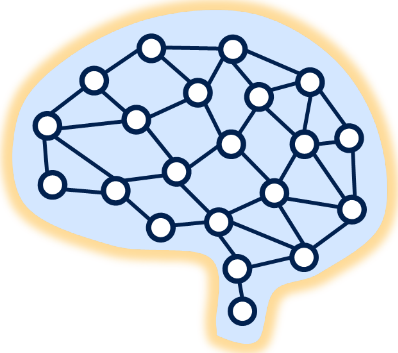


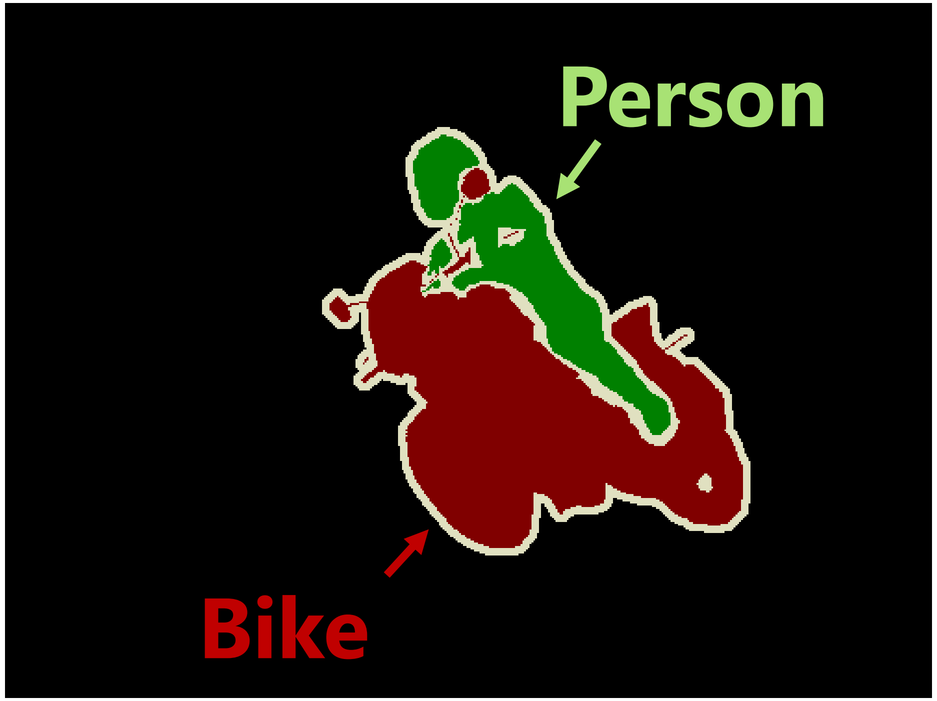
Seeing through Sounds

Visual Scene Understanding from Acoustic Signals

Go Irie
Tokyo University of Science
goirie@ieee.org

Visual Scene Understanding

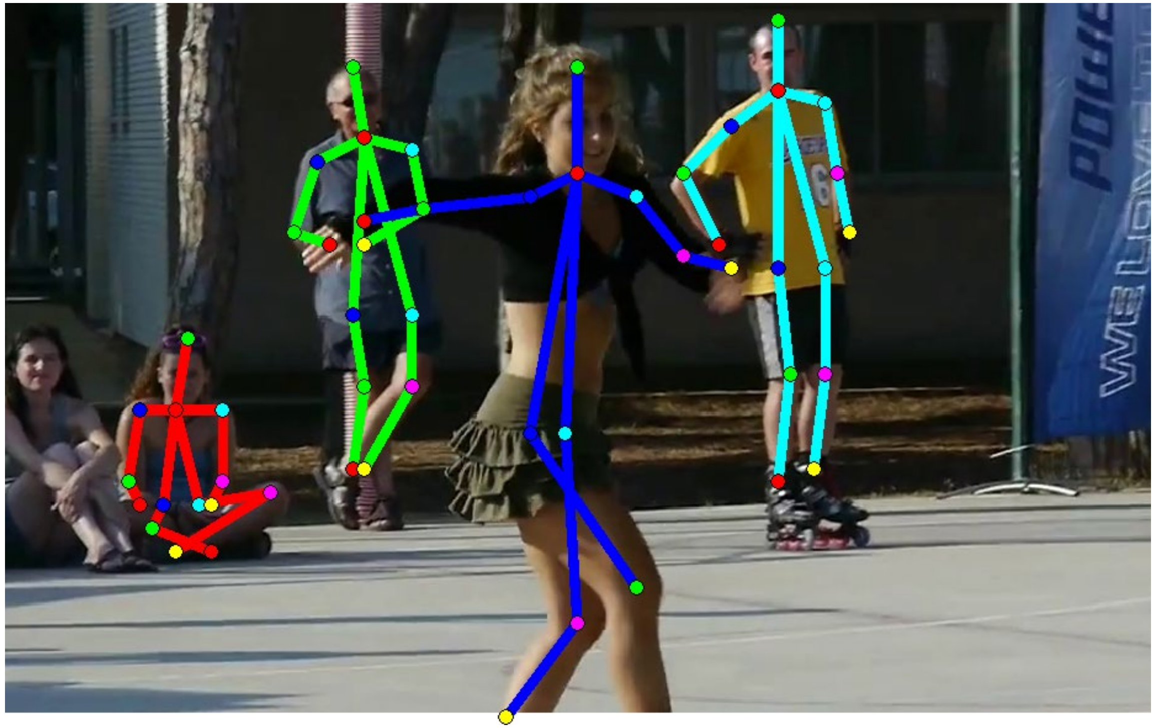
- **Recognize visible information in scene, e.g., object name, human action, etc.**
- **Object Recognition**



Images from Pascal VoC dataset

Visual Scene Understanding

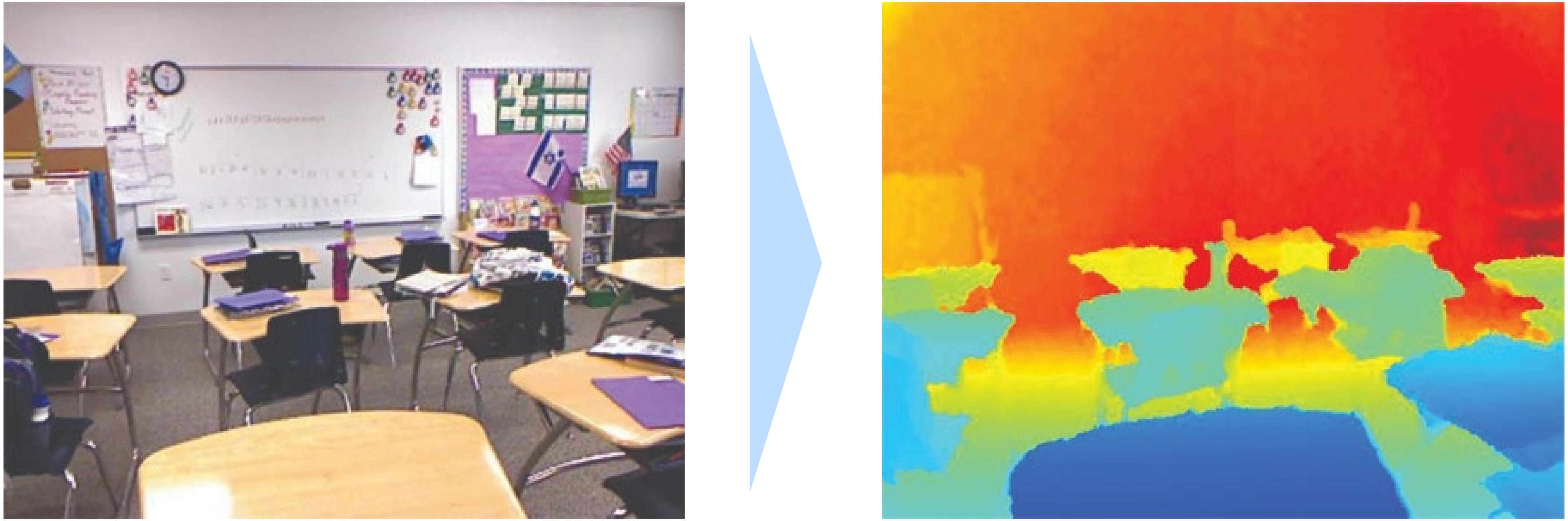
- **Recognize visible information in scene, e.g., object name, human action, etc.**
- **Object Recognition, Human Pose Estimation,**



Images from https://pose.mpi-inf.mpg.de/

Visual Scene Understanding

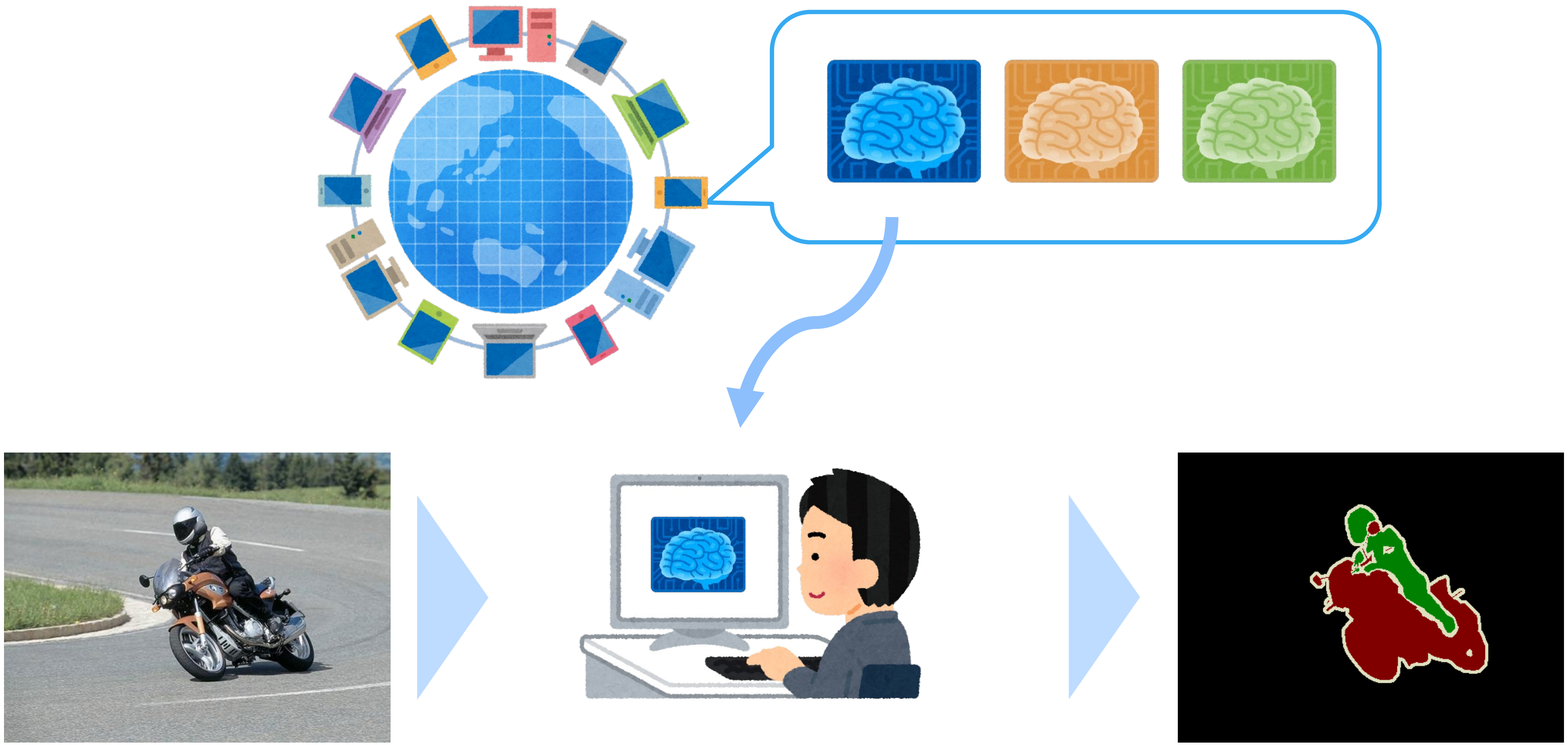
- **Recognize visible information in scene, e.g., object name, human action, etc.**
- **Object Recognition, Human Pose Estimation, Scene Depth Estimation, etc.**



Images from NYU Depth V2 dataset

Visual Scene Understanding

- SOTA models publicly available online
- Just use them to get excellent recognition results

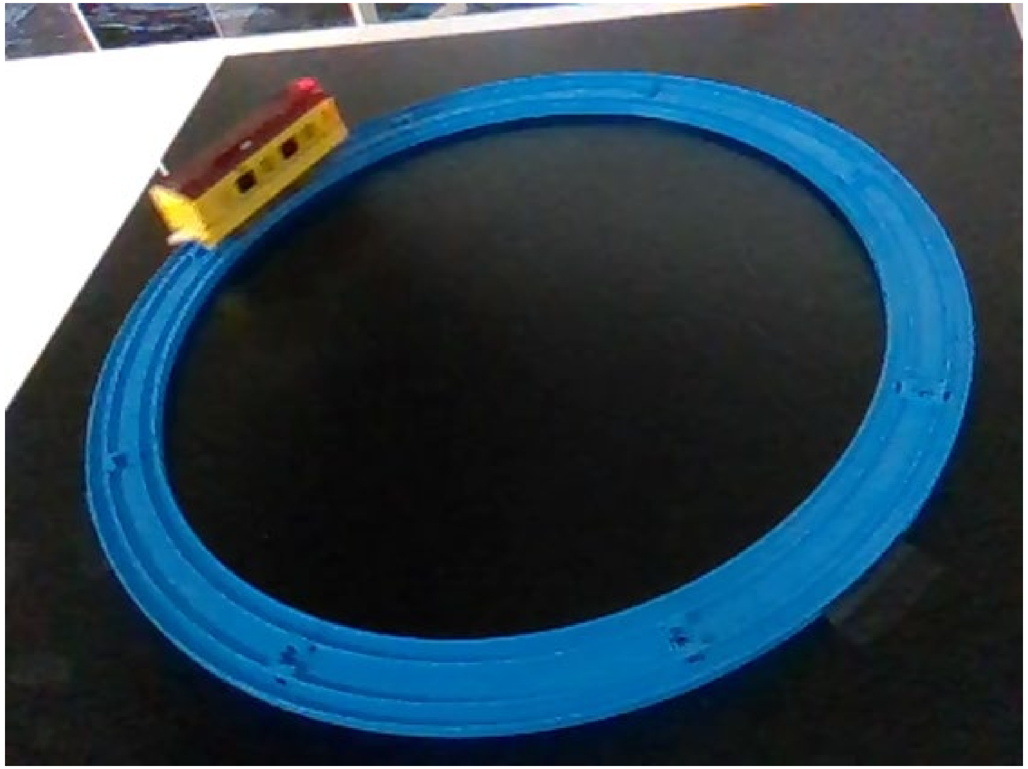


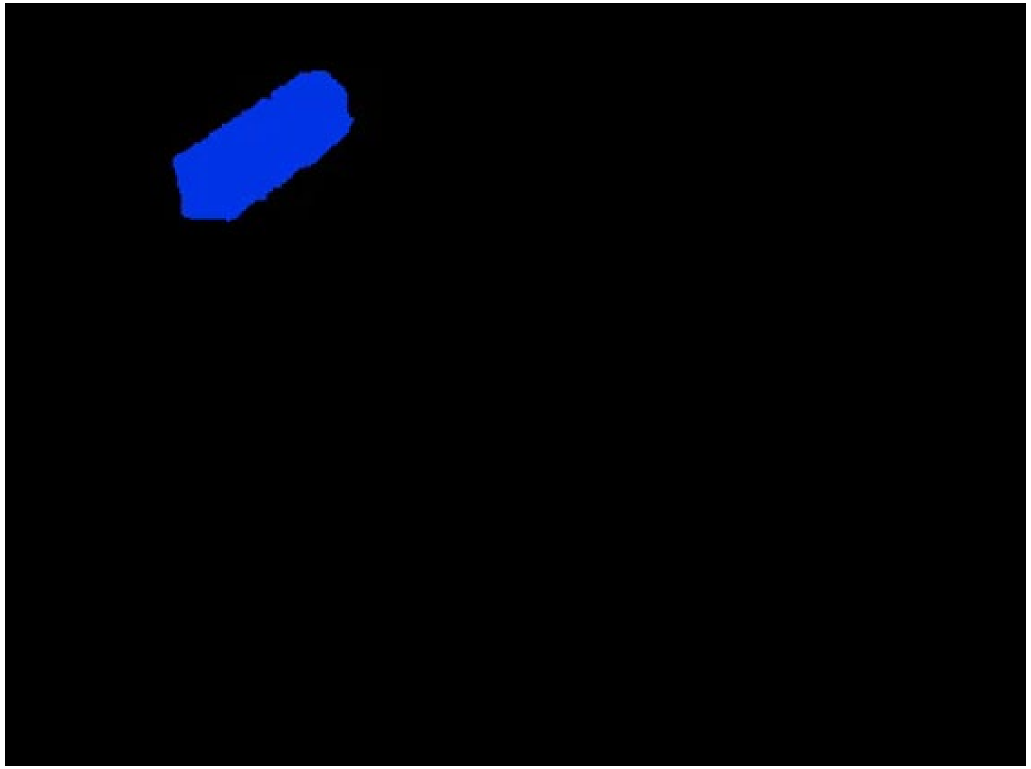
What if we cannot use camera?

- Dark room
- Highly private/public area
- Place where camera prohibited by rule
- Etc.

Our idea
Use Microphone Instead!

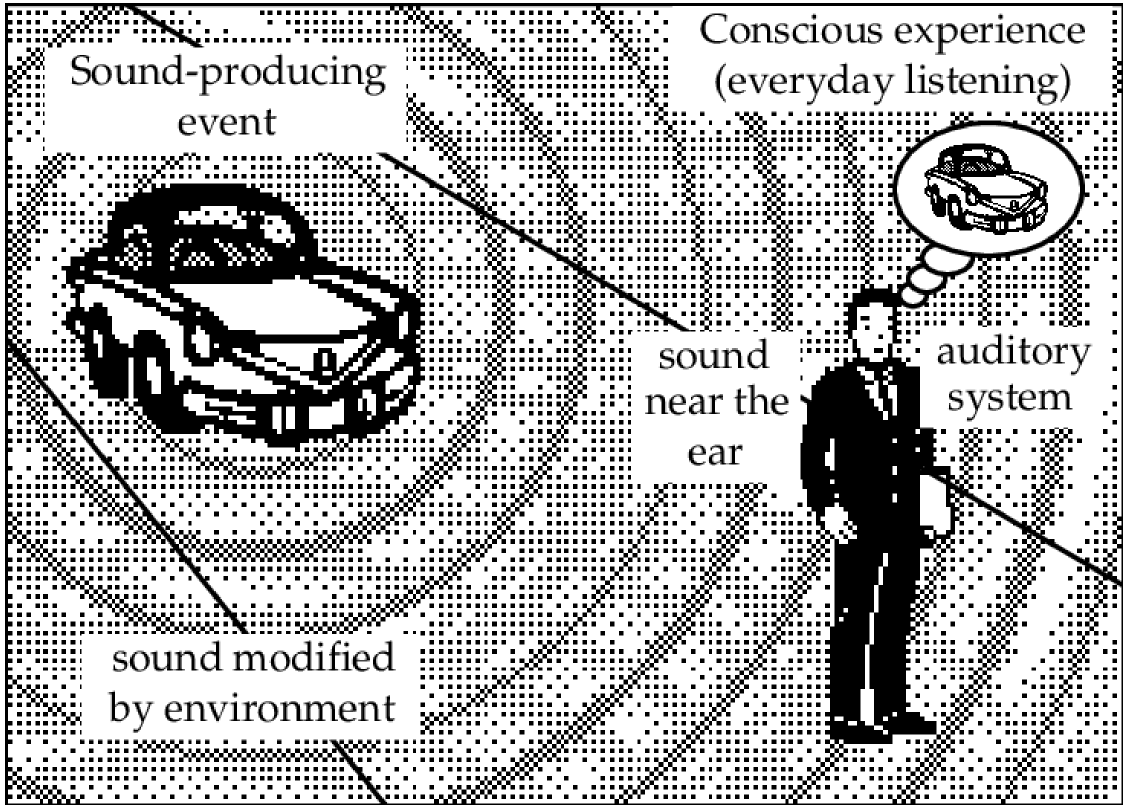
Predicting Segmentation Results from Sound





Why Audio? -- Auditory Scene Analysis

- **Human can "recognize" object around her/him from just hearing sound, without looking at source of sound**
- **Build Auditory Scene Analysis artificially?**

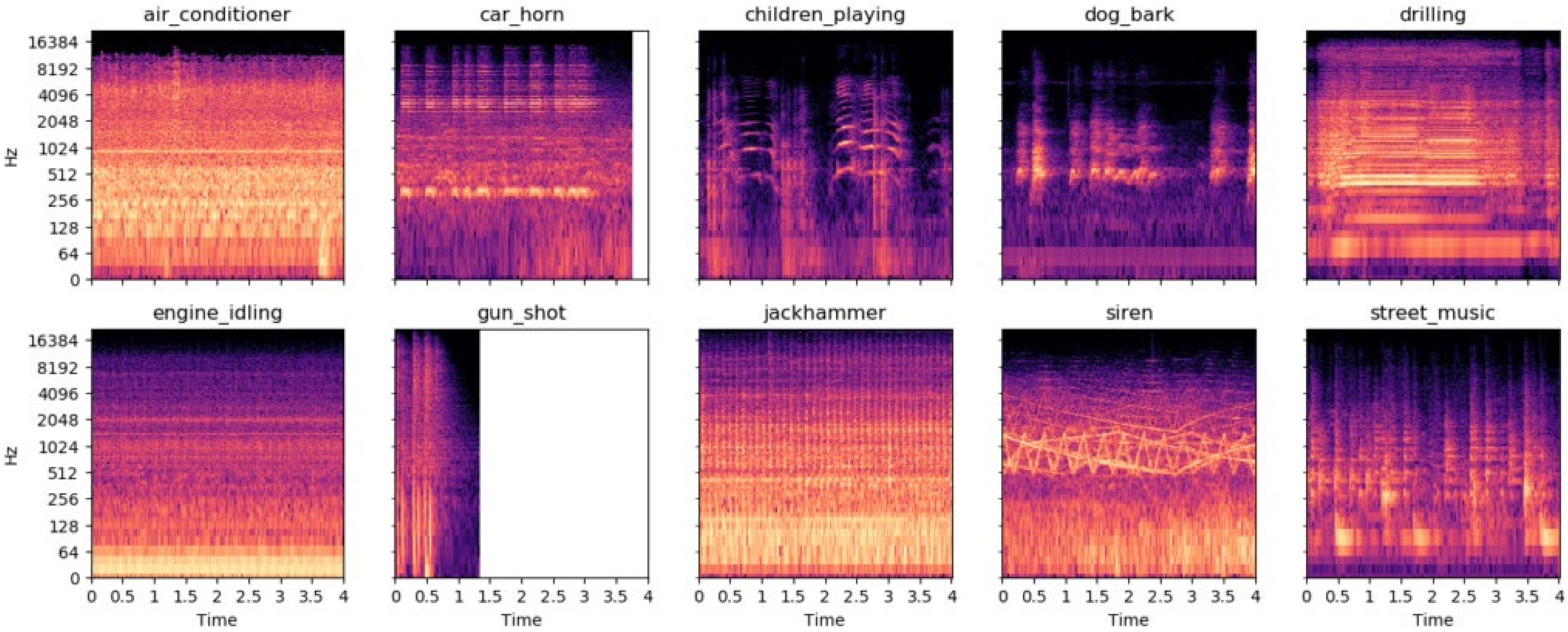


[Gaver, 1993]

Shutterstock

Why Audio? -- Semantic & Geometric Cues

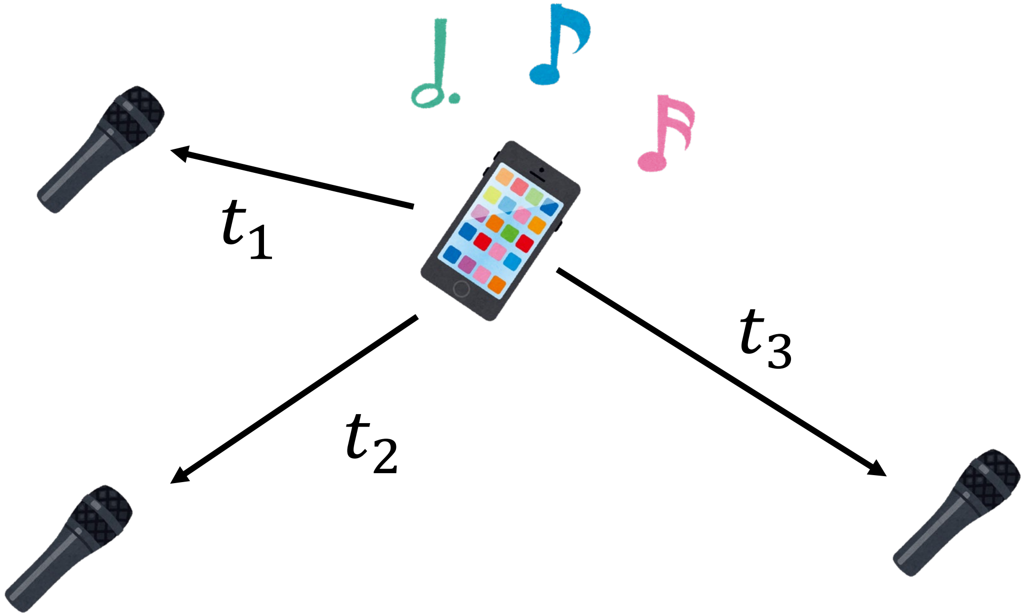
- **Different sound shows different spectral property**



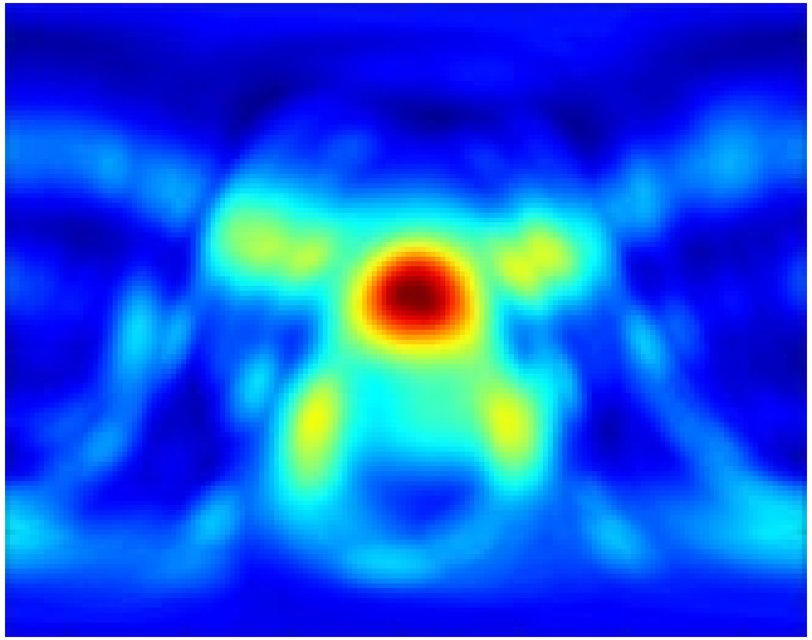
https://www.jonnor.com/2021/12/audio-classification-with-machine-learning-europython-2019/

Why Audio? -- Semantic & Geometric Cues

- **Different sound shows different spectral property**
- **Time difference of arrival gives direction of sound arrival**

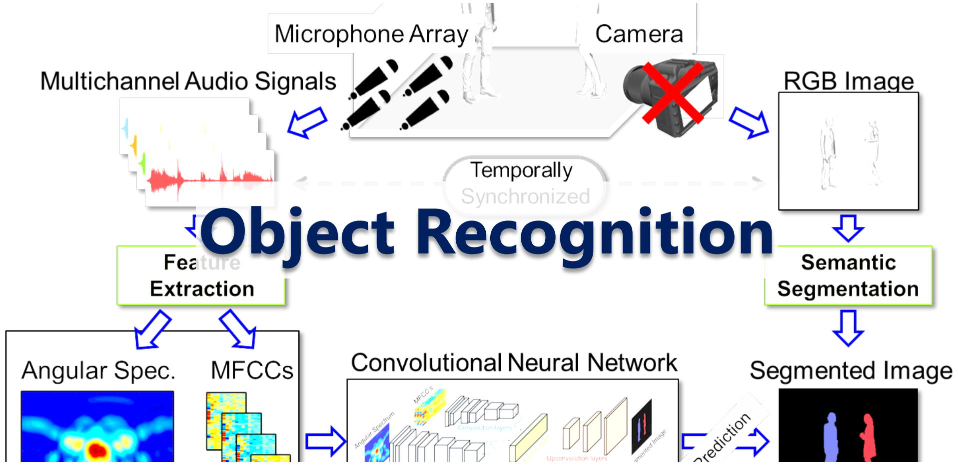


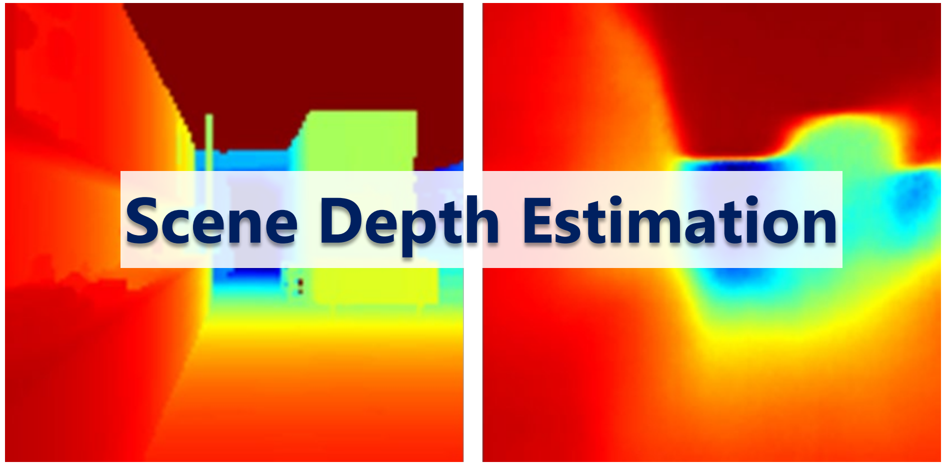
Estimated Direction of Sound

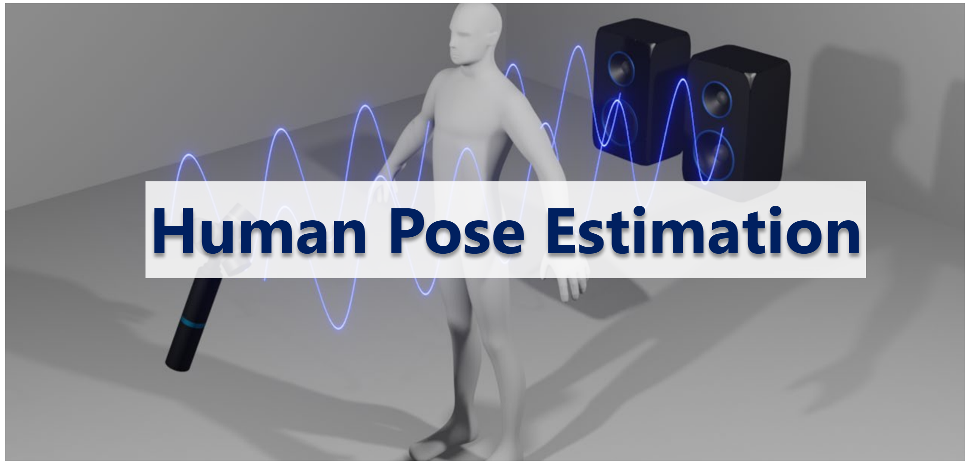


Our Projects

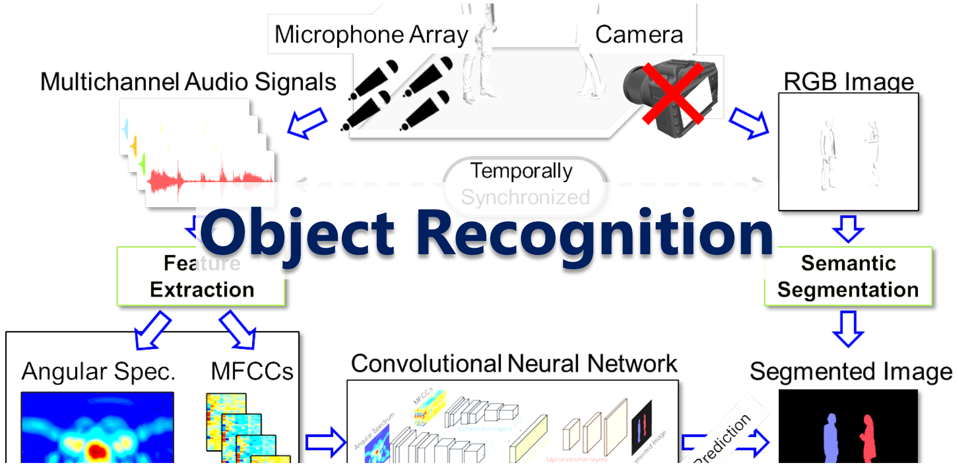
	Passive Sensing	Active Sensing
Static	Object Recognition	Scene Depth Estimation
Dynamic	Human Action Recognition	Human Pose Estimation



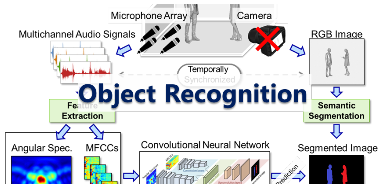


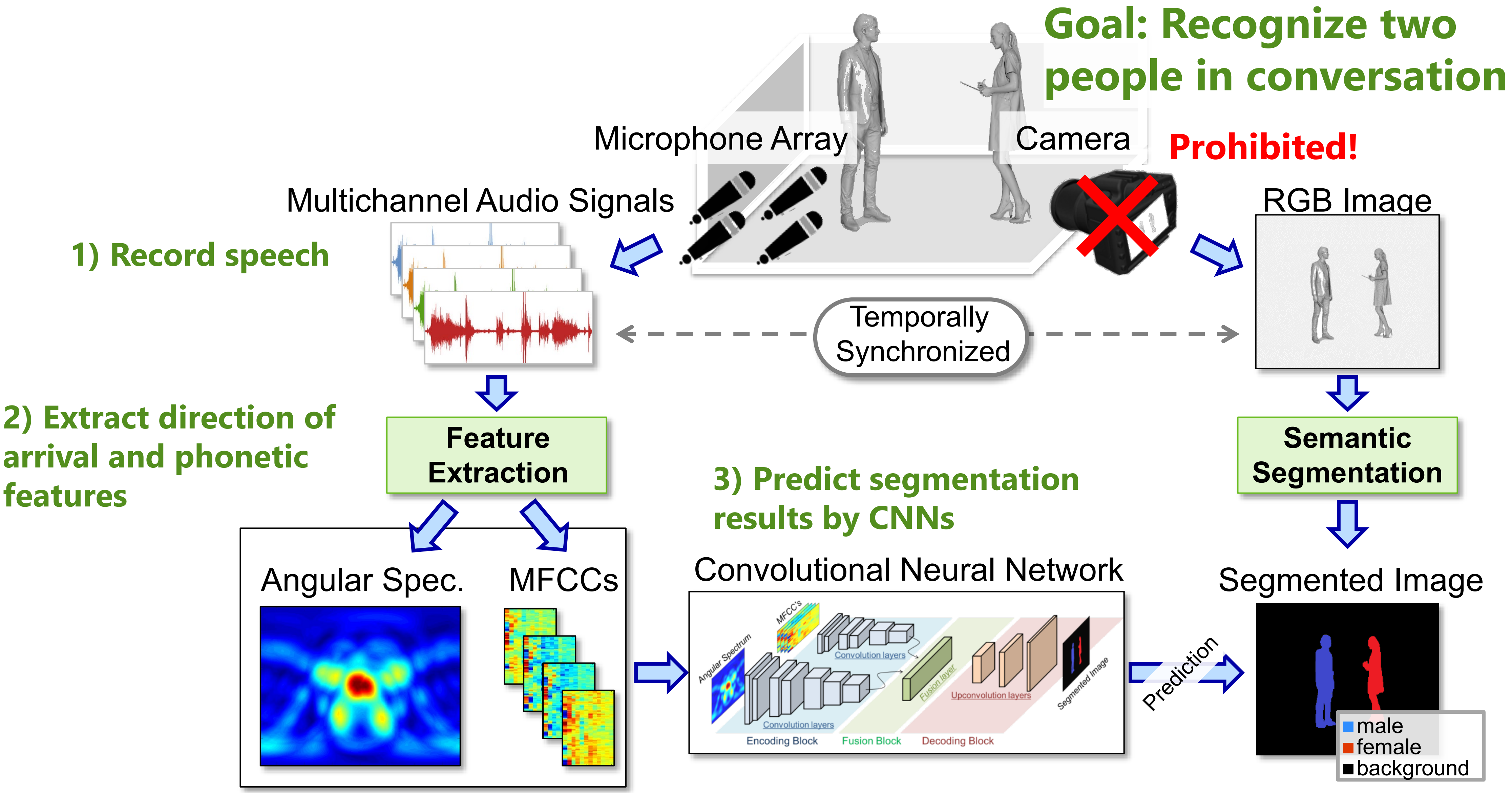


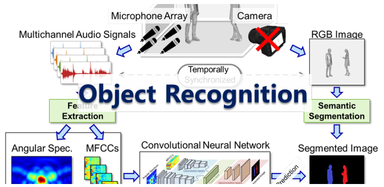
Our Projects

	Passive Sensing	Active Sensing
Static		
Dynamic		

Predicting Semantic Segmentation Results







Prediction Results

Image Recognition Results (Ground Truth) | **Prediction from Sound**

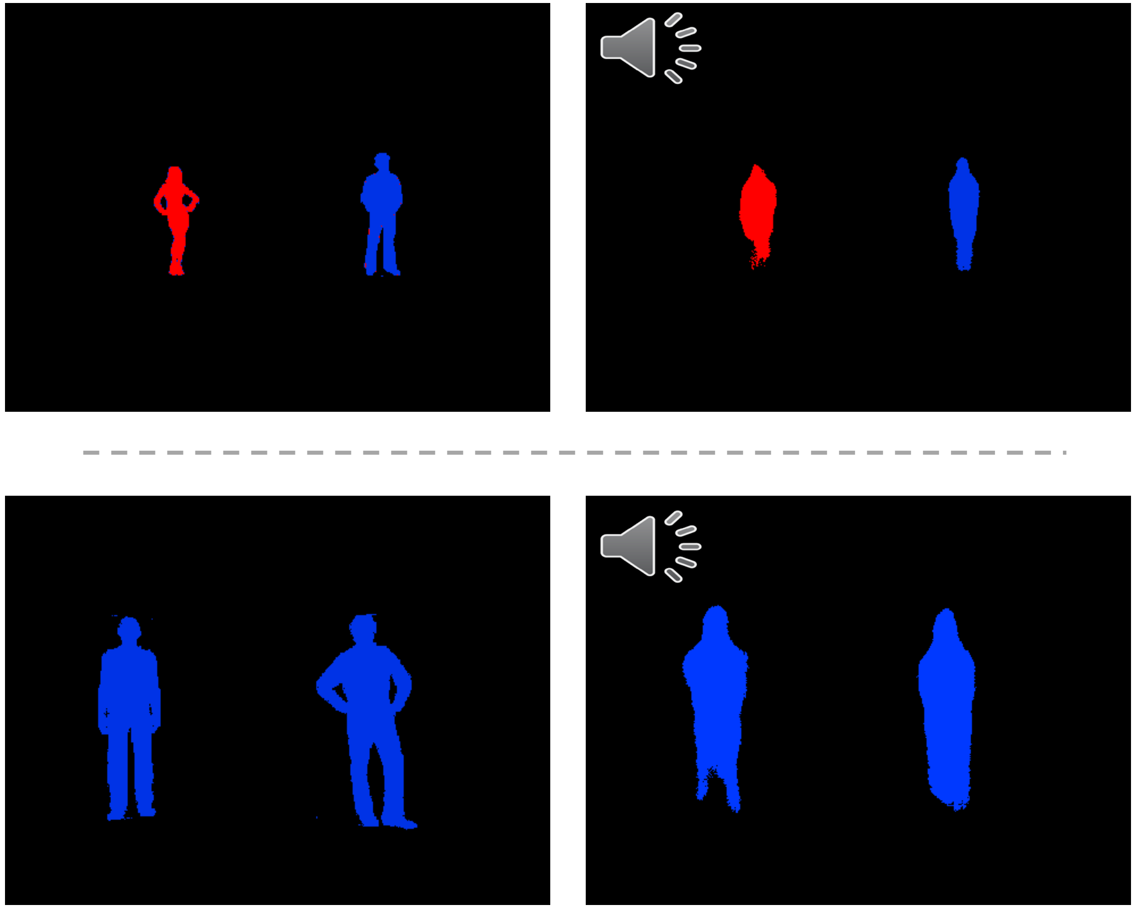
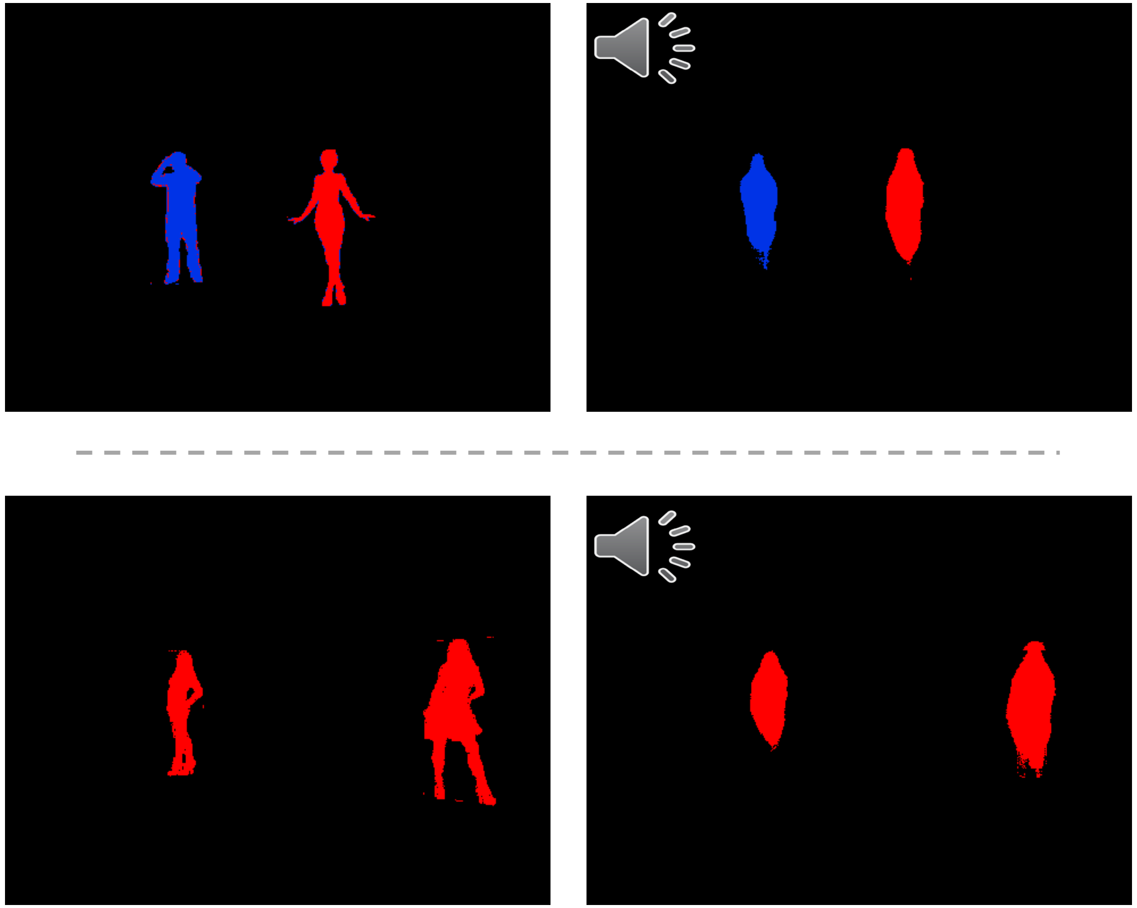
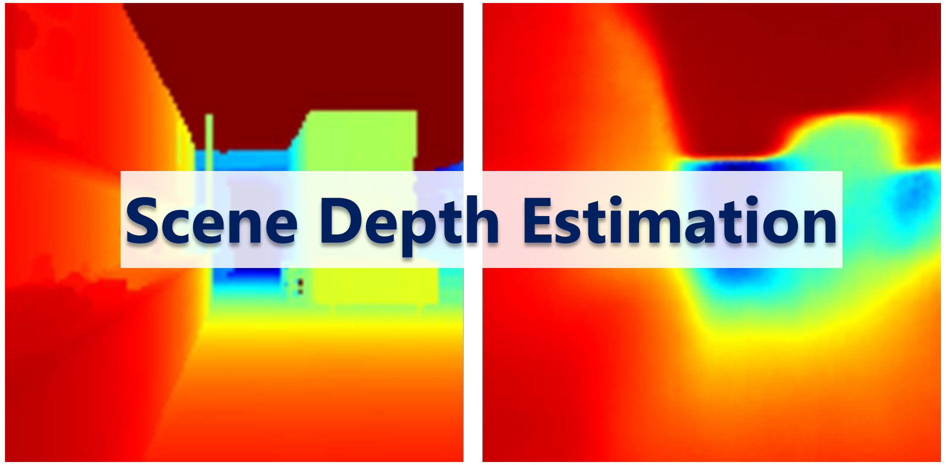
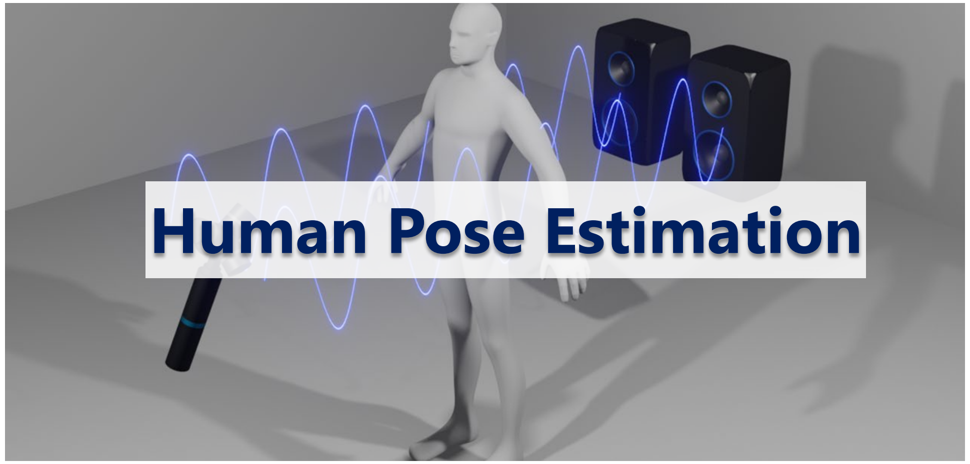


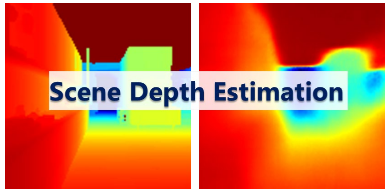
Image Recognition Results (Ground Truth) | **Prediction from Sound**

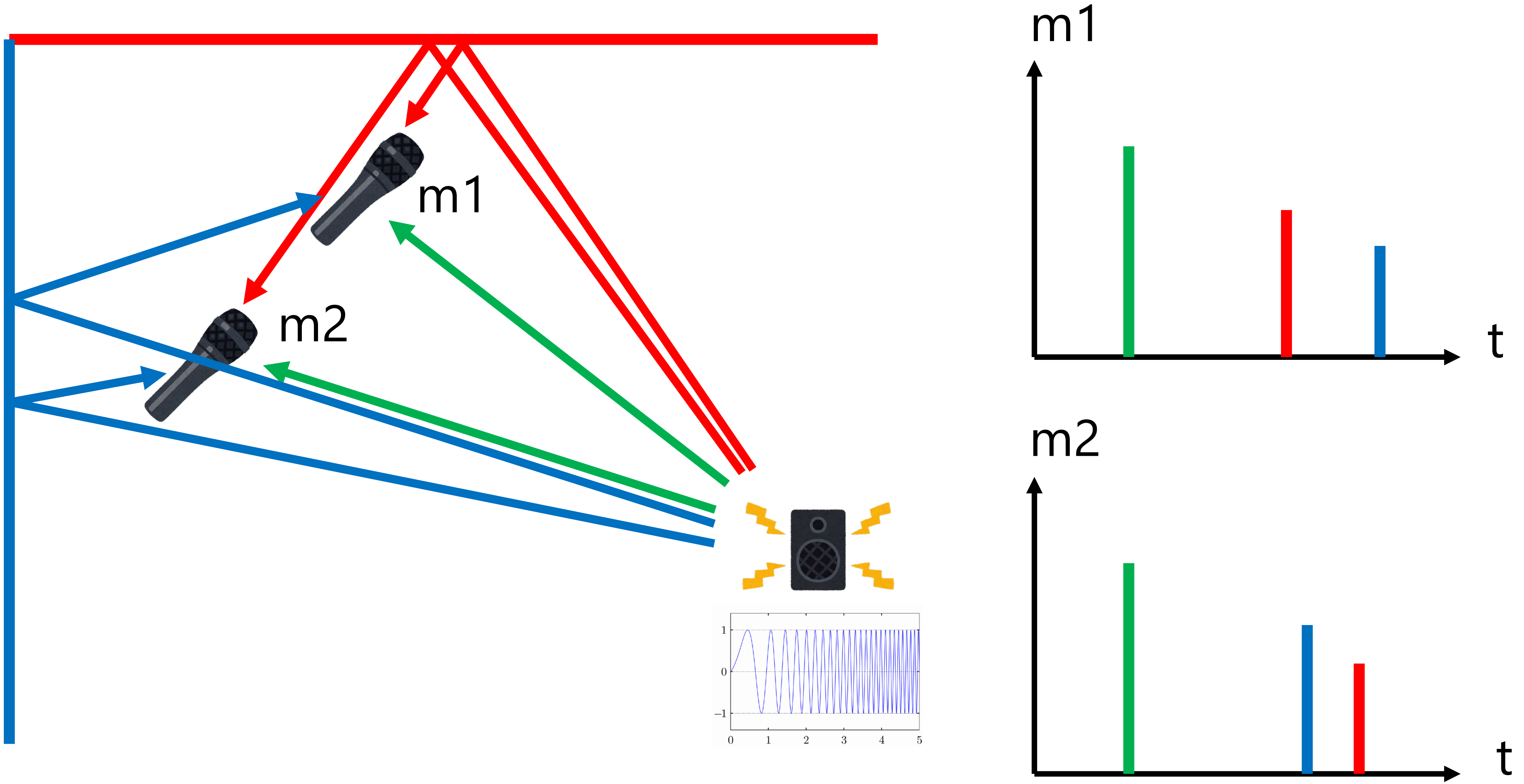


Our Projects

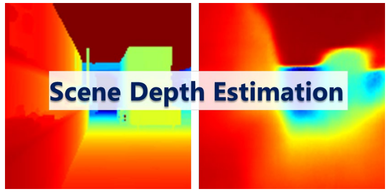
	Passive Sensing	Active Sensing
Static		
Dynamic		

Active Acoustic Sensing

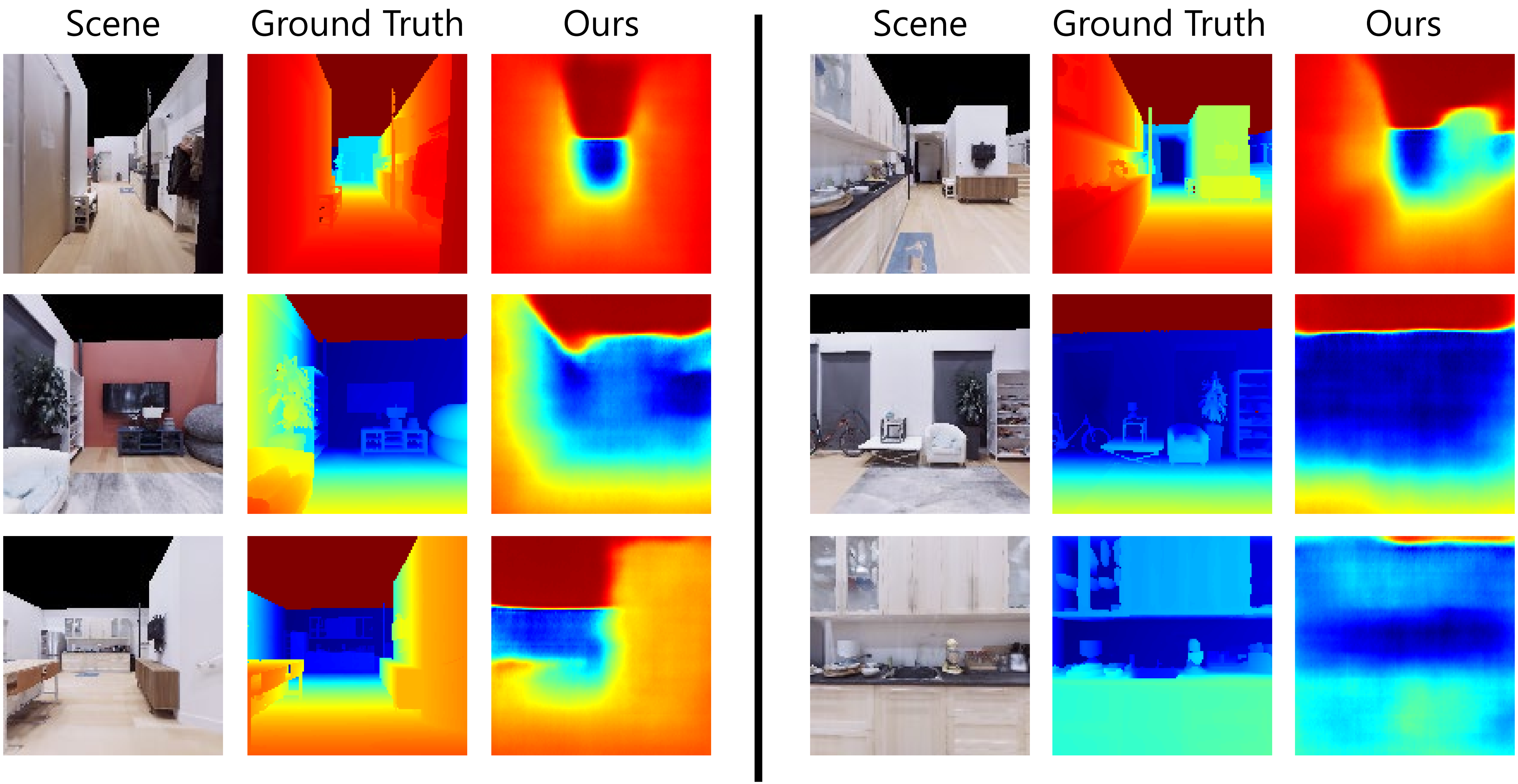




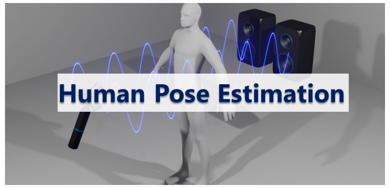
Dokmanić et al., 2013

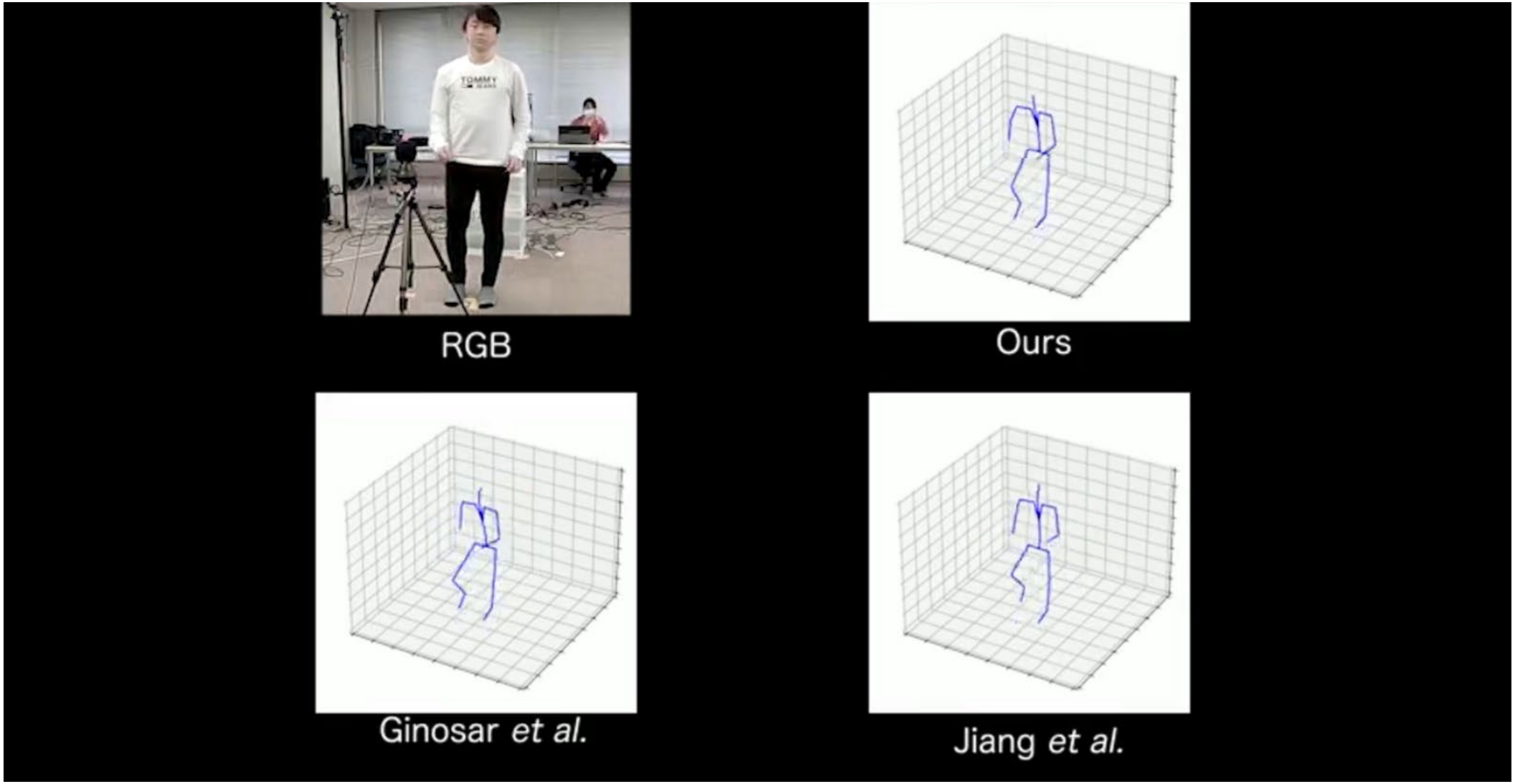


Echo-based Depth Estimation

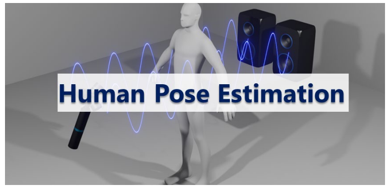


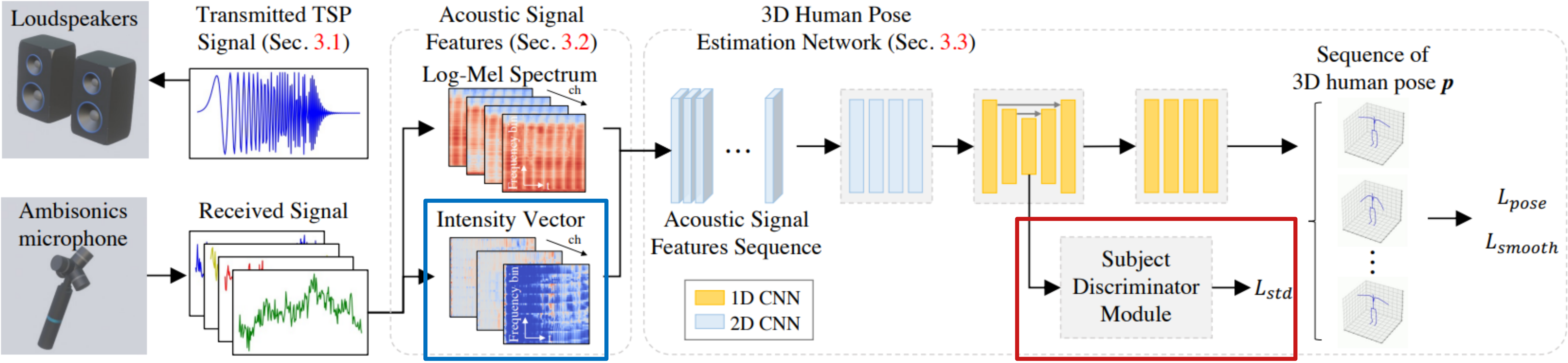
Listening Human Pose





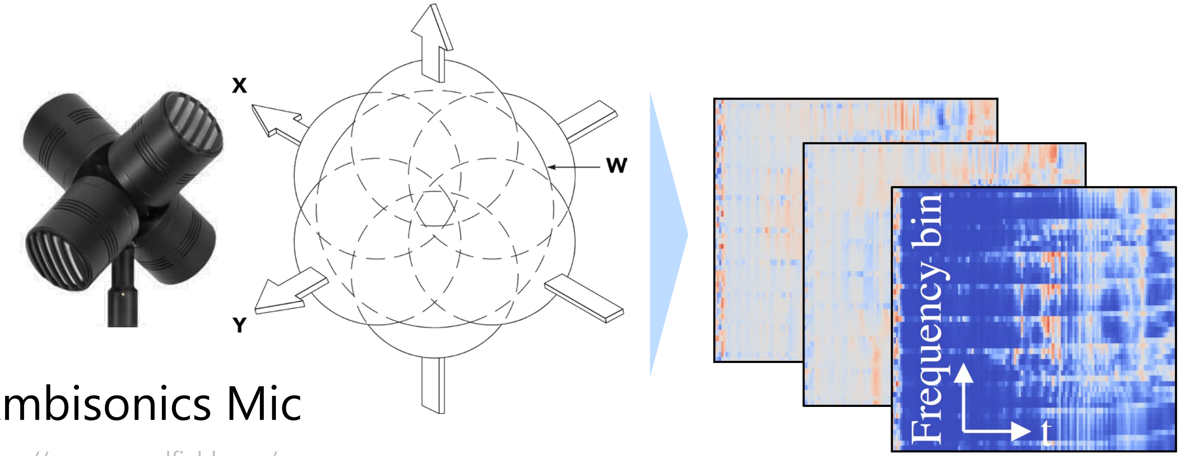
Framework





Intensity Vector

Capturing Direction of Arrival



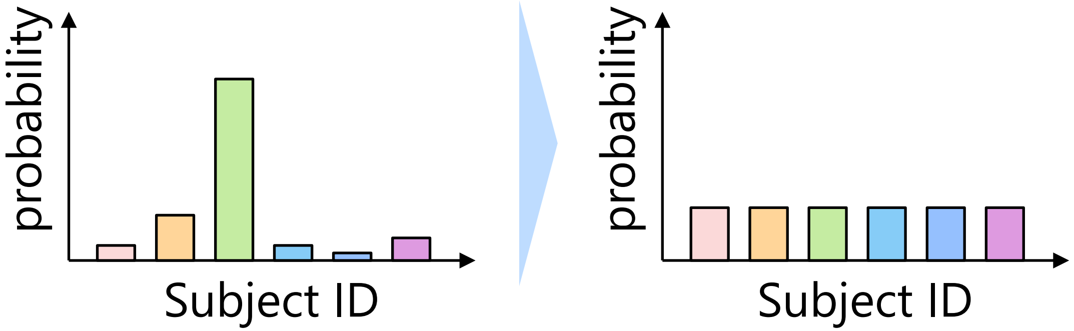
Ambisonics Mic

https://www.soundfield.com/

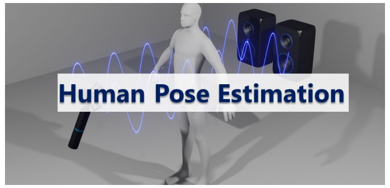
Subject-Agnostic Learning

Adversarial learning to subject classification

Diversifying subject prediction probability
to make subject unpredictable



Results



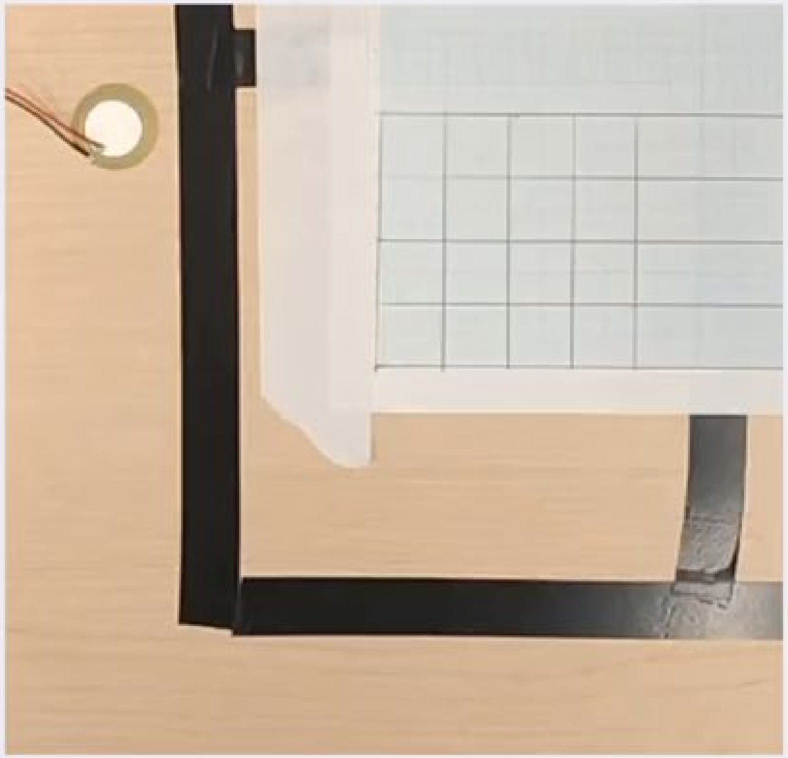
Method	Anechoic Chamber Environment					
	Single Subject			Cross Subject		
	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)	RMSE (↓)	MAE (↓)	PCKh @0.5 (↑)
Ginosar *et al*. [10]	0.44	0.23	0.90	0.83	0.51	0.60
Jiang *et al*. [18]	0.90	0.44	0.73	0.96	0.55	0.62
Ours (Method's best)	**0.42**	**0.22**	**0.90**	**0.73**	**0.45**	**0.72**

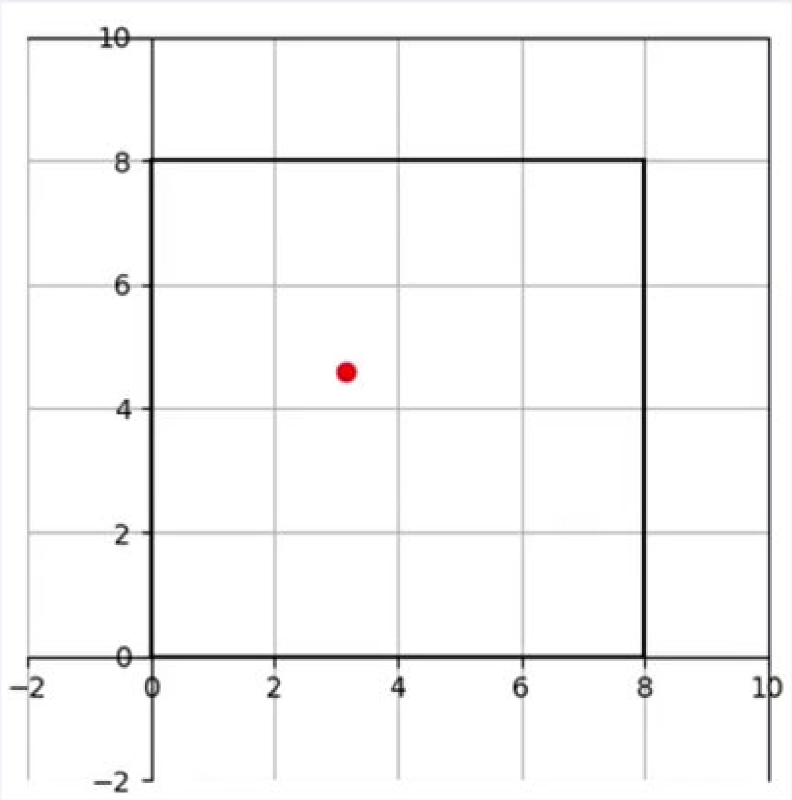
Project page:

Our Projects

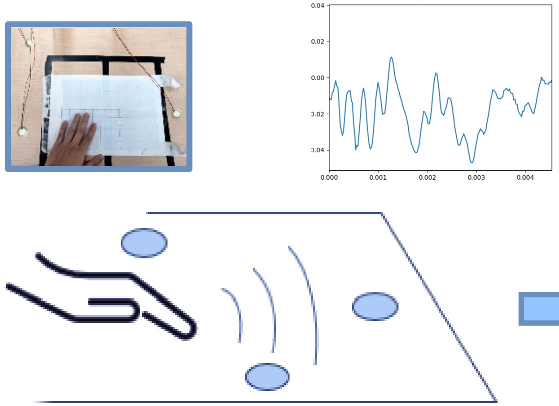
	Passive Sensing	Active Sensing
Static	Object Recognition	Scene Depth Estimation
Dynamic	**Human Action Recognition**	Human Pose Estimation

Contact Point Estimation

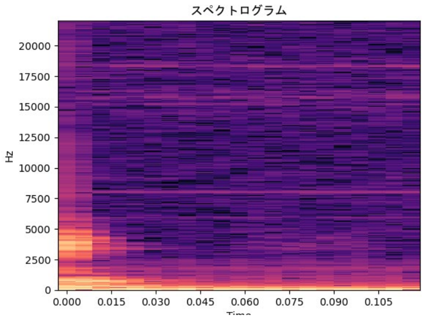




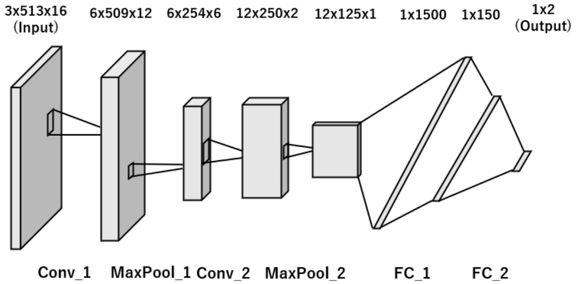
Overview



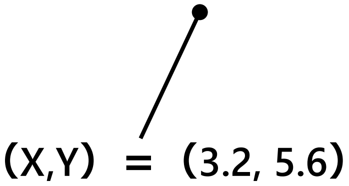
Sound Acquisition



Spectrogram

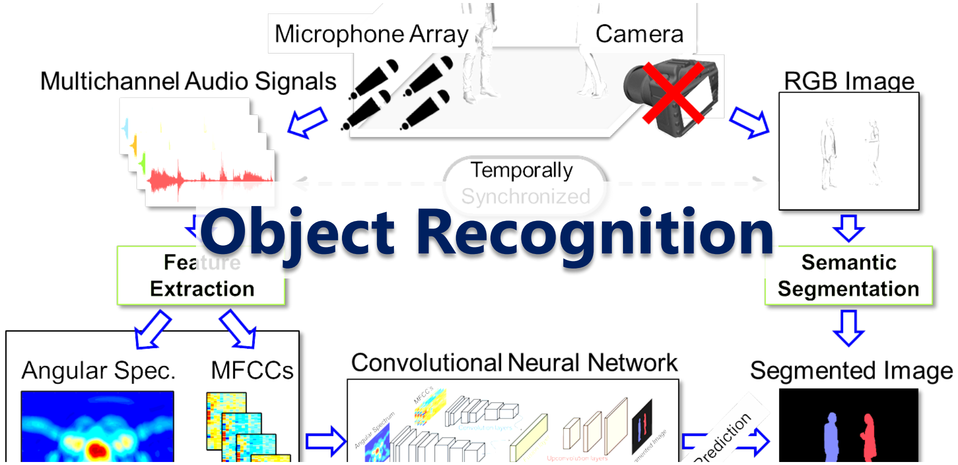
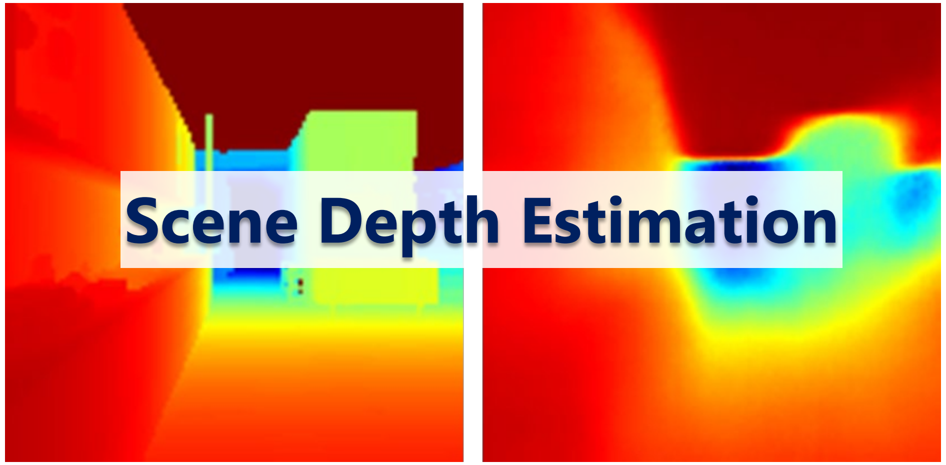
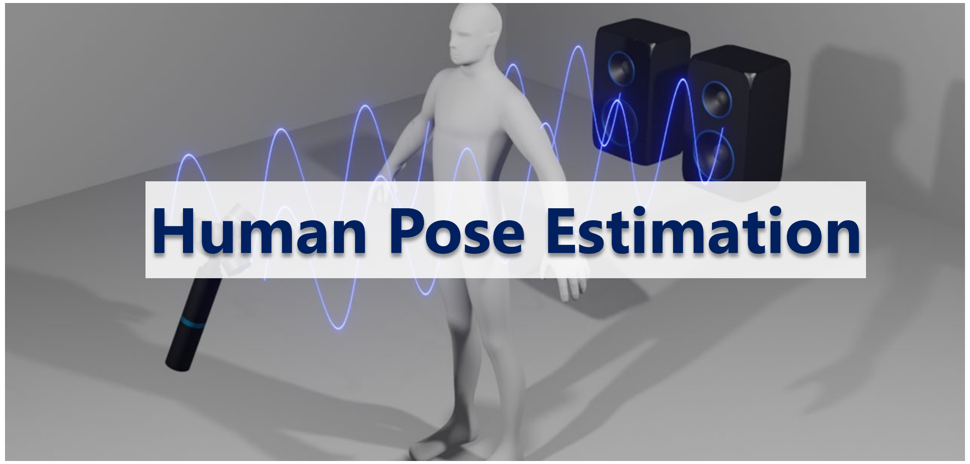


Regression Network



Estimated Position

Our Projects

	Passive Sensing	Active Sensing
Static	Object Recognition 	Scene Depth Estimation 
Dynamic	Human Action Recognition 	Human Pose Estimation 

Summing Up

- **Sound is a promising alternative to image for solving various visual scene understanding tasks, e.g., object recognition, scene depth estimation, human pose estimation, and action recognition.**

- **Light and sound are often "equivariant":**
  Audio-visual model may have a potential to learn various physical/geometric knowledge in the real world.

Seeing through Sounds

Visual Scene Understanding from Acoustic Signals

Go Irie
Tokyo University of Science
goirie@ieee.org