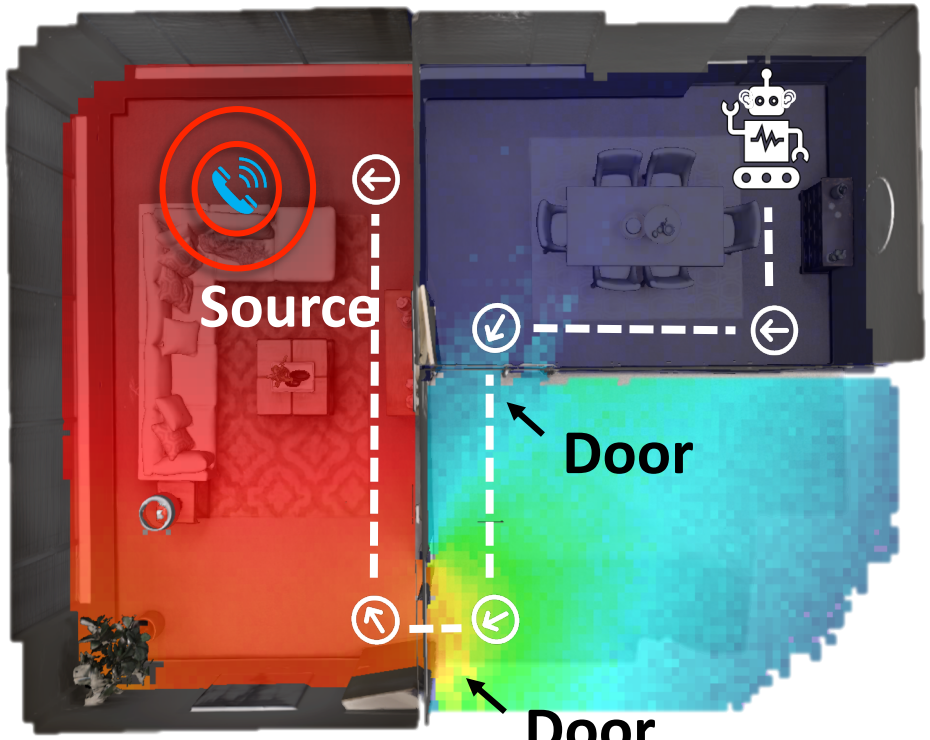
SoundSpaces: Audio-Visual Navigation in 3D Environments

Changan Chen[*1,4], *Unnat Jain*[*2,4], *Carl Schissler*[3], *Sebastia V. Amengual Gari*[3], *Ziad Al-Halah*[1], *Vamsi K. Ithapu*[3], *Philip Robinson*[3], *Kristen Grauman*[1,4]

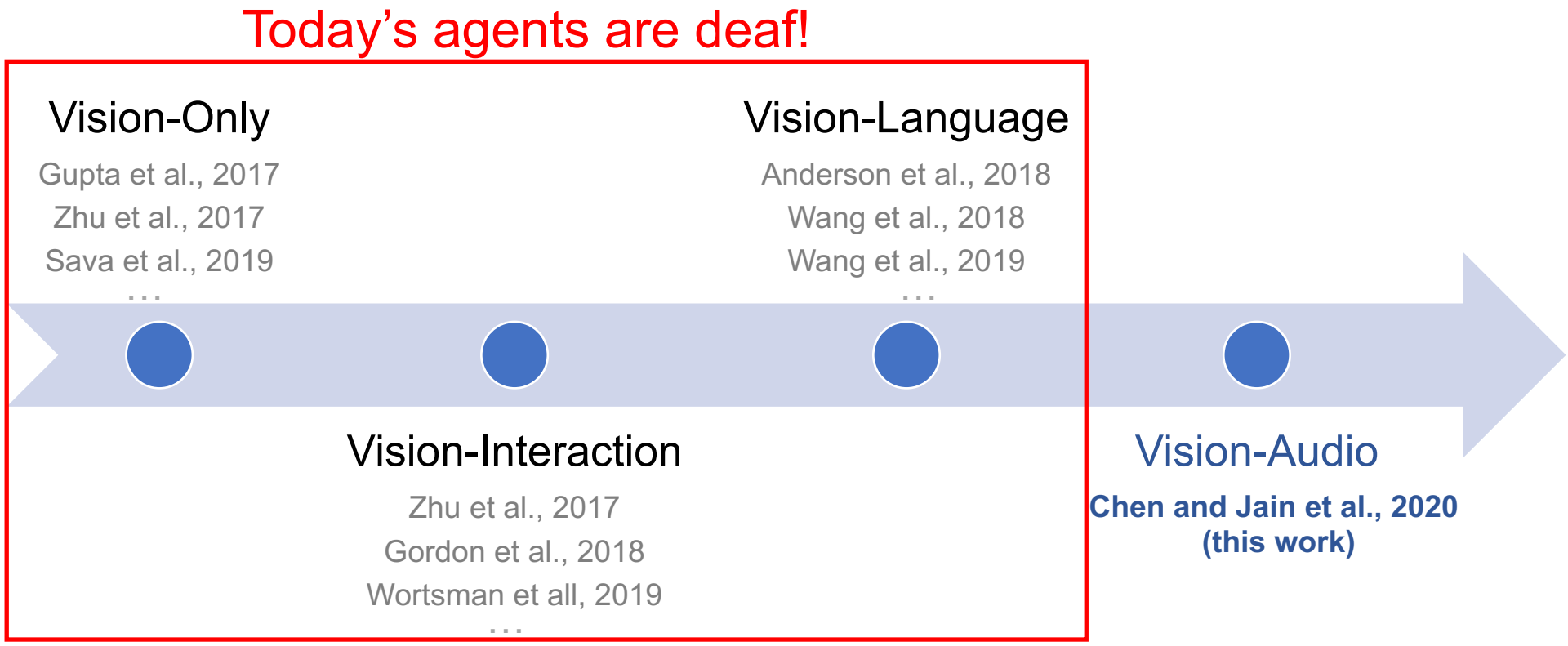
[1]UT Austin, [2]UIUC, [3]Facebook Reality Labs, [4]Facebook AI Research



ECCV'20
SEC, Glasgow
23-28 AUGUST 2020

Embodied Perception Is a Multisensory Experience

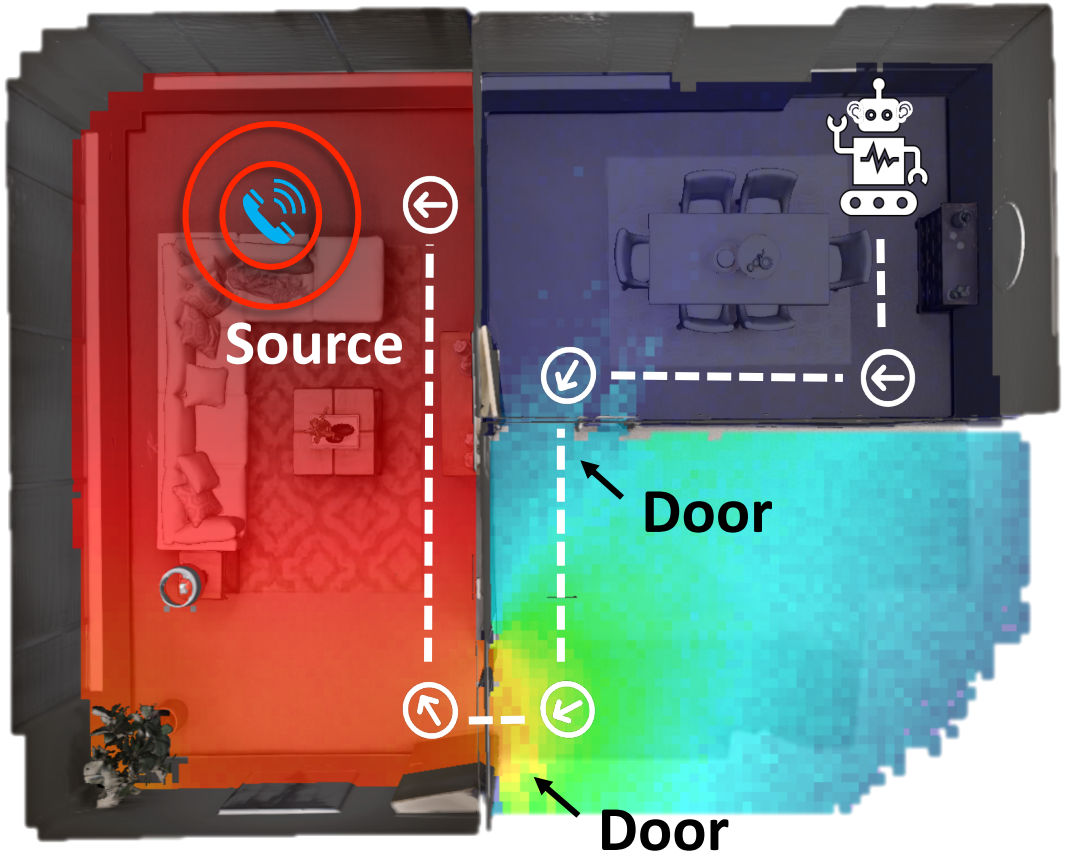
We often use *vision*, *audio*, *touch*, *smell* to move around



Our contribution: audio-visual embodied navigation --- task and simulation

Audio-Visual Navigation in 3D Environments

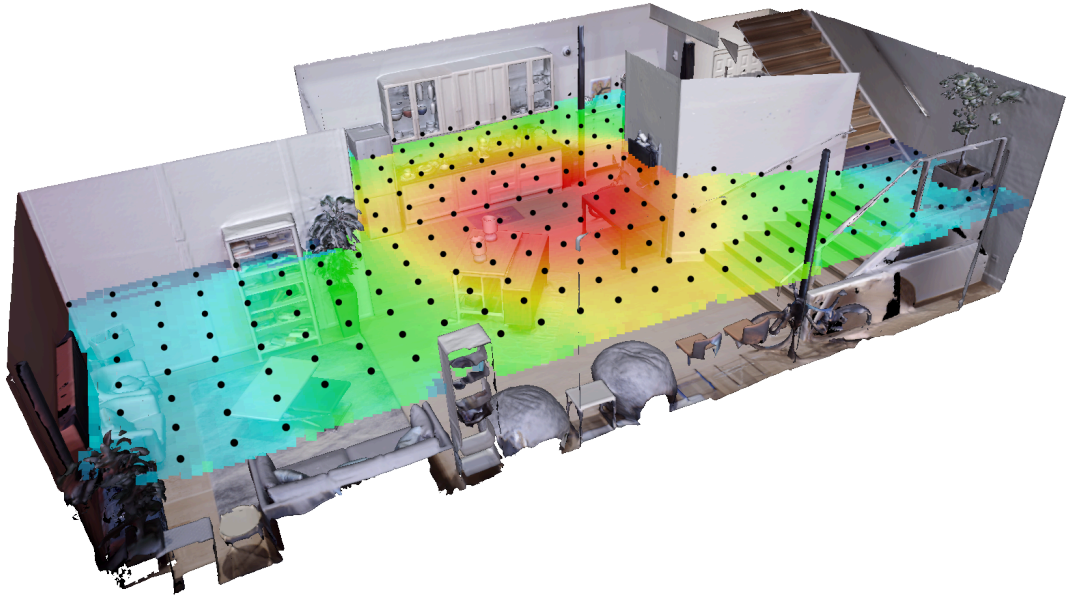
An agent navigates to a sounding object with vision and audio perception



C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

SoundSpaces: Our Audio Simulator

- We introduce SoundSpaces, an audio simulation platform to enable audio-visual navigation for two visually realistic 3D environments: Replica[1] and Matterport3D[2]



	# Scenes	Avg. Area	# Training Eps.
Replica	18	47.24 m^2	0.1M
Matterport3D	85	517.34 m^2	2M

Table: Summary of dataset statistics

[1]The Replica Dataset: A Digital Replica of Indoor Spaces, Straub et al., arXiv, 2019
[2]Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV, 2017

SoundSpaces: Our Audio Simulator

- We introduce SoundSpaces, an audio simulation platform to enable audio-visual navigation for two visually realistic 3D environments: Replica[1] and Matterport3D[2]

- Our audio simulator produces realistic audio rendering based on the room geometry, materials, and sound source location

- The platform can play varying sounds of your choice in real time by precomputing a transfer function between locations

[1]The Replica Dataset: A Digital Replica of Indoor Spaces, Straub et al., arXiv, 2019
[2]Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV, 2017

Example 1: Where Is My Phone?

Agent view

Top-down map (unknown to the agent)

Direction: left ear is louder when the agent faces upward on the top-down map

Intensity: overall intensity gets higher as the agent gets closer to the goal

Agent | Goal | Start | Shortest path | Agent path | Seen/Unseen area | Occupied area

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Example 2: Where Is The Piano?

Agent view

Top-down map (unknown to the agent)

Agent | Goal | Start | Shortest path | Agent path | Seen/Unseen area | Occupied area

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Audio-Visual Navigation Tasks

New tasks

PointGoal

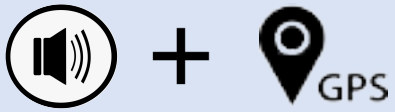
Gupta et al., 2017
Savva et al., 2019
…

The agent receives a displacement vector (Δx, Δy) pointing towards the goal at each time step

AudioGoal

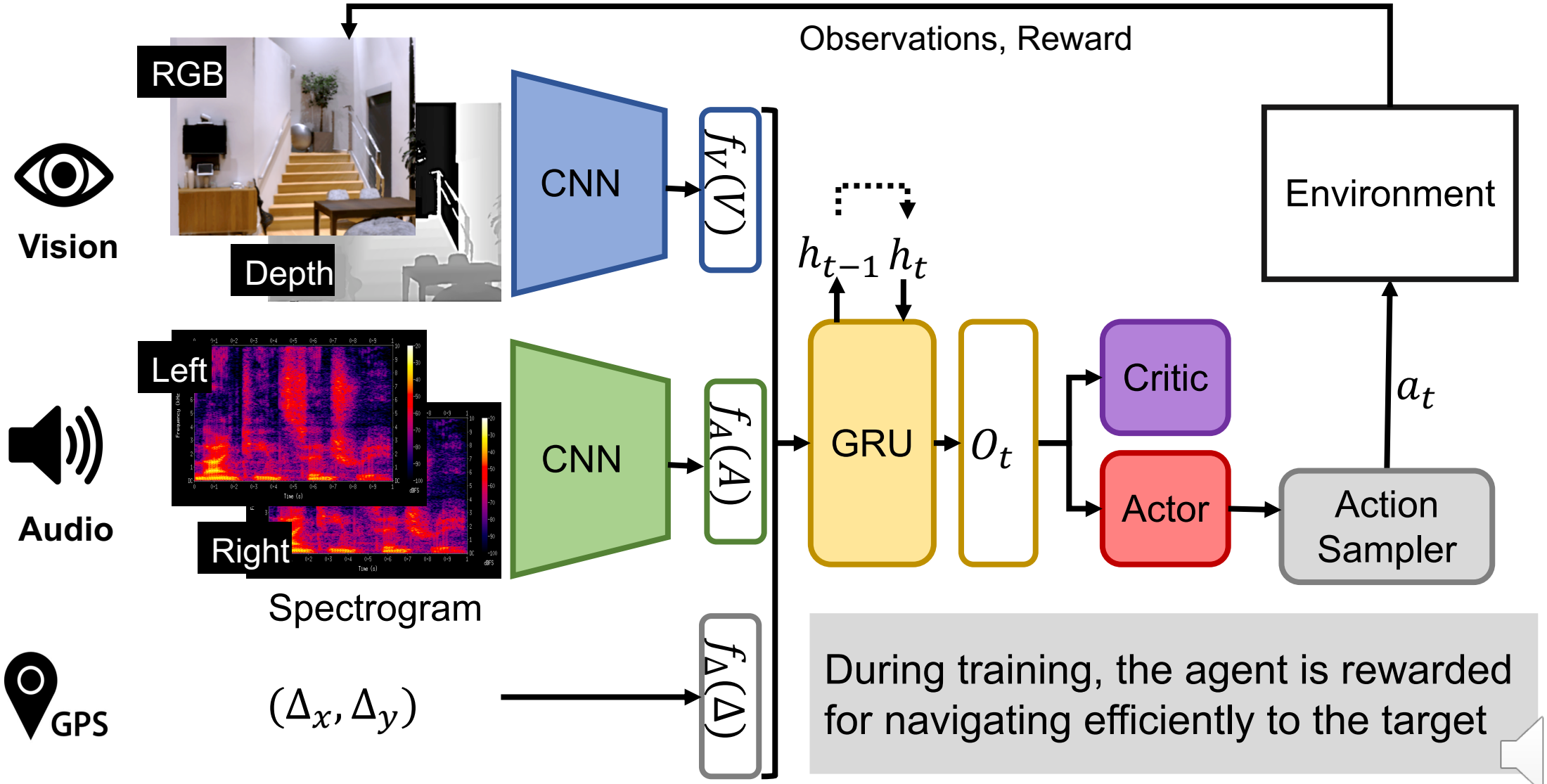
The agent receives an audio signal emitted by the sounding object at each time step

AudioPointGoal



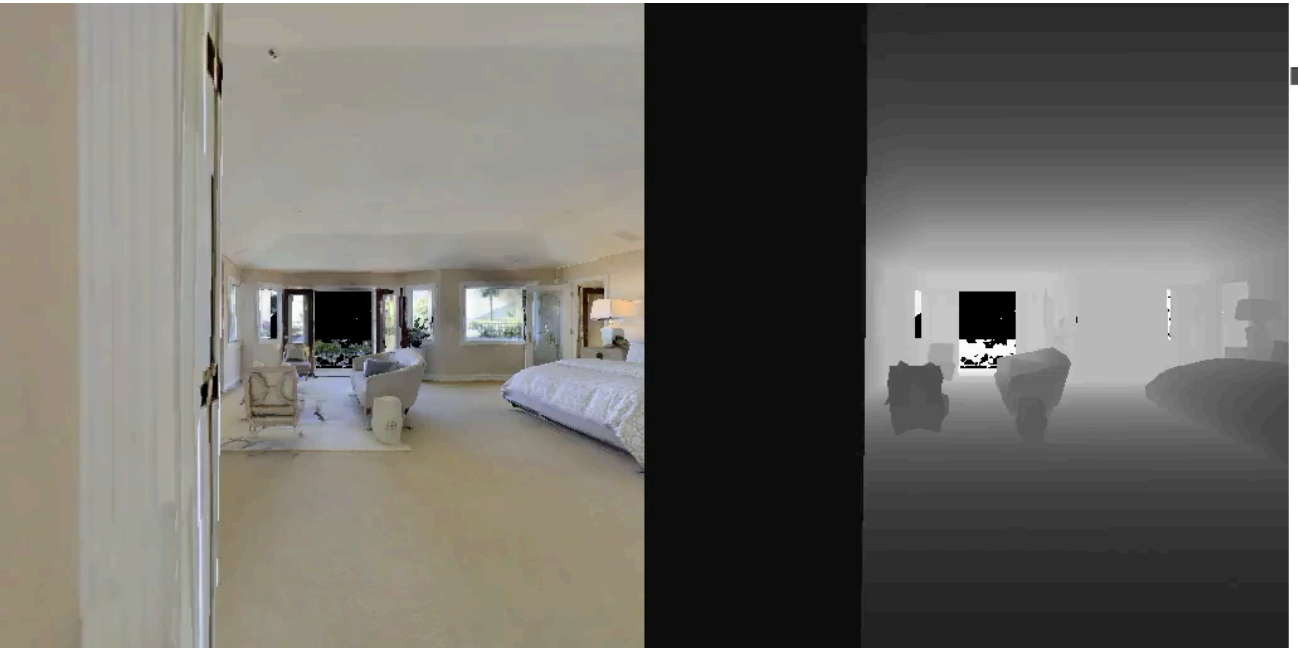
The agent receives both a displacement vector (Δx, Δy) and an audio signal at each time step

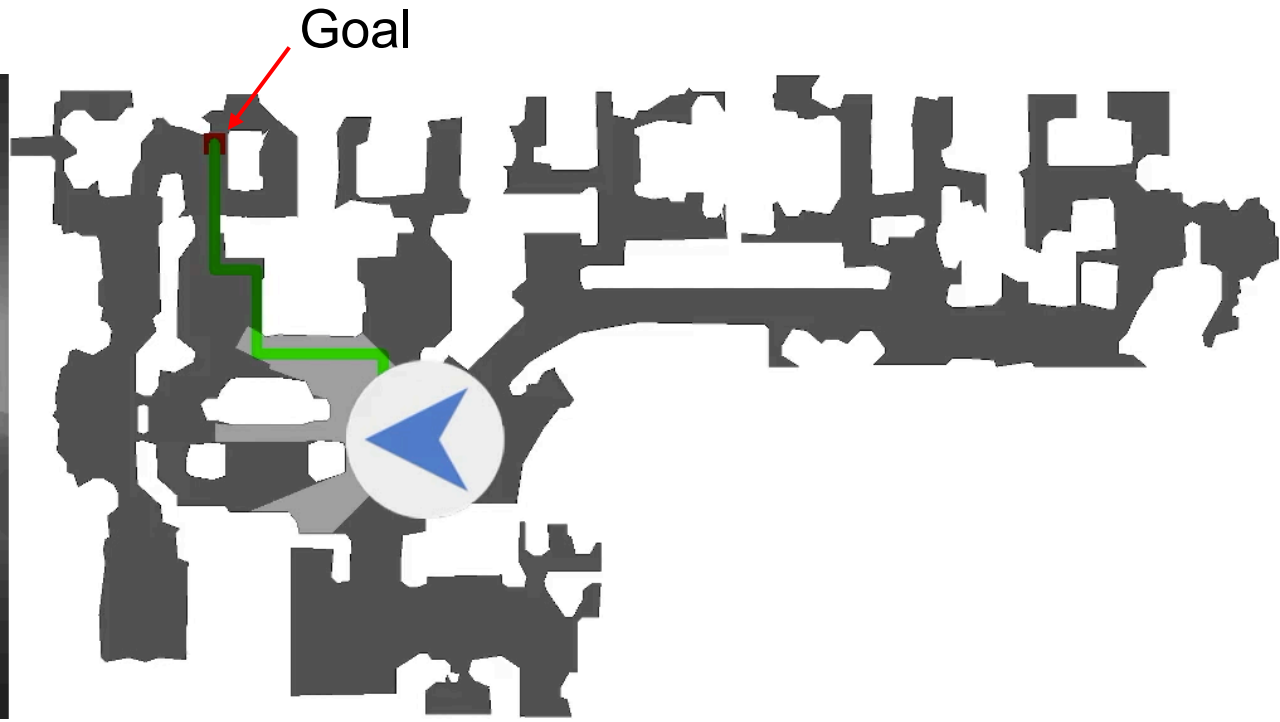
Deep RL for Audio-Visual Navigation



C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Navigation Demo - AudioPointGoal





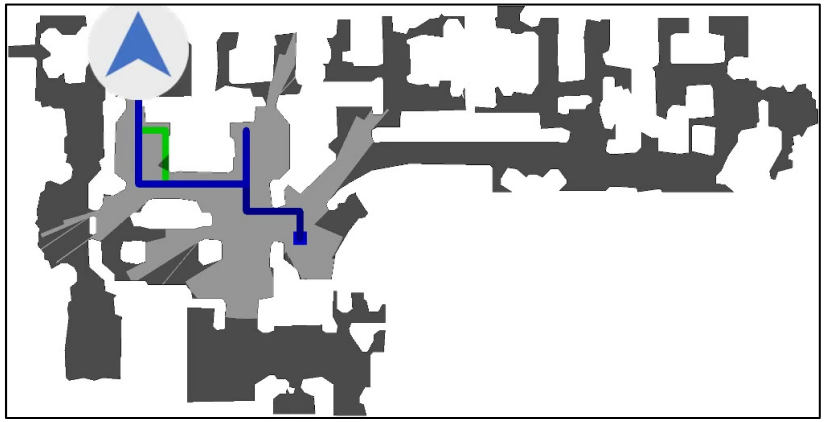
SPL: 1.00

GPS + AudioPointGoal agent leverages the complementary information in audio and GPS, and navigates to the goal efficiently

Agent Goal Start Shortest path Agent path Seen/Unseen area Occupied area Red Frame: Action

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

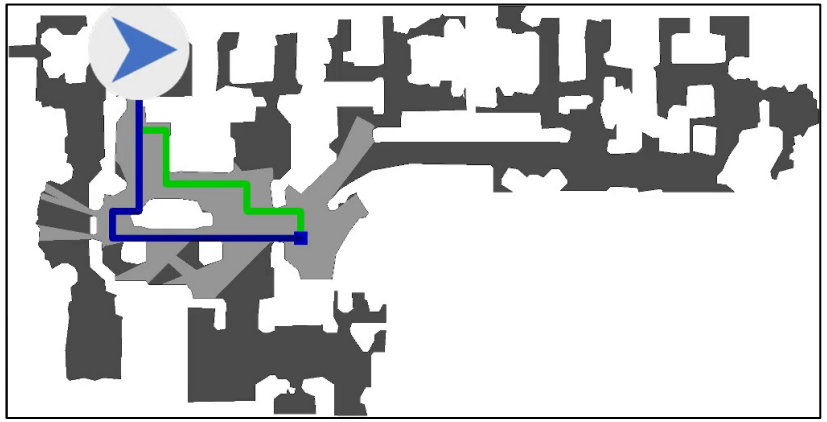
Navigation Trajectory Comparison



SPL: 0.68

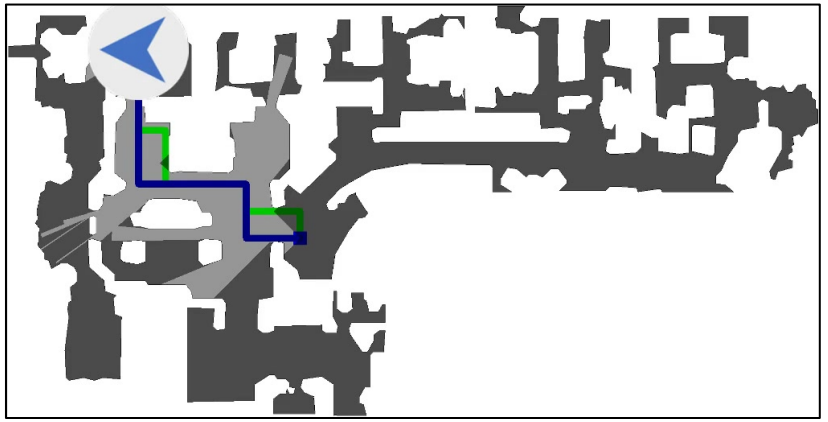
GPS

PointGoal agent gets confused about the direction and gets stuck behind the bed.



SPL: 0.87

AudioGoal agent figures out the sound comes from the front more quickly than the PointGoal agent



SPL: 1.00

GPS +

AudioPointGoal agent knows immediately it should go straight and then right and thus follows the shortest path

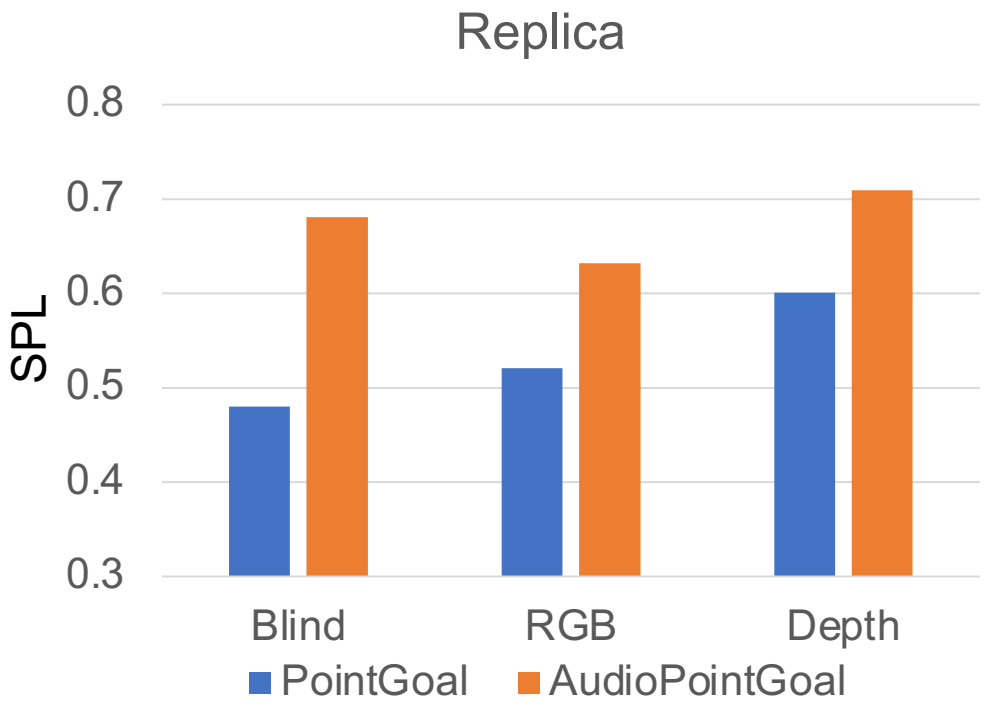
Agent Goal Start Shortest path Agent path Seen/Unseen area Occupied area Red Frame: Sound

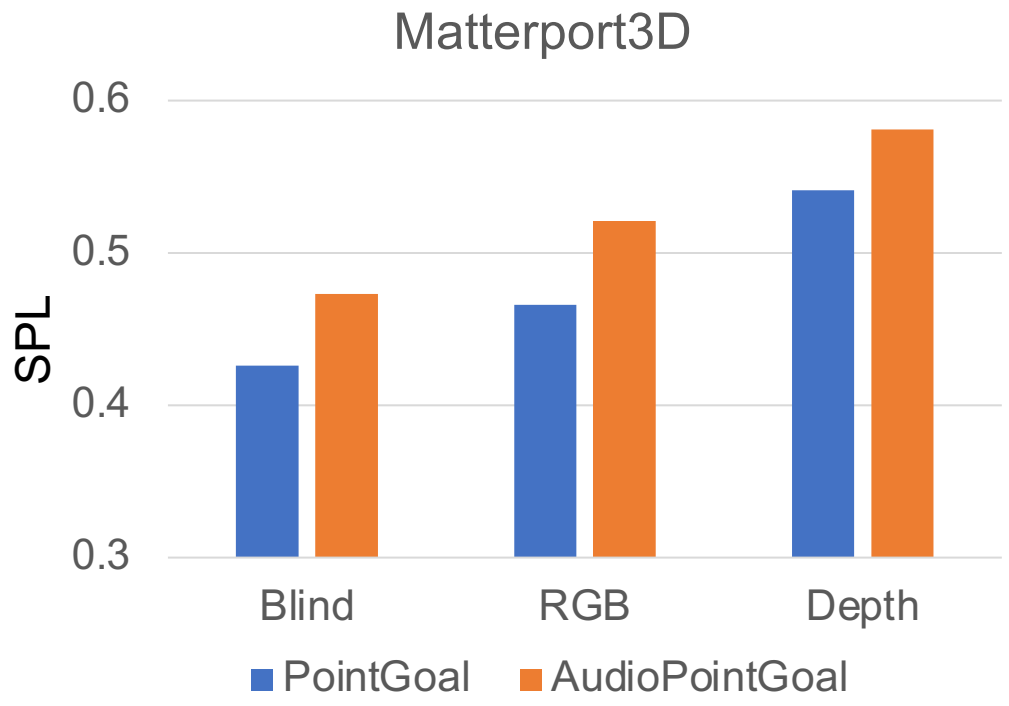
C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Does Audio Help Navigation?

Comparing PointGoal (PG) and AudioPointGoal (APG):

- Audio improves accuracy significantly



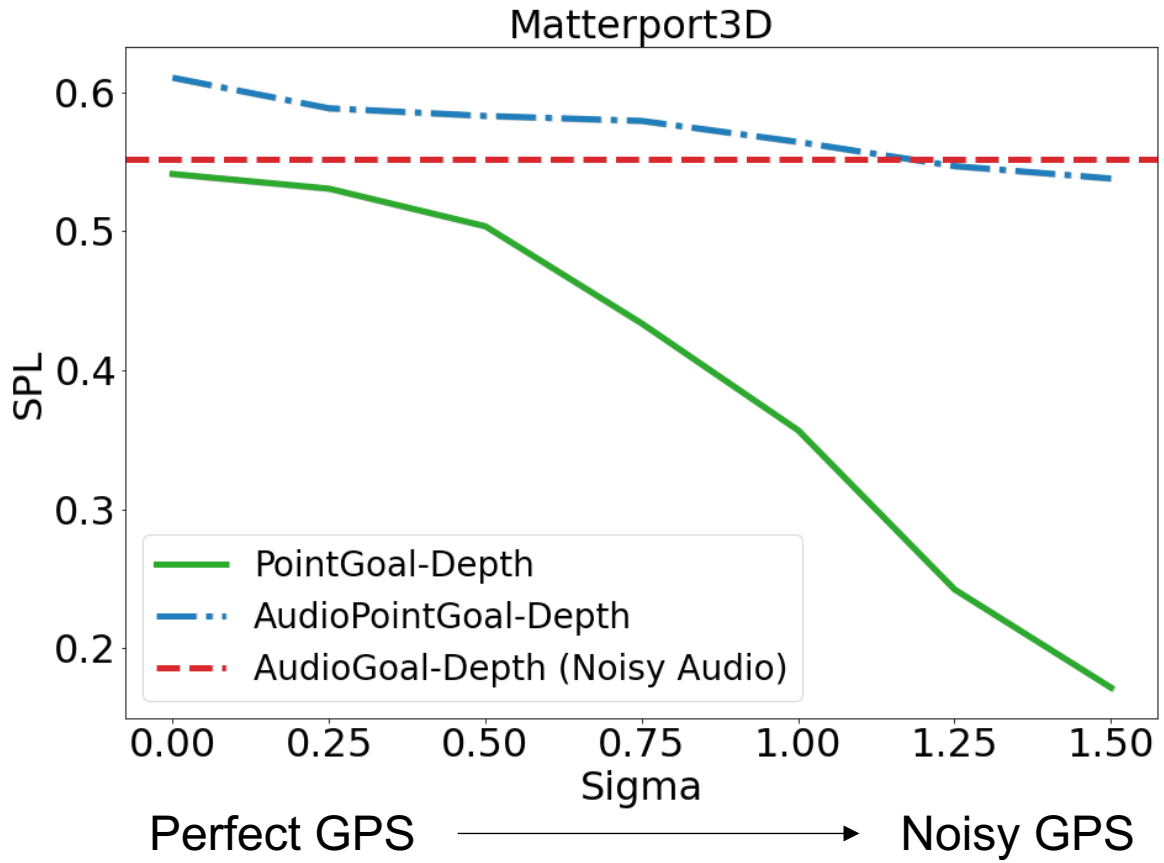


Metric: SPL (success weighted by inverse path length)

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Can Audio Supplant GPS for AudioGoal?

- AudioGoal is immune to GPS noise (localization error) and robust to microphone noise
- AudioPointGoal degrades less in the presence of GPS noise
- Audio signal gives similar or even better spatial cues than the PointGoal displacements

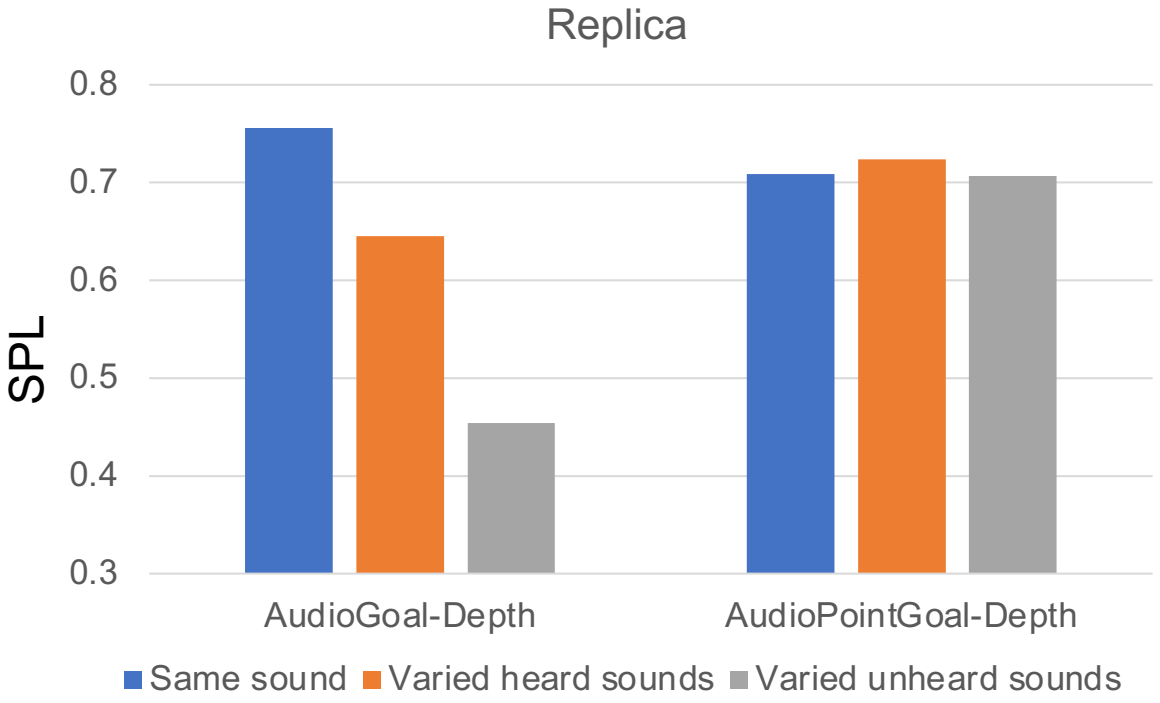


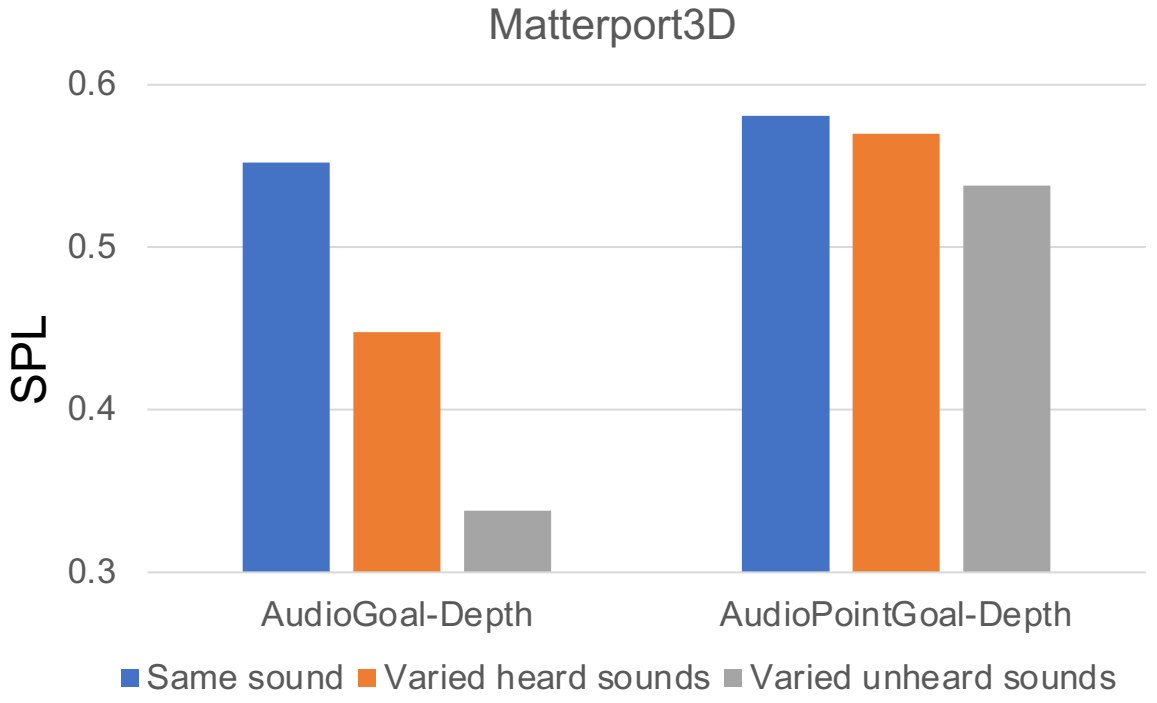
C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Effect of Different Sound Sources

From *same sound* to *varied heard sounds* to *varied unheard sounds*[1]

- AudioGoal accuracy declines with varied heard sounds to unheard sounds
- AudioPointGoal almost always outperforms AudioGoal agent

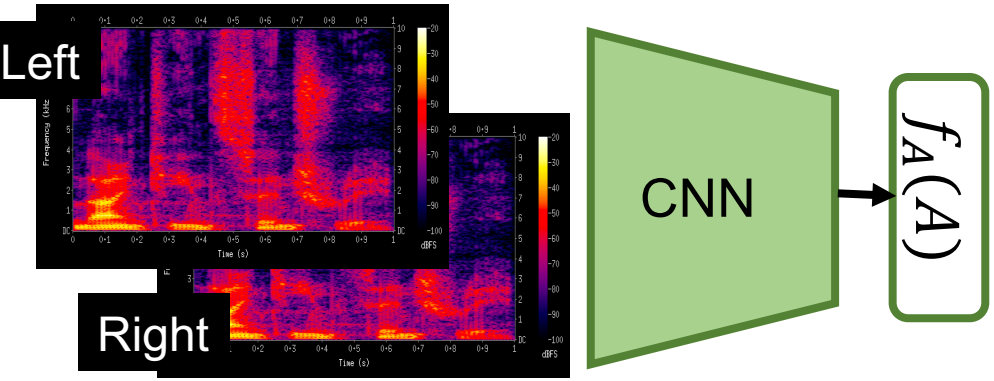


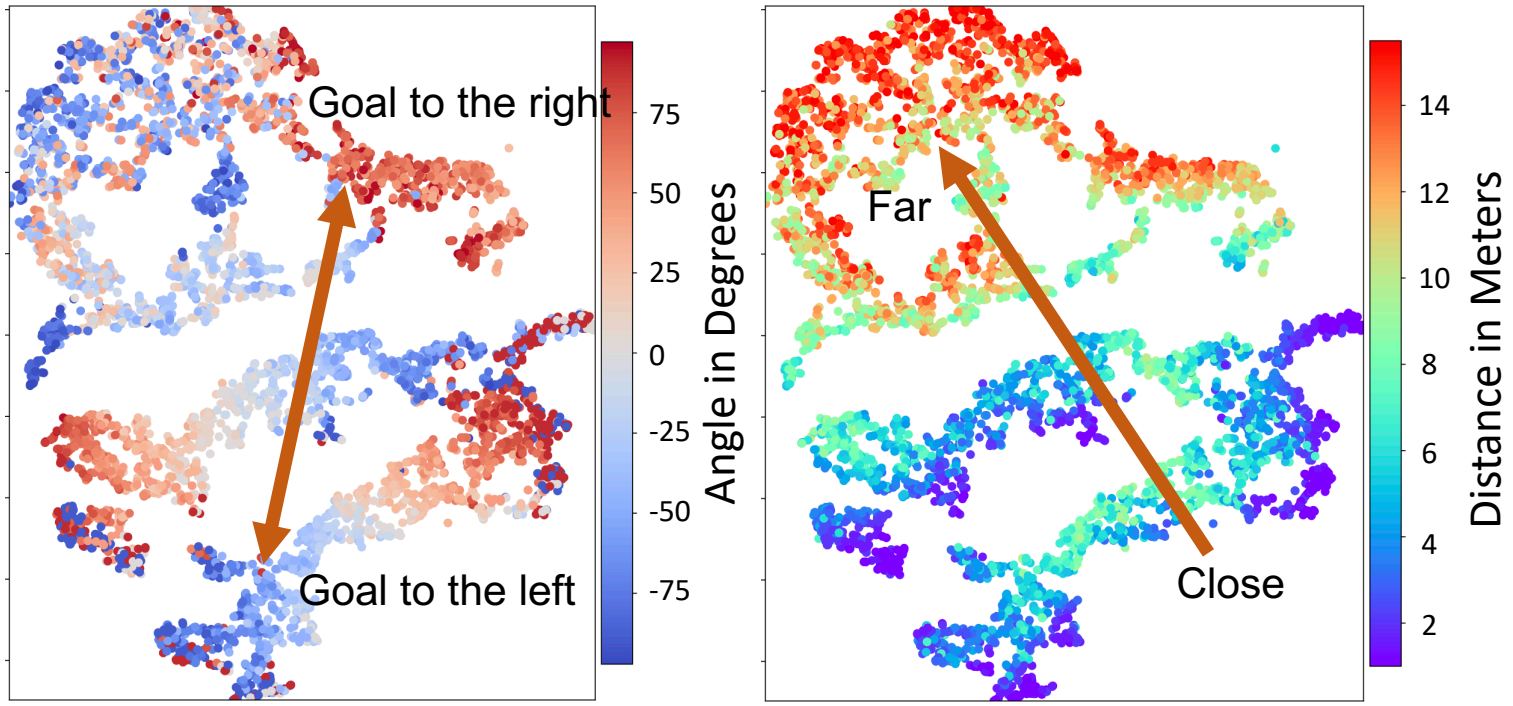


[1]102 copyright-free sounds, divided into 73/11/18 for train/val/test

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

What Do the Learned Audio Features Capture?





T-SNE of audio features from an AudioGoal agent

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Relative Importance of Audio and Vision

Each modality plays an important role in action selection, based on the environment context and goal placement

C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Conclusion

- Introduce task of audio-visual navigation in 3D environments
- Generalize a state-of-the-art deep RL model
- Introduce SoundSpaces: enabling audio rendering for Habitat
- Create a benchmark suite of tasks for audio-visual navigation

SoundSpaces: Audio-Visual Navigation in 3D Environments

Changan Chen[1,4], Unnat Jain*[2,4], Carl Schissler[3], Sebastia V. Amengual Gari[3], Ziad Al-Halah[1], Vamsi K. Ithapu[3], Philip Robinson[3], Kristen Grauman[1,4]*

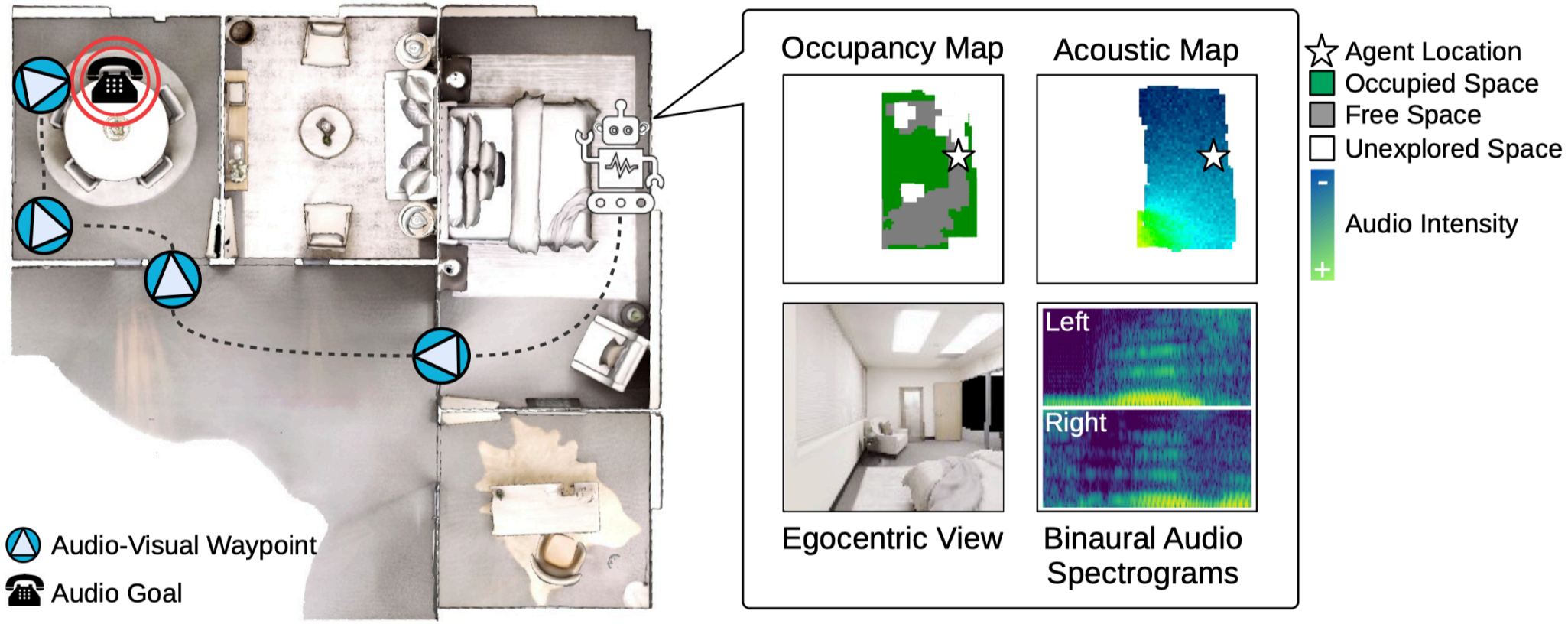
[1]UT Austin, [2]UIUC, [3]Facebook Reality Labs, [4]Facebook AI Research

Code and audio simulation data available at:
http://vision.cs.utexas.edu/projects/audio_visual_navigation

Audio-Visual Waypoints for Navigation

Changan Chen[1,2], Sagnik Majumder[1], Ziad Al-Halah[1], Ruohan Gao[1,2],
Santhosh Kumar Ramakrishnan[1,2], Kristen Grauman[1,2]

UT Austin[1], Facebook AI Research[2]



Model

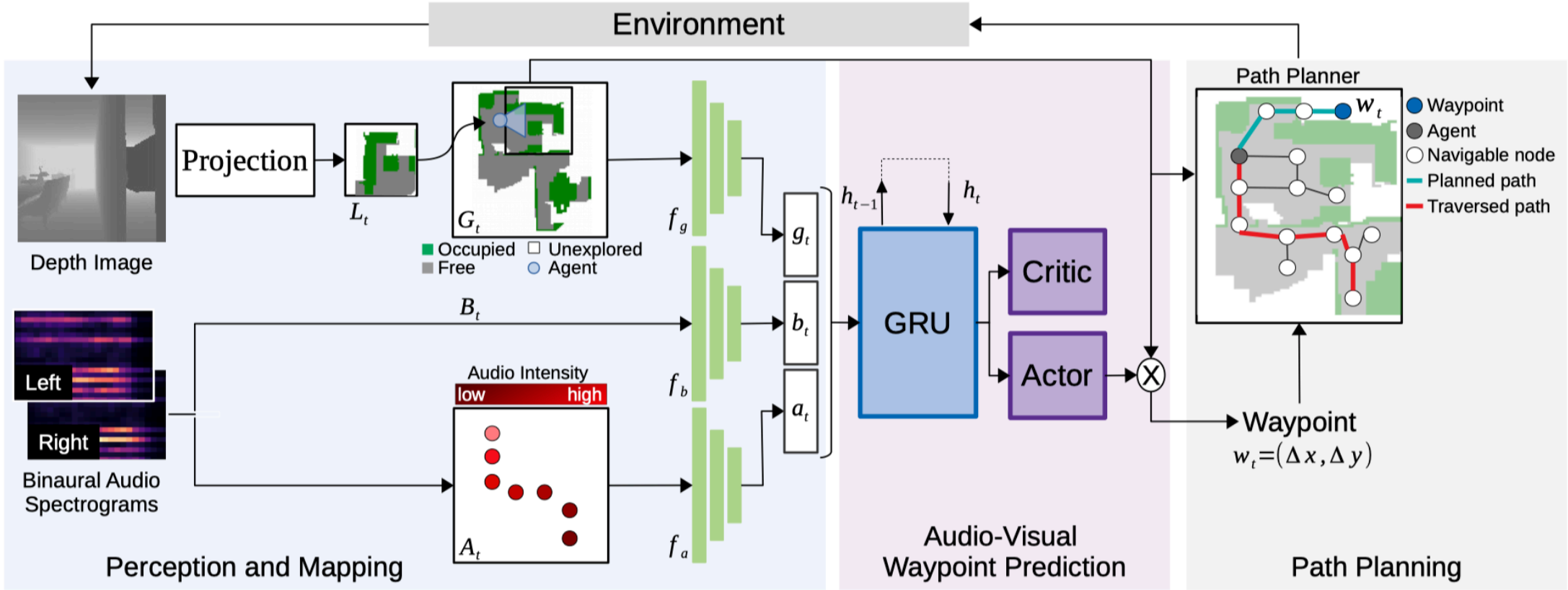
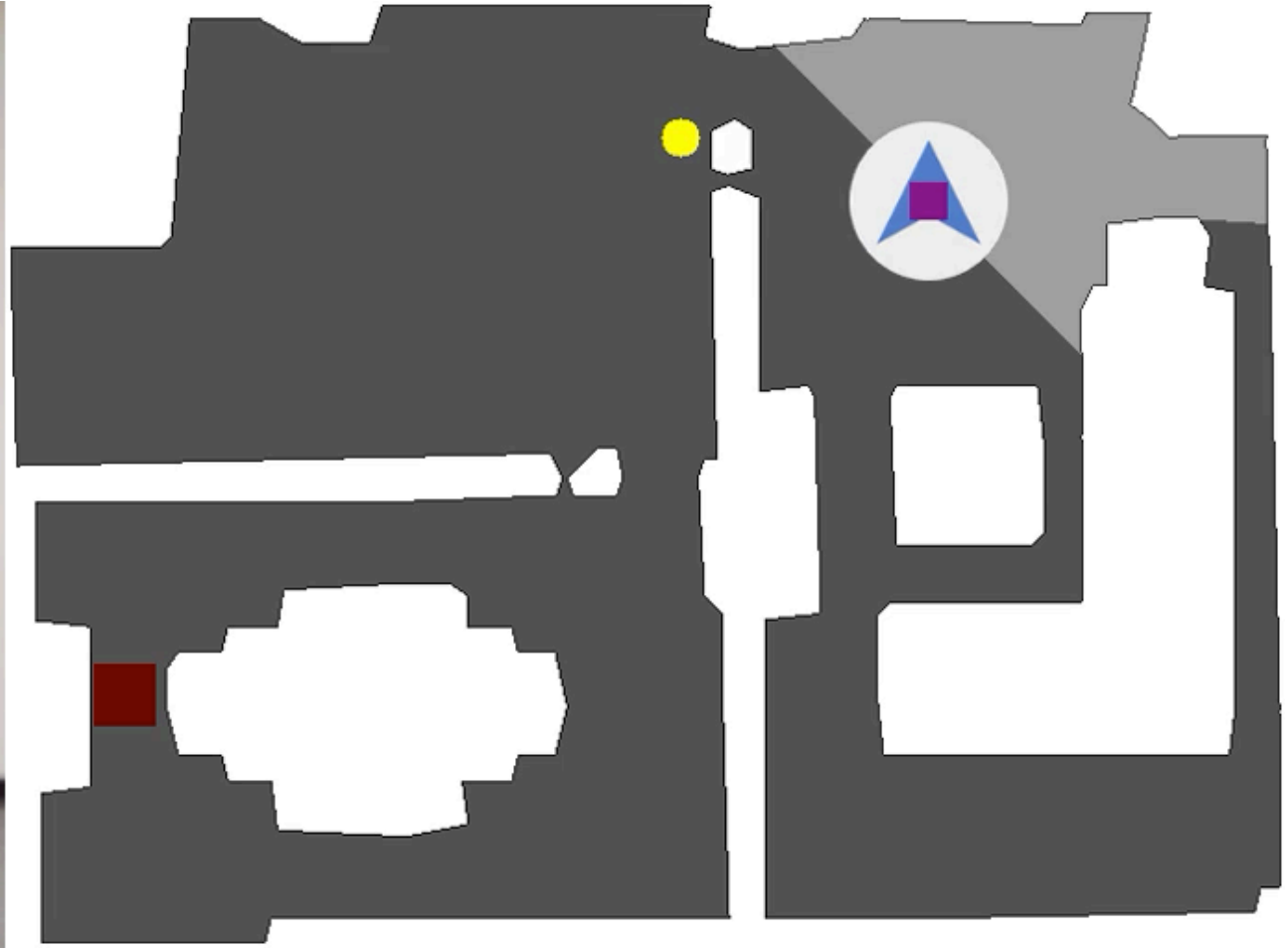


Figure 2: Model architecture. Our audio-visual navigation model uses the egocentric stream of depth images and binaural audio (B_t) to learn geometric (G_t) and acoustic (A_t) maps for the 3D environment. The multi-modal cues and partial maps (left) inform the RL policy's prediction of intermediate waypoints (center). For each waypoint, the agent plans the shortest navigable path (right). From this sequence of waypoints, the agent reaches the final AudioGoal efficiently.

Waypoint Selection and Acoustic Memory



Our model dynamically selects waypoints and builds an acoustic memory as it moves.

Agent Goal Start Waypoint Normalized intensity Seen/Unseen area Occupied area

Experiment Setup

Baselines:

- Random: randomly selects actions and automatically stop when it reaches the goal
- Audio Direction Follower: predicts the audio direction of arrival and moves in that direction
- Frontier Waypoints: intersects the predicted DoA with frontier of the explored area
- Chen and Jain et al.: uses end-to-end RL and predicts primitive actions at each step
- Gan et al.: trains goal predictor and uses geometric planner to reach the goal

Metrics:

- Success (SR): the fraction of successful episodes
- Success weighted by path length (SPL): weighs success with path length
- Normalized distance to goal (NDG): distance to goal relative to the shortest path
- Number of actions (NA): average number of actions per episode
- Success weighted by number of actions (SNA): similar to SPL

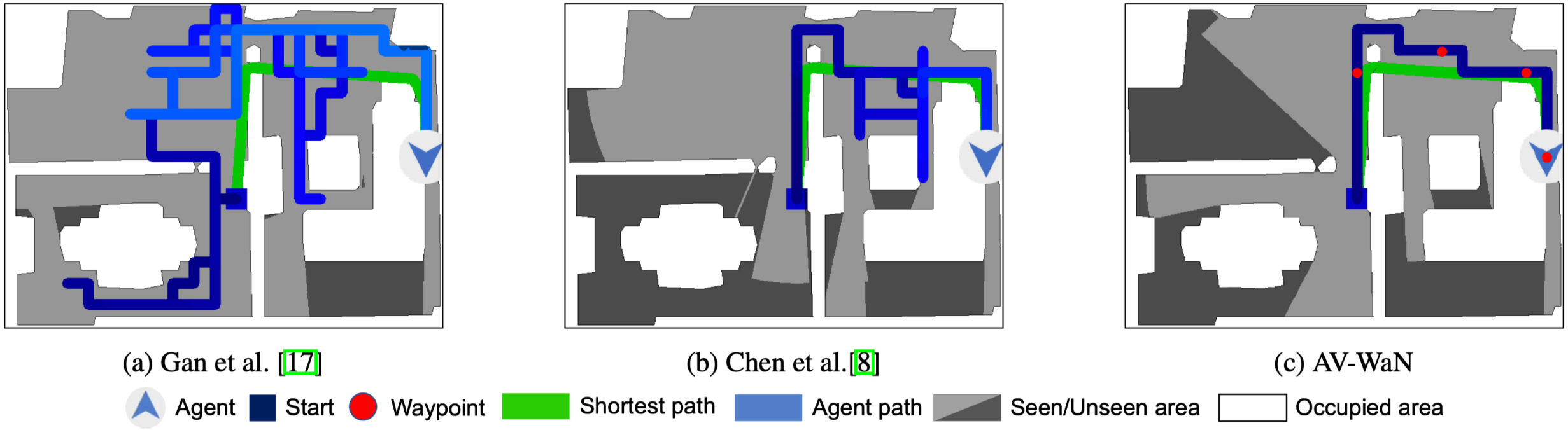
Experiment Results

- AV-WaN strongly outperforms all baselines and existing methods
- Our model improves the agent's efficiency substantially

	Heard sound					*Unheard sounds*				
Model	SPL ↑	SR ↑	NDG ↓	NA ↓	SNA↑	SPL ↑	SR ↑	NDG ↓	NA ↓	SNA↑
Random Agent	4.7	18.5	89.7	448.9	1.7	4.7	18.5	89.7	448.9	1.7
Audio Direction Follower	54.7	72.0	19.5	173.0	41.1	11.1	17.2	55.4	415.2	8.4
Frontier Waypoints	44.0	63.9	47.5	218.5	35.2	6.5	14.8	96.3	434.2	5.1
Gan et al. [17]	57.6	83.1	17.6	130.5	47.9	7.5	15.7	76.3	336.6	5.7
Chen and Jain et al. [8]	75.0	95.7	1.3	84.4	43.4	28.6	41.2	16.0	223.1	14.9
AV-WaN (**Ours**)	**85.3**	**98.0**	**0.6**	**38.8**	**70.0**	**54.4**	**71.2**	**12.3**	**173.6**	**43.8**

Navigation Trajectories

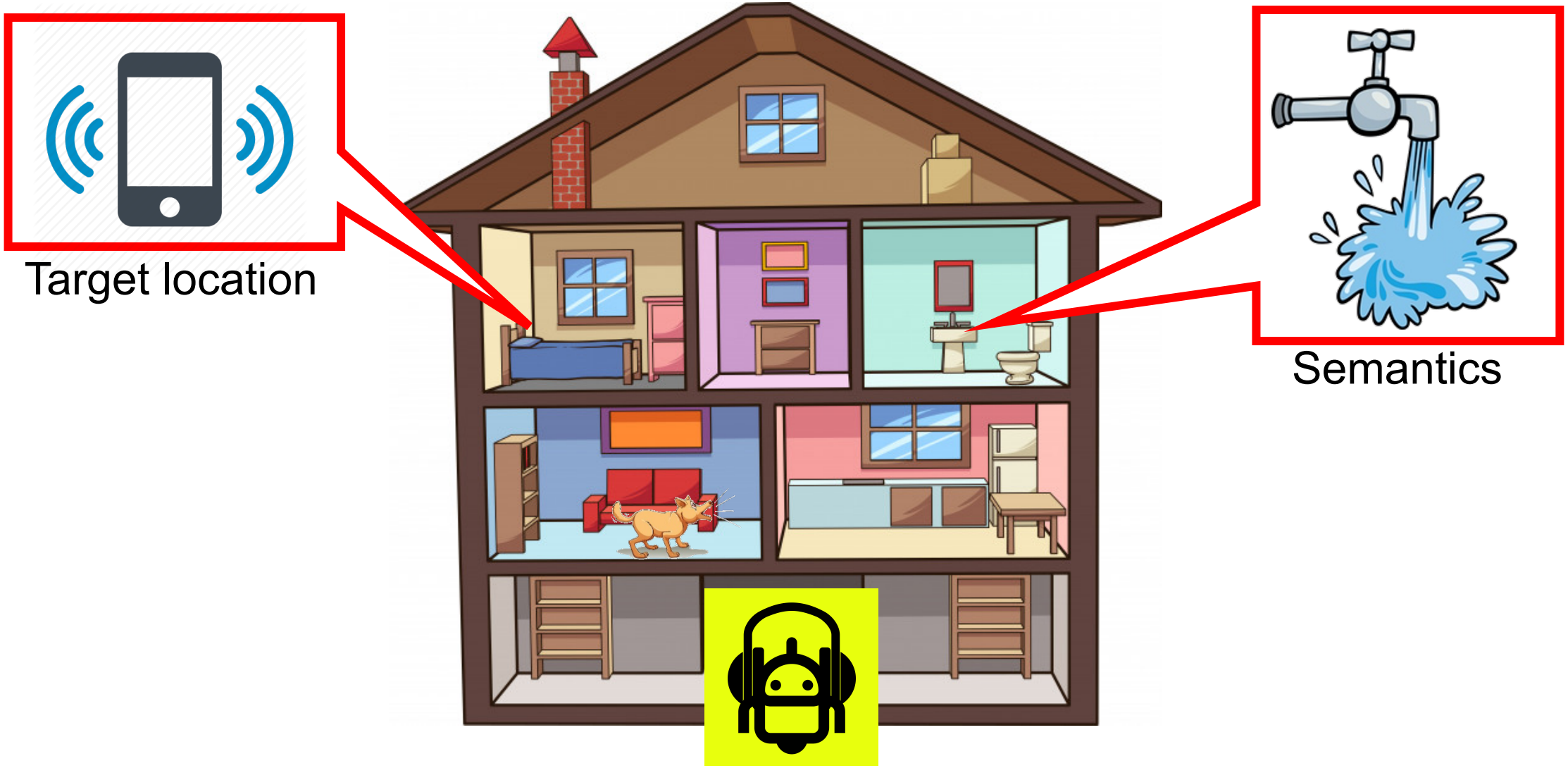
- Gan et al.: is prone to errors and leads the agent to backtrack
- Chen et al.: oscillates around obstacles
- AV-WaN (ours): reaches the goal most efficiently



(a) Gan et al. [17] (b) Chen et al.[8] (c) AV-WaN

Idea: Semantic Audio Navigation

Sound informs navigating agent about…



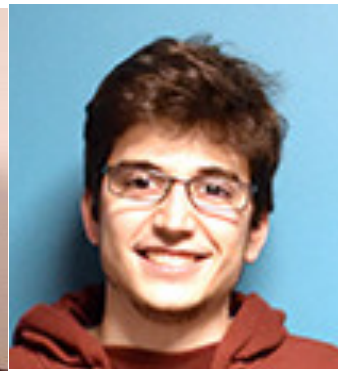
Audio-Visual Navigation

Facebook AI Research
Facebook Reality Labs
University of Texas at Austin


Changan Chen


Unnat Jain

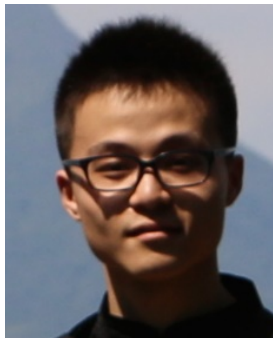

Carl Schissler


Sebastià V. Amengual Garí


Sagnik Majumder


Ziad Al-Halah


Santhosh Ramakrishnan


Ruohan Gao


Vamsi Krishna Ithapu


Philip Robinson


Kristen Grauman