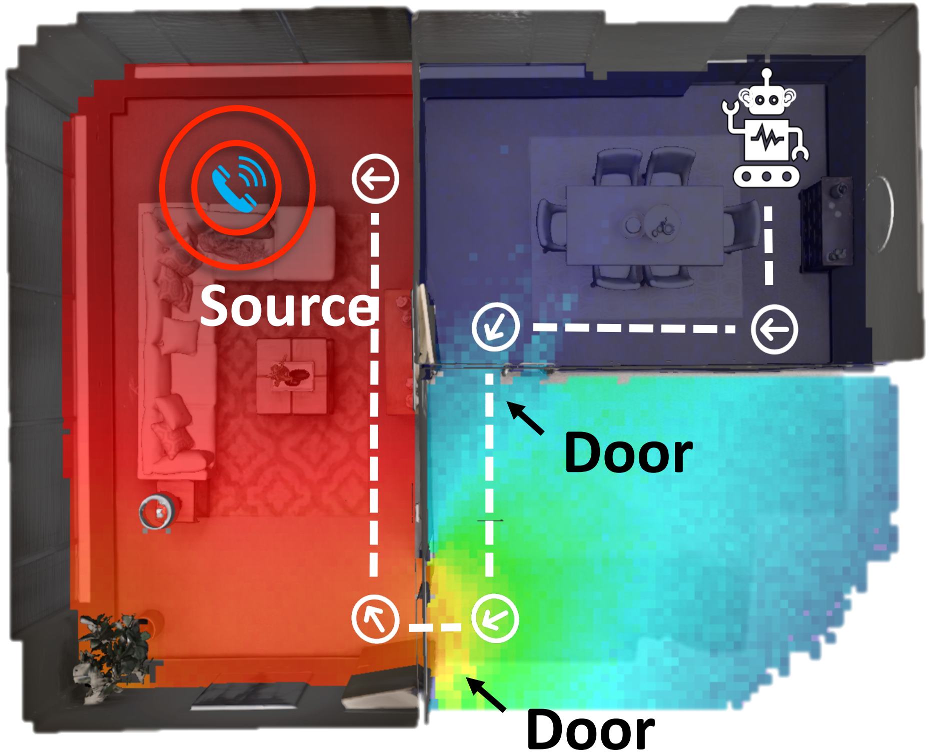
SoundSpaces: Audio-Visual Navigation in 3D Environments

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

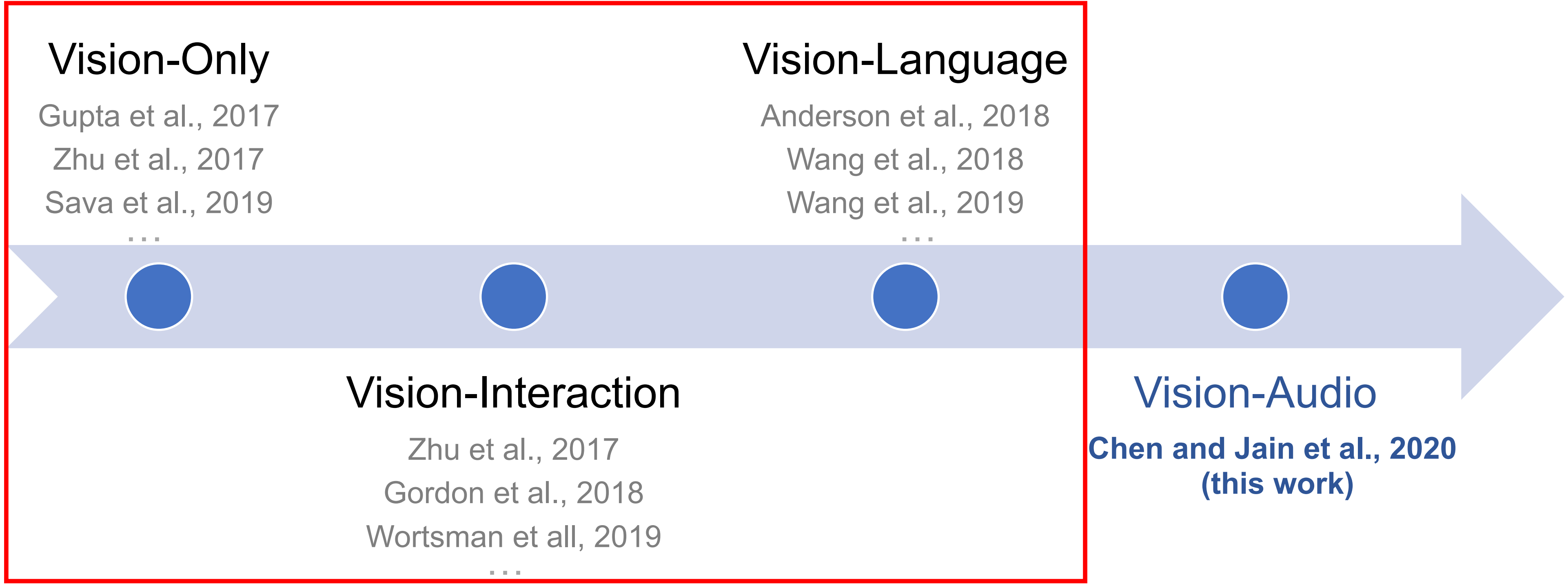


ECCV'20
SEC, Glasgow
23-28 AUGUST 2020

Embodied Perception Is a Multisensory Experience

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

Today’s agents are deaf!



Vision-Only

Gupta et al., 2017
Zhu et al., 2017
Sava et al., 2019
…

Vision-Interaction

Zhu et al., 2017
Gordon et al., 2018
Wortsman et all, 2019
…

Vision-Language

Anderson et al., 2018
Wang et al., 2018
Wang et al., 2019
…

Vision-Audio

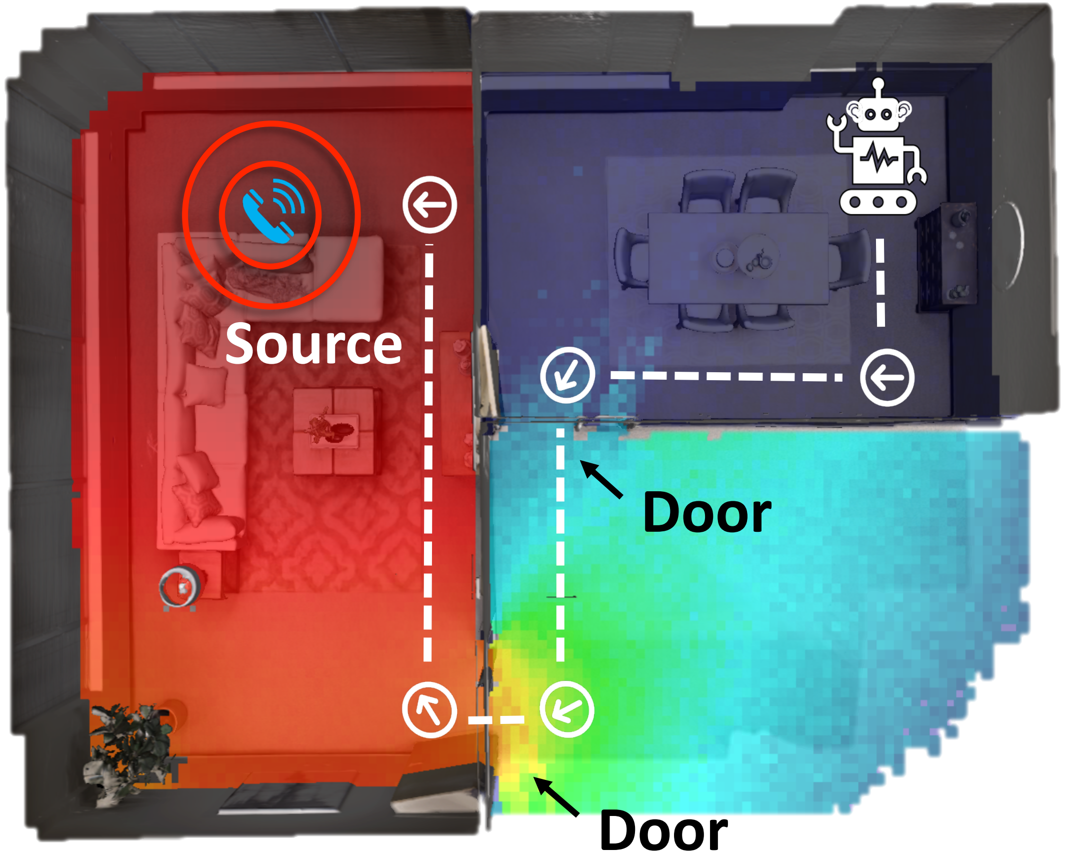
Chen and Jain et al., 2020
(this work)

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

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

Audio-Visual Navigation in 3D Environments

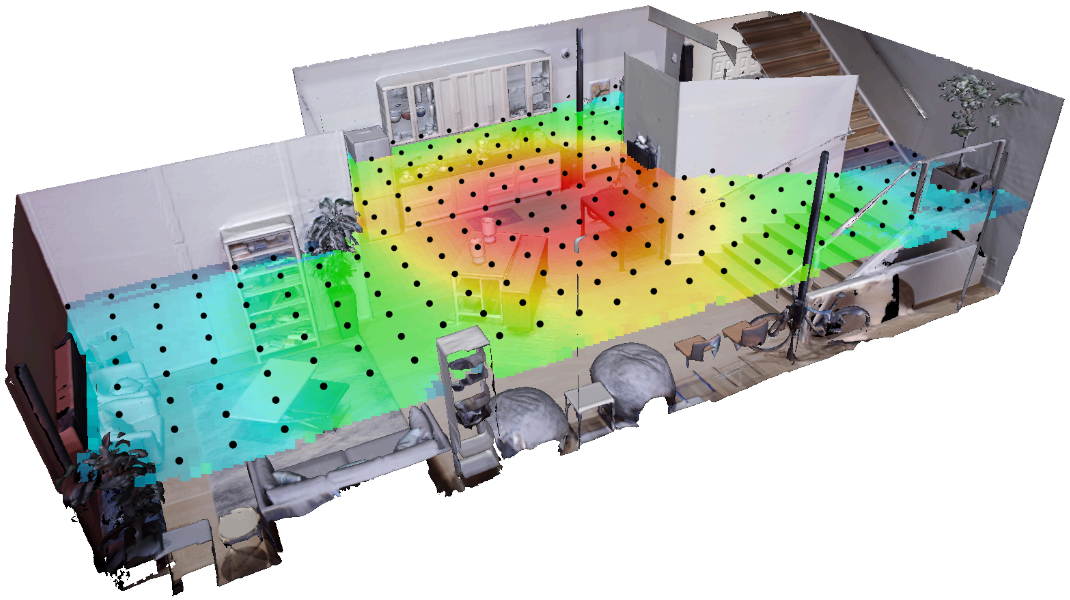
An agent navigates to a sounding object with vision and audio perception



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

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]

- 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

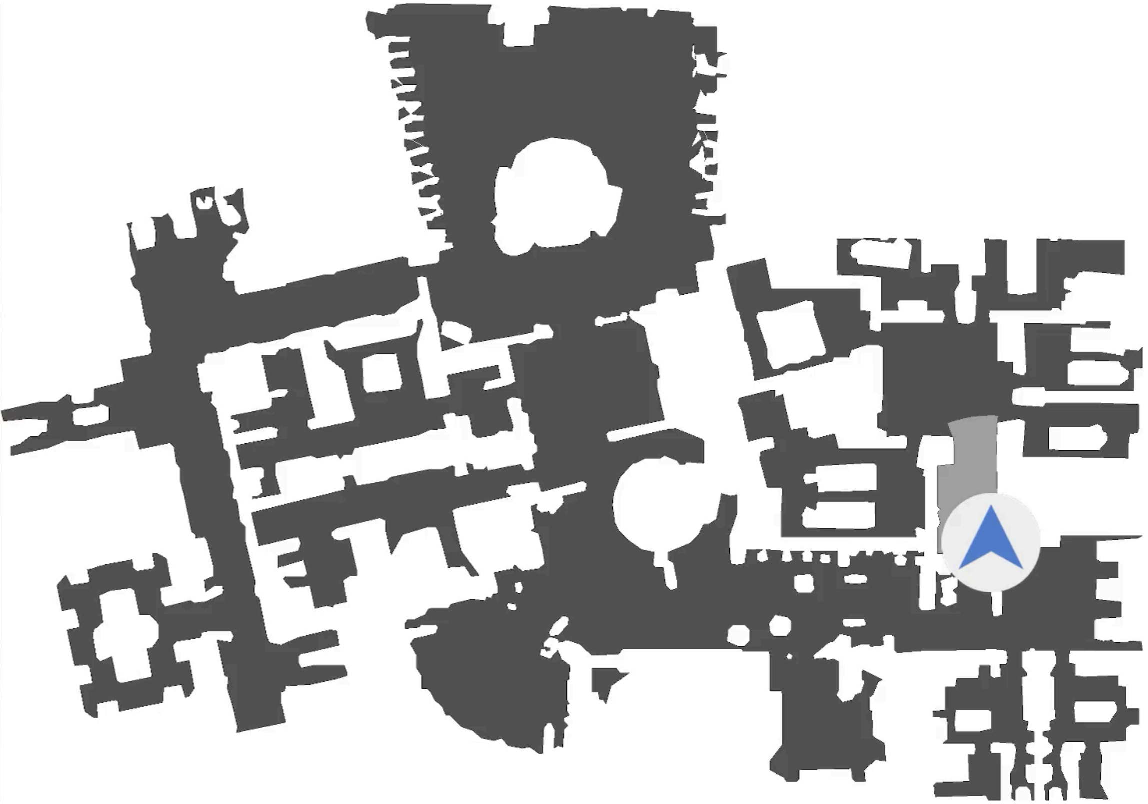
[1] The Replica Dataset: A Digital Replica of Indoor Spaces, Straub et al., arXiv, 2019
[2] Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV, 2017

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

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

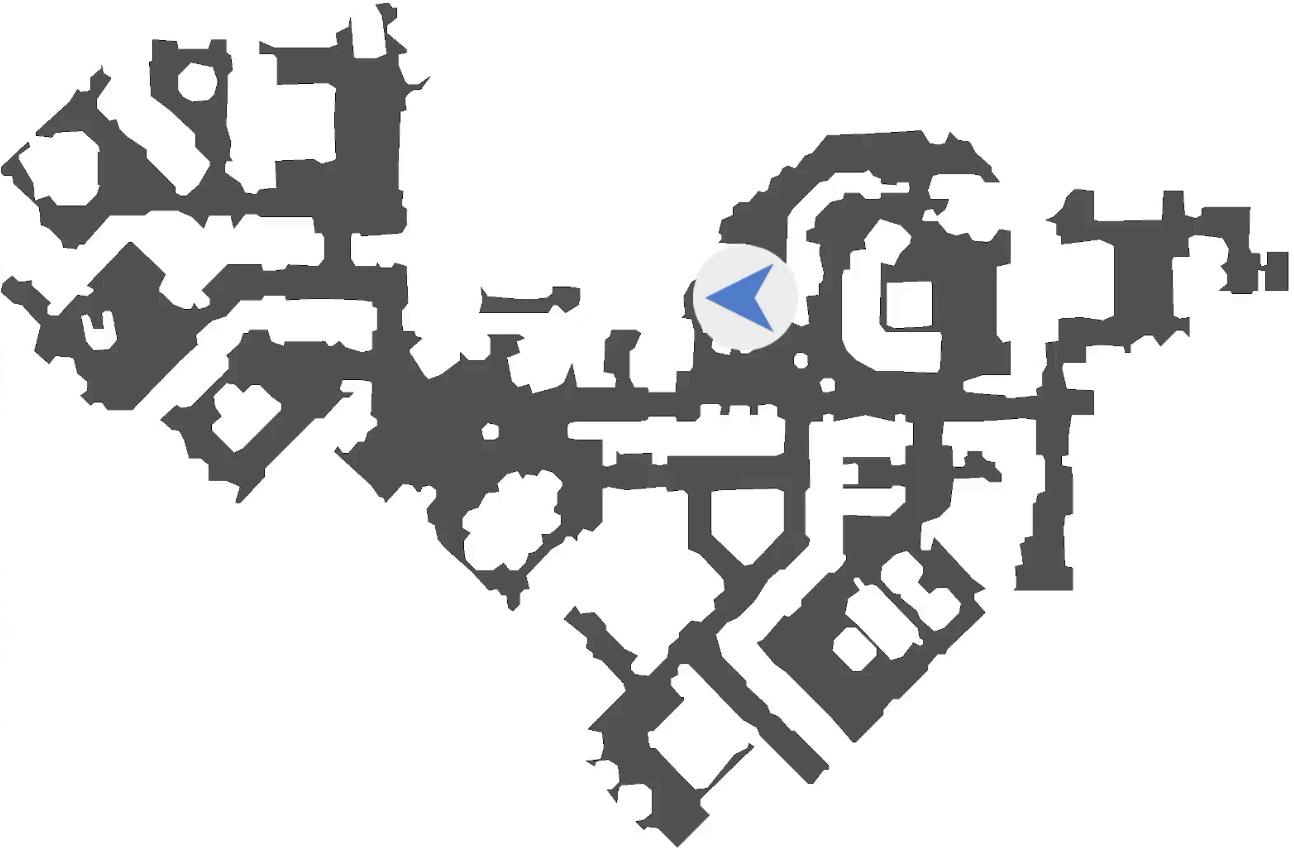
▲ Agent 🟥 Goal ■ Start 🟩 Shortest path 🟦 Agent path ◩ Seen/Unseen area □ Occupied area

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

PointGoal

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

GPS

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

New tasks

AudioGoal

🔊

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

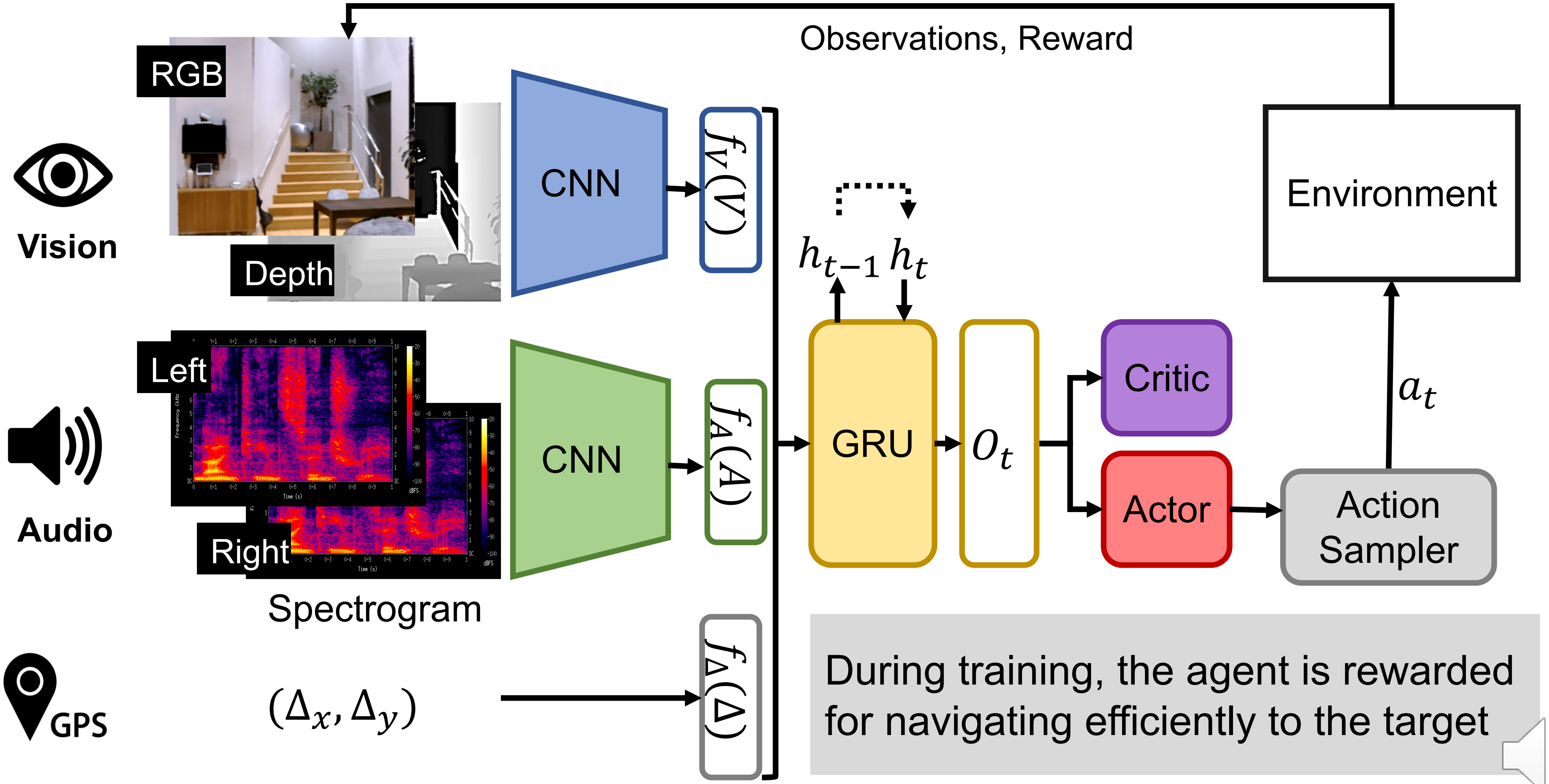
AudioPointGoal

🔊 + GPS

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

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

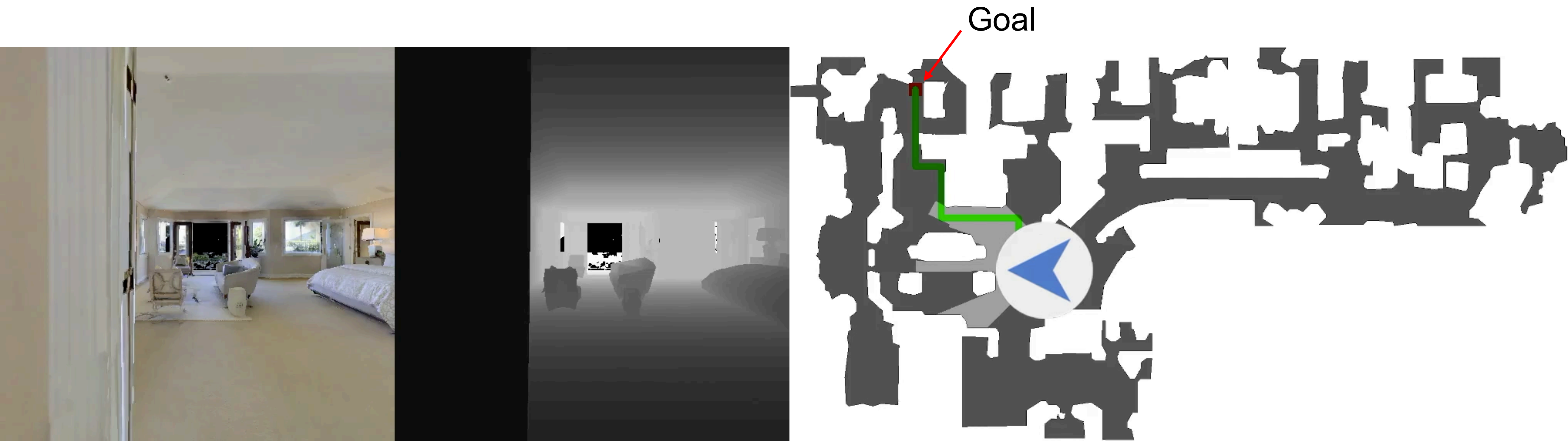
Deep RL for Audio-Visual Navigation



During training, the agent is rewarded for navigating efficiently to the target

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

Navigation Demo - AudioPointGoal



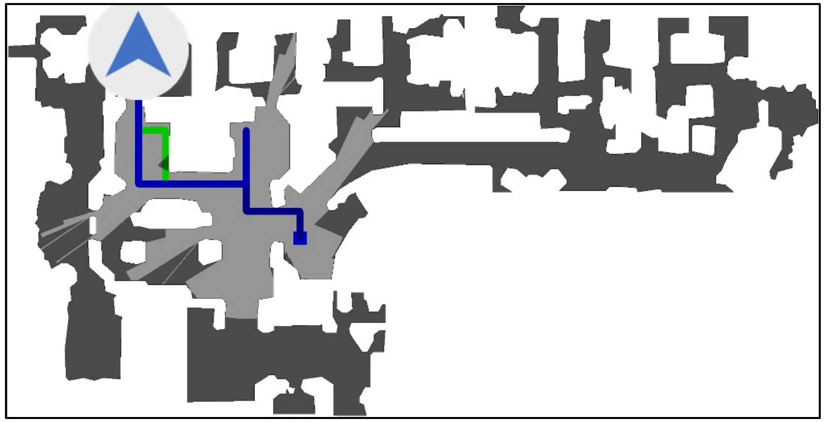
SPL: 1.00

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: ision

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

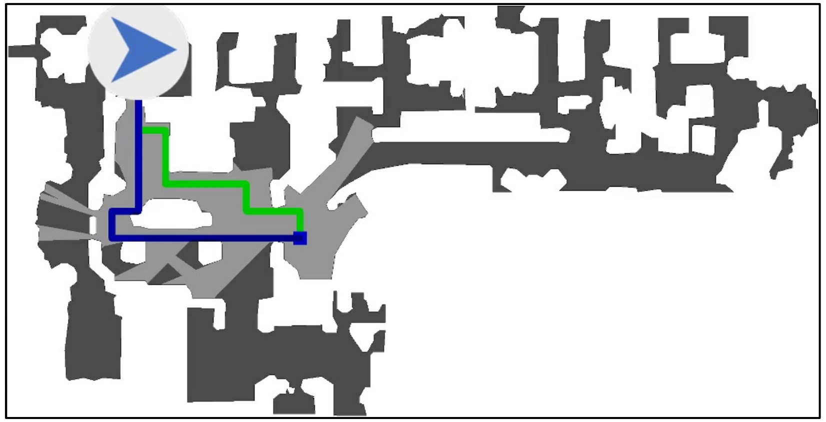
Navigation Trajectory Comparison



SPL: 0.68

📍GPS

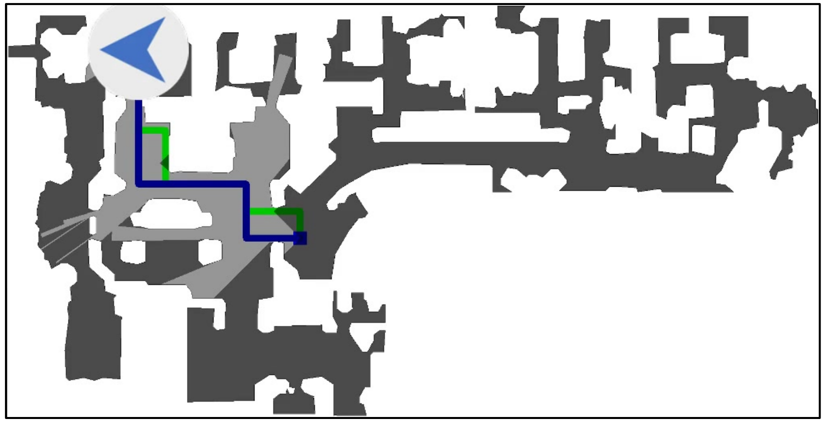
PointGoal agent gets confused about the direction and gets stuck behind the bed.



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

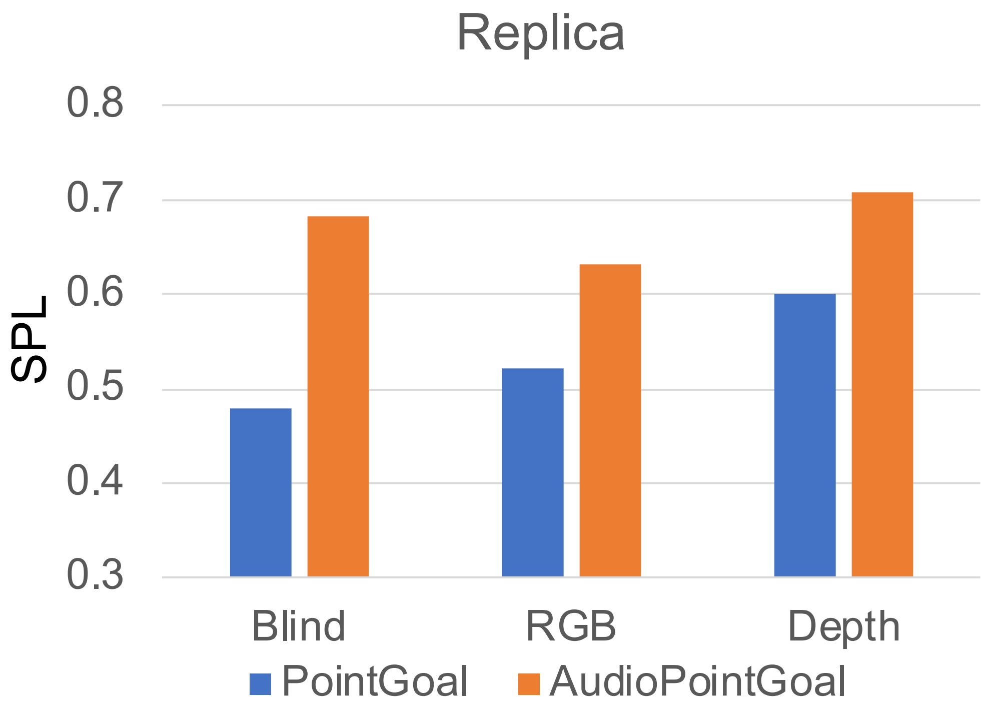
▲ Agent 🟥 Goal ■ Start 🟩 Shortest path 🟦 Agent path ▨ Seen/Unseen area □ Occupied area ▣ Red Frame: ision

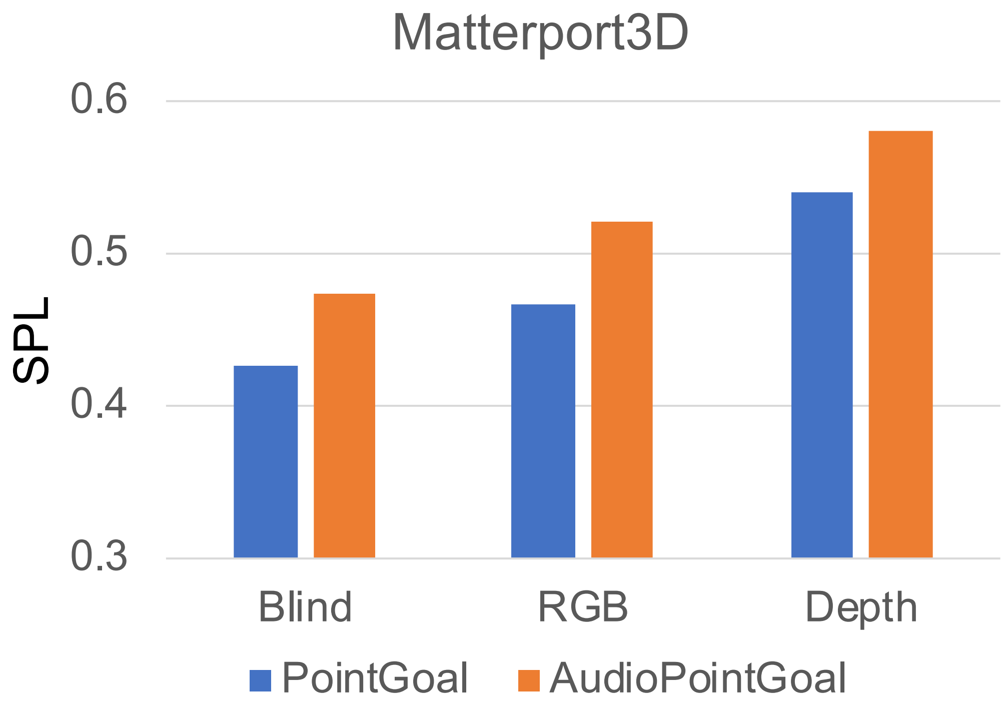
C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Does Audio Help Navigation?

Comparing PointGoal (PG) and AudioPointGoal (APG):

- Audio improves accuracy significantly



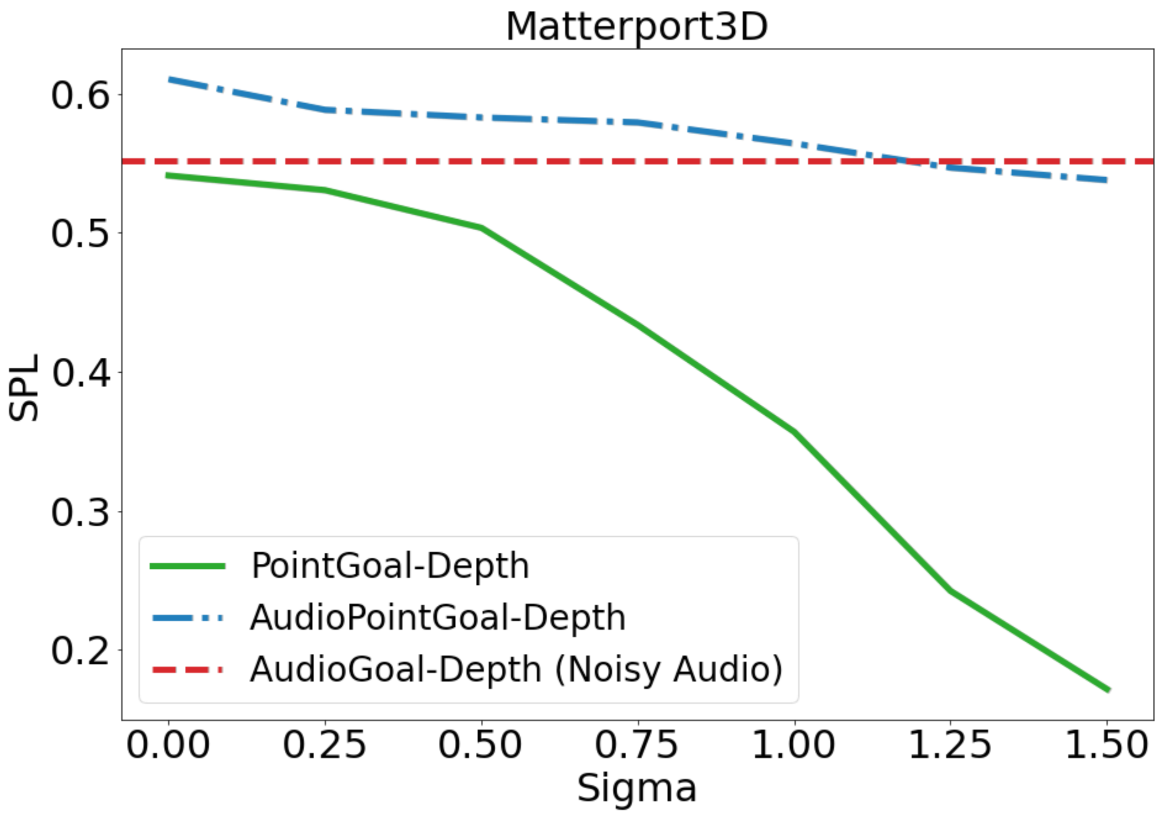


Metric: SPL (success weighted by inverse path length)

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



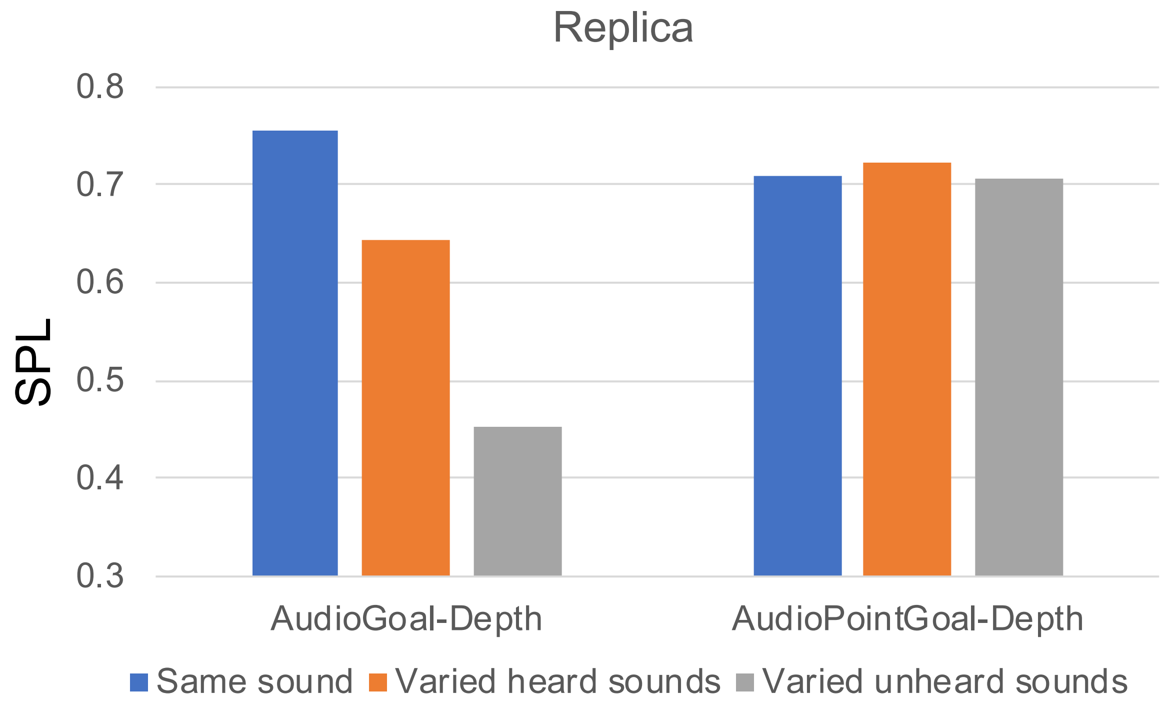
Perfect GPS → Noisy GPS

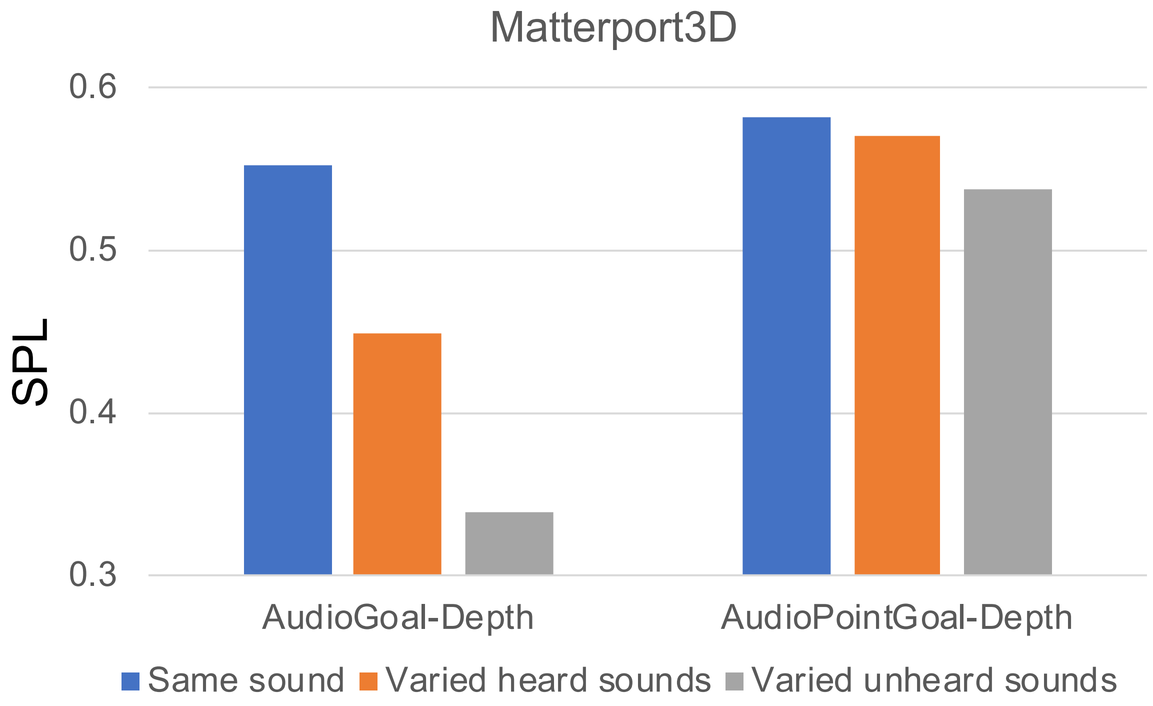
C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

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

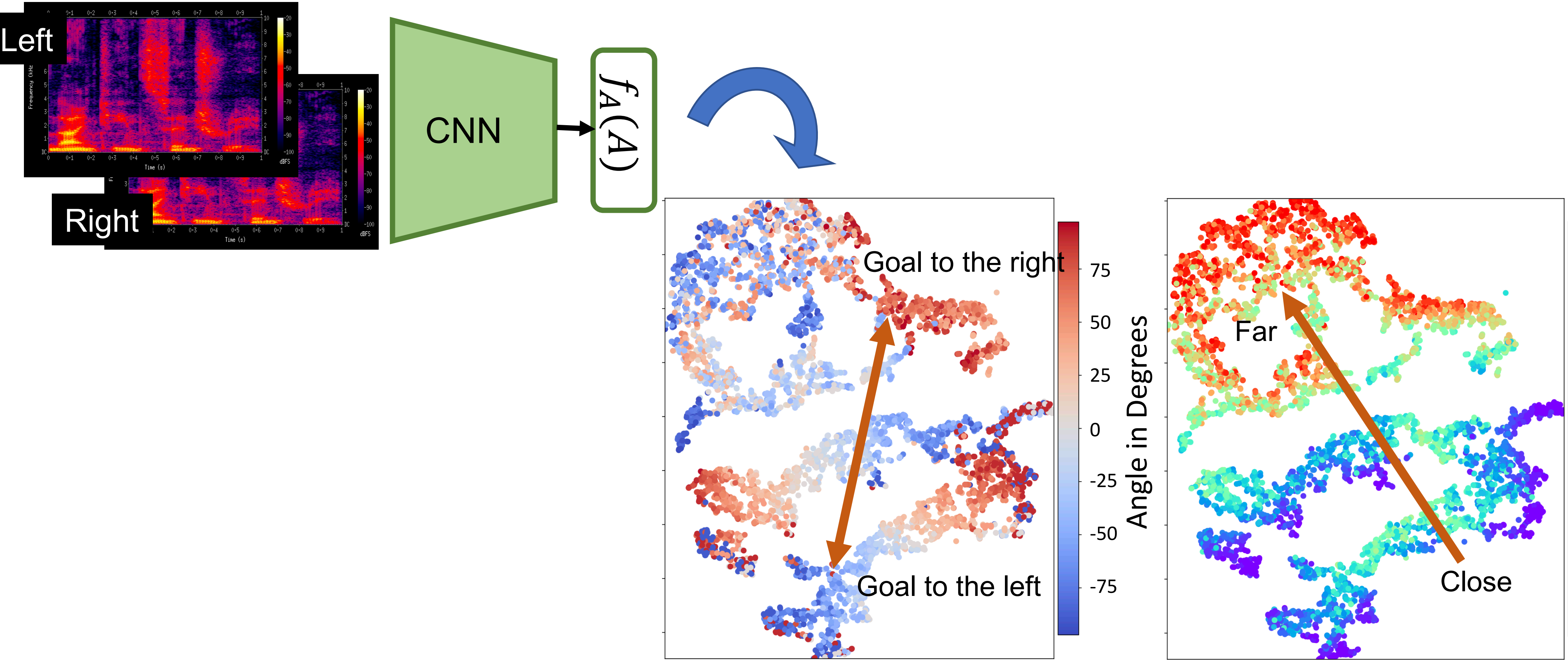




[1] 102 copyright-free sounds, divided into 73/11/18 for train/val/test

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

What Do the Learned Audio Features Capture?

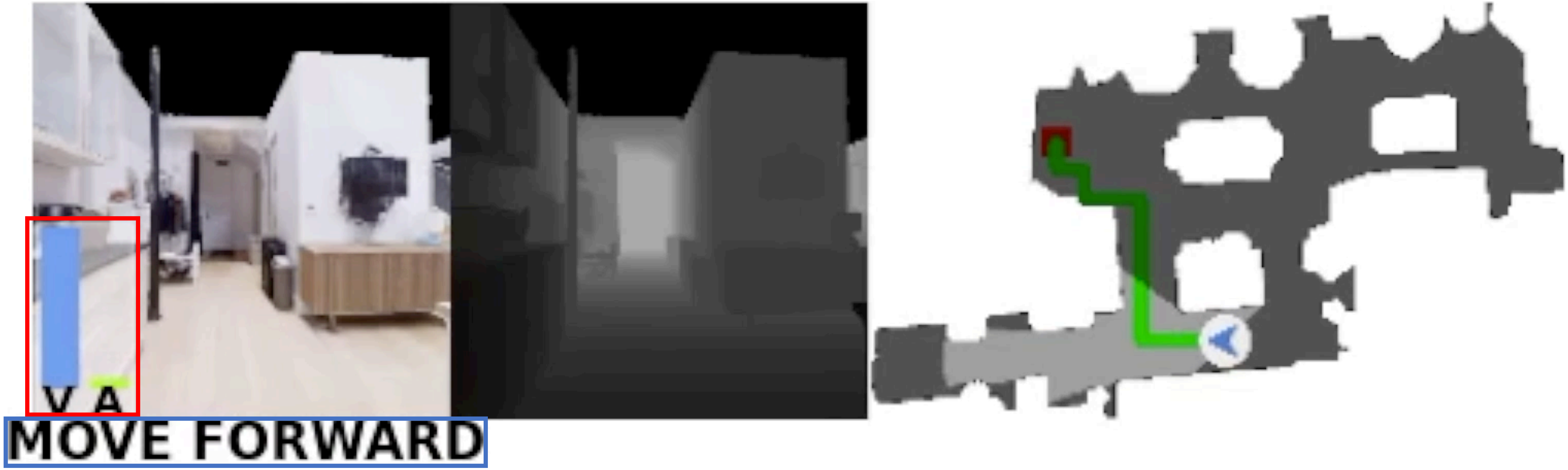


T-SNE of audio features from an AudioGoal agen

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



▲ Agent 🟥 Goal ■ Start 🟩 Shortest path ■ Agent path ▨ Seen/Unseen area □ Occupied area □ Red Frame: ision

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

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

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

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