LEARNING TO SET WAYPOINTS FOR AUDIO-VISUAL NAVIGATION

Changan Chen[1,2] **Sagnik Majumder**[1] **Ziad Al-Halah**[1] **Ruohan Gao**[1,2]
Santhosh K. Ramakrishnan[1,2] **Kristen Grauman**[1,2]
[1]UT Austin [2]Facebook AI Research

ABSTRACT

In audio-visual navigation, an agent intelligently travels through a complex, unmapped 3D environment using both sights and sounds to find a sound source (e.g., a phone ringing in another room). Existing models learn to act at a fixed granularity of agent motion and rely on simple recurrent aggregations of the audio observations. We introduce a reinforcement learning approach to audio-visual navigation with two key novel elements: 1) waypoints that are dynamically set and learned end-to-end within the navigation policy, and 2) an acoustic memory that provides a structured, spatially grounded record of what the agent has heard as it moves. Both new ideas capitalize on the synergy of audio and visual data for revealing the geometry of an unmapped space. We demonstrate our approach on two challenging datasets of real-world 3D scenes, Replica and Matterport3D. Our model improves the state of the art by a substantial margin, and our experiments reveal that learning the links between sights, sounds, and space is essential for audio-visual navigation. Project: `http://vision.cs.utexas.edu/projects/audio_visual_waypoints`. Please see the full paper for more details.

1 INTRODUCTION

Intelligent robots must be able to move around efficiently in the physical world. In addition to geometric maps and planning, work in embodied AI shows the promise of agents that *learn* to map and navigate. Sensing directly from egocentric images, they jointly learn a spatial memory and navigation policy in order to quickly reach target locations in novel, unmapped 3D environments (Gupta et al., 2017b;a). High quality simulators have accelerated this research direction to the point where policies learned in simulation can (in some cases) successfully translate to robotic agents deployed in the real world (Gupta et al., 2017a; Chaplot et al., 2020).

Much current work centers around visual navigation by a PointGoal agent that has been told where to find the target (Gupta et al., 2017a; Sax et al., 2018; Mishkin et al., 2019; Savva et al., 2019; Chaplot et al., 2020). However, in the recently introduced AudioGoal task, the agent must use both visual and auditory sensing to travel through an unmapped 3D environment to find a sound-emitting object, without being told where it is (Chen et al., 2020; Gan et al., 2020). As a learning problem, AudioGoal not only has strong motivation from cognitive and neuroscience (Gougoux et al., 2005; Lessard et al., 1998), it also has compelling real-world significance: a phone is ringing somewhere upstairs; a person is calling for help from another room; a dog is scratching at the door to go out.

What role should audio-visual inputs play in learning to navigate? There are two existing strategies. One employs deep reinforcement learning to learn a navigation policy that generates step-by-step actions (TurnRight, MoveForward, etc.) based on both modalities (Chen et al., 2020). This has the advantage of unifying the sensing modalities, but can be inefficient when learning to make long sequences of individual local actions. The alternative approach separates the modalities—treating the audio stream as a beacon that signals the goal location, then planning a path to that location using a visual mapper (Gan et al., 2020). This strategy has the advantage of modularity, but the disadvantage of restricting audio's role to localizing the target. Furthermore, both existing methods make strong assumptions about the granularity at which actions should be predicted, either myopically for each step (0.5 to 1 m) (Chen et al., 2020) or globally for the final goal location (Gan et al., 2020).

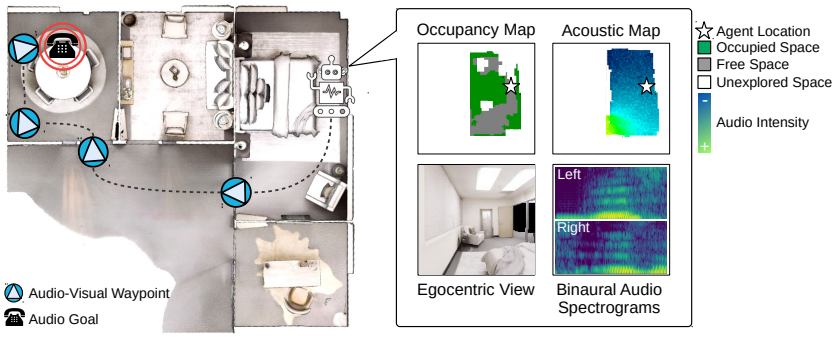


Figure 1: Waypoints for audio-visual navigation: Given egocentric audio-visual sensor inputs (depth and binaural sound), the proposed agent builds up both geometric and acoustic maps (top right) as it moves in the unmapped environment. The agent learns encodings for the multi-modal inputs together with a modular navigation policy to find the sounding goal (e.g., phone ringing in top left corner room) via a series of dynamically generated audio-visual waypoints. For example, the agent in the bedroom may hear the phone ringing, identify that it is in another room, and decide to first exit the bedroom. It may then narrow down the phone location to the dining room, decide to enter it, and subsequently find it. Whereas existing hierarchical navigation methods rely on heuristics to determine subgoals, our model learns a policy to set waypoints jointly with the navigation task.

We introduce a new approach for AudioGoal navigation where the agent instead predicts non-myopic actions with self-adaptive granularity. Our key insight is to *learn to set audio-visual waypoints*: the agent dynamically sets intermediate goal locations based on its audio-visual observations and partial map—and does so in an end-to-end manner with learning the navigation task. Intuitively, it is often hard to directly localize a distant sound source from afar, but it can be easier to identify the general direction (and hence navigable path) along which one could move closer to that source. See Figure 1.

Both the audio and visual modalities are critical to identifying waypoints in an unmapped environment. Audio input suggests the general goal direction; visual input reveals intermediate obstacles and free spaces; and their interplay indicates how the geometry of the 3D environment is warping the sounds received by the agent, such that it can learn to trace back to the hidden goal. In contrast, subgoals selected using only visual input are limited to mapped locations or clear line-of-sight paths.

To realize our idea, our first contribution is a novel deep reinforcement learning approach for AudioGoal navigation with audio-visual waypoints (Figure 2). The model is hierarchical, with an outer policy that generates waypoints and an inner module that plans to reach each waypoint. Hierarchical policies for 3D navigation are not new, e.g., Chaplot et al. (2020); Stein et al. (2018); Bansal et al. (2019). However, whereas existing visual navigation methods employ heuristics to define subgoals, the proposed agent *learns to set useful subgoals in an end-to-end fashion for the navigation task.* This is a new idea for 3D visual navigation subgoals in general, not specific to audio goals. As a second technical contribution, we introduce an *acoustic memory* to record what the agent hears as it moves, complementing its visual spatial memory. Whereas existing models aggregate audio evidence purely based on an unstructured memory (GRU), our proposed acoustic map is structured, interpretable, and integrates audio observations throughout the reinforcement learning pipeline.

We demonstrate our approach on the complex 3D environments of Replica and Matterport3D using SoundSpaces audio (Chen et al., 2020). It outperforms the state of the art for AudioGoal navigation by a substantial margin (8 to 49 points in SPL on heard sounds, Table 1), and generalizes much better to the challenging cases of unheard sounds, noisy audio, and distractor sounds. Our results show learning to set waypoints in an end-to-end fashion outperforms current subgoal approaches, while the proposed acoustic memory helps the agent set goals more intelligently.

2 Experiments

Environments We test with SoundSpaces for Replica and Matterport environments in the Habitat simulator. We follow the protocol of the SoundSpaces AudioGoal benchmark (Chen et al., 2020), with train/val/test splits of 9/4/5 scenes on Replica and 73/11/18 scenes on Matterport3D. We stress that the test and train/val environments are disjoint, requiring the agent to learn generalizable behaviors. Furthermore, for the same scene splits, we experiment with training and testing on disjoint sounds, requiring the agent to generalize to unheard sounds. For heard-sound experiments, the telephone ringing is the sound source; for unheard, we draw from 102 unique sounds.

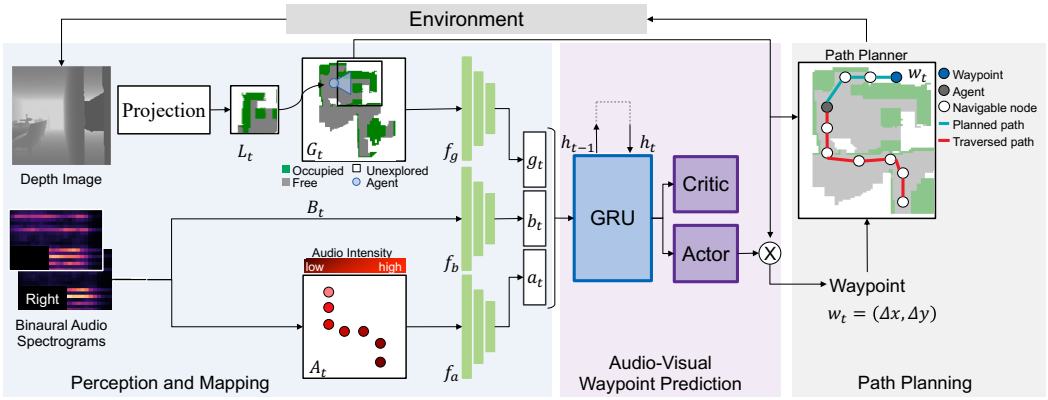


Figure 2: Model architecture. Our audio-visual navigation model uses the egocentric stream of depth images and binaural audio (B_t) to learn geometric (G_t) and acoustic (A_t) maps for the 3D environment. The multi-modal cues and partial maps (left) inform the RL policy's prediction of intermediate waypoints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

Metrics We evaluate the following navigation metrics: 1) success rate (SR), the fraction of successful episodes, i.e., episodes in which the agent stops exactly at the audio goal location on the grid; 2) success weighted by path length (SPL), the standard metric (Anderson et al., 2018) that weighs successes by their adherence to the shortest path; 3) success weighted by number of actions (SNA), which penalizes rotation in place actions, which do not lead to path changes. Please see the full paper for more details on these comprehensive metrics.

Existing methods and baselines We compare the following methods:

– **Random**: an agent that randomly selects each action and signals *Stop* when it reaches the goal.

– **Direction Follower**: a hierarchical model that sets intermediate goals K meters away in the audio's predicted direction of arrival (DoA), and repeats. K is estimated through a hyperparameter search on the validation split, which yields $K = 2$ in Replica and $K = 4$ in Matterport. We train a separate classifier based on audio input to predict when this agent should stop.

– **Frontier Waypoints**: a hierarchical model that intersects the predicted DoA with the frontiers of the explored area and selects that point as the next waypoint. Frontier waypoints are commonly used in the visual navigation literature, e.g., (Stein et al., 2018; Chaplot et al., 2020), making this a broadly representative baseline for standard practice.

– **Supervised Waypoints**: a hierarchical model that uses the RGB frame and audio spectrogram to predict waypoints in its field of view (FoV) with supervised (non-end-to-end) learning. This model is inspired by Bansal et al. (2019), which learns to predict waypoints in a supervised fashion.

– **Chen et al. (2020)**: a state-of-the-art end-to-end AudioGoal RL agent that selects actions using audio-visual observations. It lacks any geometric or acoustic maps. We run the authors' code.

– **Gan et al. (2020)**: a state-of-the-art AudioGoal agent that predicts the audio goal location from binaural spectrograms alone and then navigates with an analytical path planner on an occupancy map it progressively builds by projecting depth images. It uses a separate audio classifier to stop. We adapt the model to improve its performance on Replica and Matterport, since the authors originally tested on a game engine simulator.

Navigation results We consider two settings: 1) *heard sound*—train and test on the telephone sound, following (Chen et al., 2020; Gan et al., 2020), and 2) *unheard sounds*—train and test with disjoint sounds, following (Chen et al., 2020). In both cases, the test environment is always unseen, hence both settings require generalization.

Table 1 shows the results. We refer to our model as AV-WaN (**A**udio-**V**isual **Wa**ypoint **N**avigation). Random does poorly due to the challenging nature of the AudioGoal task and the complex 3D environments. For the heard sound, AV-WaN strongly outperforms all the other methods—with 8.4% and 29% SPL gains on Replica compared to Chen et al. (2020) and Gan et al. (2020), and 17.2% and

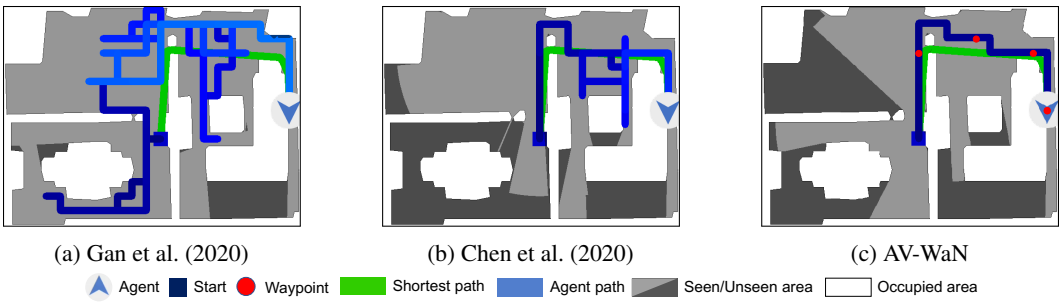
Table 1: AudioGoal navigation results. Our audio-visual waypoints navigation model (AV-WaN) reaches the goal faster (higher SPL) and it is more efficient (higher SNA) compared to the state-of-the-art. SPL, SR, SNA are shown as percentages. For all metrics, higher is better. (H) denotes a hierarchical model.

	Replica						*Matterport3D*					
	Heard			*Unheard*			*Heard*			*Unheard*		
Model	SPL	SR	SNA	SPL	SR	SNA	SPL	SR	SNA	SPL	SR	SNA
Random Agent	4.9	18.5	1.8	4.9	18.5	1.8	2.1	9.1	0.8	2.1	9.1	0.8
Direction Follower (H)	54.7	72.0	41.1	11.1	17.2	8.4	32.3	41.2	23.8	13.9	18.0	10.7
Frontier Waypoints (H)	44.0	63.9	35.2	6.5	14.8	5.1	30.6	42.8	22.2	10.9	16.4	8.1
Supervised Waypoints (H)	59.1	88.1	48.5	14.1	43.1	10.1	21.0	36.2	16.2	4.1	8.8	2.9
Gan et al. (2020)	57.6	83.1	47.9	7.5	15.7	5.7	22.8	37.9	17.1	5.0	10.2	3.6
Chen et al. (2020)	78.2	94.5	52.7	34.7	50.9	16.7	55.1	71.3	32.6	25.9	40.1	12.8
AV-WaN (**Ours**) (H)	**86.6**	**98.7**	**70.7**	**34.7**	**52.8**	**27.1**	**72.3**	**93.6**	**54.8**	**40.9**	**56.7**	**30.6**

49.5% gains on Matterport. This result shows the advantage of our dynamic audio-visual waypoints and structured acoustic map, compared to the myopic action selection in Chen et al. (2020) and the final-goal prediction in Gan et al. (2020). We find that the RL model of Chen et al. (2020) fails when it oscillates around an obstacle. Meanwhile, predicting the final audio goal location, as done by Gan et al. (2020), is prone to errors and leads the agent to backtrack or change course often to redirect itself towards the goal. This result emphasizes the difficulty of the audio-visual navigation task itself; simply reducing the task to PointGoal after predicting the goal location from audio (as done in Gan et al. (2020)) is much less effective than the proposed model. See Figure 3.

Our method also surpasses all three other hierarchical models. This highlights our advantage of directly *learning* to set waypoints, versus the heuristics used in current hierarchical visual navigation models. Even the Supervised Waypoints model does not generalize as well to unseen environments as AV-WaN. We expect this is due to the narrow definition of the optimal waypoint posed by supervision compared to our model, which learns from its own experience what is the best waypoint for the given navigation task in an end-to-end fashion.

In the unheard sounds setting covering 102 sounds (Table 1, right), our method again strongly outperforms all existing methods on both datasets and in almost every metric. The only exception is our 2.8% lower SPL vs. Chen et al. (2020) on Replica, though our model still surpasses Chen et al. (2020) in terms of SNA on that dataset, meaning we have better accuracy when normalizing for total action count. Absolute performance declines for all methods, though, due to the unfamiliar audio spectrogram patterns. The acoustic memory is critical for this important setting; it successfully abstracts away the specific content of the training sounds to better generalize.



(a) Gan et al. (2020) (b) Chen et al. (2020) (c) AV-WaN

Figure 3: Navigation trajectories on top-down maps vs. all existing AudioGoal methods. Agent path fades from dark blue to light blue as time goes by. Green is the shortest geodesic path in continuous space. All agents have reached the goal. Our waypoint model navigates to the goal more efficiently. The agent's inputs are egocentric views (Fig. 1); figures show the top-down view for ease of viewing the full trajectories.

REFERENCES

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

Somil Bansal, Varun Tolani, Saurabh Gupta, Jitendra Malik, and Claire Tomlin. Combining optimal control and learning for visual navigation in novel environments. In *Conference on Robot Learning (CoRL)*, 2019.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *ICLR*, 2020.

Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. SoundSpaces: Audio-visual navigation in 3D environments. In *ECCV*, 2020.

Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020.

Frédéric Gougoux, Robert J Zatorre, Maryse Lassonde, Patrice Voss, and Franco Lepore. A functional neuroimaging study of sound localization: visual cortex activity predicts performance in early-blind individuals. *PLoS biology*, 2005.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017a.

Saurabh Gupta, David Fouhey, Sergey Levine, and Jitendra Malik. Unifying map and landmark based representations for visual navigation. *arXiv preprint arXiv:1712.08125*, 2017b.

Nadia Lessard, Michael Paré, Franco Lepore, and Maryse Lassonde. Early-blind human subjects localize sound sources better than sighted subjects. *Nature*, 1998.

Dmytro Mishkin, Alexey Dosovitskiy, and Vladlen Koltun. Benchmarking classic and learned navigation in complex 3d environments. *arXiv preprint arXiv:1901.10915*, 2019.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019.

Alexander Sax, Bradley Emi, Amir R Zamir, Leonidas Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. *arXiv preprint arXiv:1812.11971*, 2018.

Gregory J. Stein, Christopher Bradley, and Nicholas Roy. Learning over subgoals for efficient navigation of structured, unknown environments. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto (eds.), *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pp. 213–222. PMLR, 29–31 Oct 2018.