Learning Semantic-Agnostic and Spatial-Aware Representation for Generalizable Visual-Audio Navigation

Hongcheng Wang, Yuxuan Wang, Fangwei Zhong, *Member, IEEE*, Mingdong Wu, Jianwei Zhang, *Member, IEEE*, Yizhou Wang, *Member, IEEE*, and Hao Dong, *Member, IEEE*

***Abstract*—Visual-audio navigation (VAN) is attracting more and more attention from the robotic community due to its broad applications, e.g., household robots and rescue robots. In this task, an embodied agent must search for and navigate to the sound source with egocentric visual and audio observations. However, the existing methods are limited in two aspects: 1) poor generalization to unheard sound categories; 2) sample inefficient in training. Focusing on these two problems, we propose a brain-inspired plug-and-play method to learn a semantic-agnostic and spatial-aware representation for generalizable visual-audio navigation. We meticulously design two auxiliary tasks for respectively accelerating learning representations with the above-desired characteristics. With these two auxiliary tasks, the agent learns a spatially-correlated representation of visual and audio inputs that can be applied to work on environments with novel sounds and maps. Experiment results on realistic 3D scenes (Replica and Matterport3D) demonstrate that our method achieves better generalization performance when zero-shot transferred to scenes with unseen maps and unheard sound categories.**

***Index Terms*—Vision-based navigation, representation learning, reinforcement learning.**

Manuscript received 27 December 2022; accepted 18 April 2023. Date of publication 2 May 2023; date of current version 17 May 2023. This letter was recommended for publication by Associate Editor G. Costante and Editor E. Marchand upon evaluation of the reviewers' comments. This work was supported in part by MOST under Grant 2022ZD0114900, in part by the NSFC under Grants 62006006, 62136001, and 62061136001, and in part by Qualcomm University Research Grant. *(Hongcheng Wang and Yuxuan Wang contributed equally to this work.) (Corresponding author: Hao Dong.)*

Hongcheng Wang, Mingdong Wu, and Hao Dong are with the School of Computer Science, Peking University, Beijing 100871, China (e-mail: whc.1999@pku.edu.cn; wmingd@pku.edu.cn; hao.dong@pku.edu.cn).

Yuxuan Wang is with the Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China (e-mail: yuxwang@pku.edu.cn).

Fangwei Zhong is with the National Key Laboratory of General Artificial Intelligence, School of Intelligence and Technology, BIGAI, Peking University, Beijing 100871, China (e-mail: zfw@pku.edu.cn).

Jianwei Zhang is with the TAMS, Department of Informatics, Universität Hamburg, 22527 Hamburg, Germany (e-mail: zhang@informatik.uni-hamburg.de).

Yizhou Wang is with the Center on Frontiers of Computing Studies, Institute for Artificial Intelligence, Peking University, Beijing 100871, China, and also with the Nat'l Eng. Research Center of Visual Technology, Beijing 100871, China (e-mail: yizhou.wang@pku.edu.cn).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2023.3272518, provided by the authors.

Digital Object Identifier 10.1109/LRA.2023.3272518

I. Introduction

EMBODIED agents should be able to navigate to different locations to complete downstream tasks such as goal-specific tidying and delivering items. Most robot navigation is currently limited to pure visual input from scenes [1], [2], [3], [4], [5]. From a bionic perspective [6], [7], [8], we humans can integrate audio information with visual observations to improve the ability to perceive objects and scenes, such as locating the position of an invisible object [9]. Consequently, it is advisable for an intelligent agent to learn how to perceive and leverage multi-modal information, including vision and audio, to achieve better navigation performance.

With the recent development of the Soundspaces [10] simulation environment, researchers have begun to study leveraging both audio and visual information for navigation [10], [11], [12]. In visual-audio navigation (VAN) task, testing sets include heard and unheard sound categories: 1) *heard sound categories* mean the same sound categories group with the training set, 2) and *unheard sound categories* are never heard sound categories by the agent during the training procedure. For training sets, in some specific and typical scenes, we can provide almost any kind of sound that might be present in these scenes. For example, in a restaurant, a service robot may only need to learn to listen to service bells and customer greetings. However, in some atypical and complex scenes, we cannot provide all possible sound categories to learn because of the wide range of sounds that the agent will confront, such as a guard robot that should be able to react to odd sounds, activate the guard procedure and find where the odd sounds occur. Therefore, intelligent agents need to handle unheard sound categories. Even though the state-of-the-art (SOTA) methods attain $\sim 90\%$ success rate [10], [11] in Replica environments [13] with heard sound categories, their success rates drop to $\sim 50\%$ when navigating to unheard sound. Besides, existing methods use pure reinforcement learning loss (e.g. critic loss and actor loss) to train an agent in a simulator and thus need about 3 M$\sim$13 M steps to converge due to low sample efficiency, which takes several days. It is important to develop an algorithm with high sample efficiency for this task.

Humans are sensitive to sounds, and even infants who know nothing about sound categories can perceive the general orientation of sound [14]. Motivated by the previous observation, in this

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

paper, we refer to the human auditory processing mechanism. A dual-pathway model of auditory processing exists in the human brain where sound semantic information (*what* path) and sound spatial information (*where* path) are segregated into different brain areas [15], [16], [17], [18]. Semantic information contains sound category and other category-related information, such as the percussive feeling of metal [19]. Spatial information includes the distance and direction of sounds and other location-related information, such as the phase difference between two ears [16], [20]. Semantic information changes with the sound category, leading to difficulties in learning generalizable semantic representations of unheard sound categories. In contrast, spatial information does not change [21], [22], [23], enabling the potential for generalizing to unheard sound categories. As a result, we opt to maintain different attention levels to different information in the features, *i.e.*, to neglect semantic information and enhance spatial information.

Concretely, based on the human auditory mechanism, we propose a plug-and-play method encouraging agents to learn task-relevant representations from multi-modal inputs. To improve sample efficiency and generalization in the VAN task, we design two auxiliary tasks that provide additional training signals. These two tasks enable the agent to discover the intrinsic spatial correlations between visual and audio inputs. That can make it possible to apply the learned representation to environments with unseen sounds and maps. In one auxiliary task, we use a gradient reversal layer to create an adversarial relationship between an audio encoder and an audio classifier to ignore semantic information. In the other auxiliary task, we use temporal information from visual and auditory inputs to predict the relative direction of a sound, thereby enhancing spatial information. Because our method is plug-and-play, it can be applied to various VAN backbone algorithms using the same settings. In our experiments, we use two SOTA algorithms, AV-Nav [10] and AV-Wan [11] as the backbones. We demonstrate the superiority of our proposed method on two realistic 3D scene datasets, Replica [13] and Matterport3D [24], with strong generalization to scenarios with unheard sound categories and fewer training steps. In summary, our contributions are listed as follows:

1) We observe that paying different attention to semantic and spatial components in sounds can improve the sample efficiency and the generalization of visual-audio navigators on unheard sound categories.
2) We meticulously design two auxiliary tasks. One task uses an adversarial mechanism to neglect semantic information, and the other task predicts a relative direction to enhance spatial information.
3) The experiments on two sets of realistic 3D scenes, Replica and Matterport3D, show that our method can achieve better generalization performance in fewer training steps.

II. Related Work

Visual-Audio Navigation: In this task, an agent should navigate to the sound source by utilizing egocentric visual and audio observations. The task is challenging because of the complexity of the room structure itself and its effect on sound propagation, which leads to the fact that the agent cannot precisely estimate the loudness and direction of the sound to make decisions. Several existing studies [10], [11], [12], [25], [26] demonstrate the importance of fusing visual and audio modalities in navigation tasks and show good performance in scenes with heard sound categories. Some works [10], [11], [12] do not explicitly focus on sound semantics and perform better on heard sound categories than unheard sound categories. Semantic-aware methods [25], [26] explicitly exploit the sound semantic information and learn the association between semantic information and scene representations to reason about the sound source location, e.g., hearing water dripping means the agent may need to go to the kitchen or bathroom. However, these semantic-aware methods [25], [26] can only deal with heard sound **categories**, including heard sound **instances** and unheard sound **instances**, while our method focuses on the generalization towards unheard sound **categories**. We argue that neglecting semantic information enhances the navigation generalization on unheard sound categories and does little harm or even improves the performance of heard sound categories.

Auxiliary Task: It is not a new concept to train a reinforcement learning (RL) agent with auxiliary tasks. Auxiliary tasks are commonly used to improve the sample efficiency and attempt to build up state representations by predicting supplemental variables about important aspects of RL tasks, such as terminal state prediction [27], agent modeling [28], [29], [30], return prediction [31], [32], and depth prediction [33]. Designing auxiliary tasks for a specific goal can be challenging, especially when the input contains multiple modalities. It is important to ensure consistency between the auxiliary tasks and the main task; otherwise, the auxiliary tasks will only train the agent to accomplish the auxiliary goals or hinder performance on the main task. Our method introduces two auxiliary tasks for visual-audio navigation by referring to the human auditory mechanism. One is to predict the relative direction between the agent and the sound source location. Furthermore, the other is to force the agent to omit semantic information in sounds by adversarial learning.

III. Method

We follow the basic settings in AV-Nav [34] and AV-Wan [11] for the AudioGoal Navigation task. The task initializes an agent in the environment (a scene with single or multiple rooms) without the map of the environment. In each episode, a sound source is set in the environment, continuously emitting sounds that the agent can receive. The agent is required to navigate to the sound source using visual and audio information. All initial settings for the episodes are pre-generated, including the agent's initial position, the location of the target sound, the category of sound, and the room used for navigation, in order to avoid overly simplistic episodes.

In order to improve sample efficiency and make the navigation policy generalizable to unheard sound categories, we focus on extracting the generalizable components of the sounds referring to the human auditory mechanism. The contents of sounds contain two main components: semantic information and spatial

information. When the sound source location and robot position remain constant, the semantic information changes with the sound category, but the spatial information remains the same. Our method is therefore composed of two main tasks for learning the generalizable representation: 1) Semantic-Agnostic Learning (denoted in green in Fig. 2) learning semantic-agnostic representation by an adversarial mechanism between audio encoder and audio classifier, and 2) Spatial-Aware Learning (denoted in red in Fig. 2) learning spatial-aware representation by predicting the angle of sound relative to the agent by using a temporal representation containing visual and auditory information.

Since the initial settings for each episode are pre-generated rather than randomly selected at the beginning of the episode [10], without the Semantic-Agnostic Learning, the navigation policy will implicitly memorize the sounds used in each training episode (*i.e.* over-fitting on the training episodes), so its generalization will be weakened. Without Spatial-Aware Learning, Semantic-Agnostic Learning may mistakenly neglect the spatial information, making it also ignored by the agent (the most extreme case is that the audio encoder will output the same features for any audio input).

The additional processing of representations by these two tasks allows the agent to learn task-relevant features much faster, thus improving sample efficiency.

A. Semantic-Agnostic Learning

When receiving a sound, a human may not know what the sound category exactly is but can estimate the sound source location [35], [36], [37], and even an infant who knows nothing about the world can roughly localize the sound source [38], [39], which shows that spatial information alone is sufficient for humans to locate sounds. Inspired by the research above, we argue that in AudioGoal navigation tasks for intelligent agents, spatial information of the sound is enough for locating and perceiving the sound. While semantic information changes with the sound categories, it increases the difficulty for agents to learn generalizable semantic representations. Moreover, for some atypical scenes (e.g., guard robots facing odd sounds), sounds and scenes are not closely related. Therefore, learning semantic-agnostic representations should not harm the navigation performance on both heard sound categories but could enhance the generalization of unheard sound categories.

Concretely, learning semantic-agnostic representations means that, with an agent fixed in a certain location and sound source in another certain location, the method outputs the same representation when taking sounds with different semantics. To equip the representations learned by the method with the semantic-agnostic property, we design an auxiliary task in which an audio encoder needs to weaken the ability of the audio classifier to distinguish the current sound semantic category while the audio classifier attempts to distinguish the sound semantic category corresponding to an audio feature. The adversarial training forces the audio encoder to learn semantic-irrelevant representations.

Therefore, we use an adversarial mechanism between an audio encoder parameterized by θ_A and a 4-layer fully connected network audio classifier (AC) parameterized by θ_C. To implement this adversarial mechanism, we employ a gradient reversal layer [40] between the audio classifier and the audio encoder by multiplying a factor $-\lambda$ on gradient flow reflecting the adversarial intensity:

$$\lambda = \frac{2b}{1.0 + e^{-10 * \frac{n}{N}}} - b \tag{1}$$

where n denotes the number of currently completed episodes, N denotes the number of total episodes and b denotes the bound of the adversarial intensity. And the parameters are optimized as follows:

$$\theta_C \longleftarrow \theta_C - \mu \frac{\partial \mathcal{L}_C}{\partial \theta_C} \tag{2}$$

$$\theta_A \longleftarrow \theta_A - \mu \left(\frac{\partial \mathcal{L}_O}{\partial \theta_A} - \lambda \frac{\partial \mathcal{L}_C}{\partial \theta_A} \right) \tag{3}$$

where μ denotes the learning rate, $\mathcal{L}_C$ denotes Cross Entropy Loss, and $\mathcal{L}_O$ denotes other loss related to θ_A such as Actor and Critic Loss in reinforcement learning.

B. Spatial-Aware Learning

Semantic-agnostic learning ignores navigation-irrelevant information but does not encourage the agent to learn navigation-relevant representations. Although reinforcement learning provides reward signals to help the agent extract navigation-relevant features, during the initial exploration phase, the agent may not catch sight of reward signals but can rapidly learn neglecting the semantic information of the sound from the adversarial audio classifier to minimize the adversarial optimization objective. This rapid learning could lead to the audio encoder incorrectly ignoring spatial information as well, resulting in its output being insensitive to changes in the agent's position. On this occasion, the agent cannot navigate to the sound source. Predicting sound location as an auxiliary task can effectively provide an additional training signal to help the agent extract spatial information and assist in navigation policy learning.

We use a 4-layer fully connected network as the location predictor (LP) with temporal features generated by a Time-series Model as input to predict the pitch and yaw angles of the sound source relative to the agent, denoted as β and α in Fig. 1, respectively. In practice, we do not predict the angle directly but predict the *sine* and *cosine* of the angle. The *sine* and *cosine* predictions avoid the periodicity of the angle that leads to the non-uniqueness. We use the Mean-Squared Loss as the auxiliary loss function. The gradients generated by the loss of the LP are utilized to update the Audio Encoder, the Visual Encoder, and the Time-series Model. These models can thus learn to extract features containing spatial information for RL's actor and critic to learn navigation policy better.

C. Training Details

We use SoundSpaces [10] as our simulator, enabling realistic audio rendering. The SoundSpaces simulator discretizes scenes into uniformly distributed navigability graphs so that the agent can only move one node to a naviagble neighboring node in the

TABLE I
TESTING RESULTS ON HEARD AND UNHEARD SOUND CATEGORIES

Method	Replica						MP3D					
	Heard Sound Category			Unheard Sound Category			Heard Sound Category			Unheard Sound Category		
	SR(↑)	SPL(↑)	SNA(↑)	SR(↑)	SPL(↑)	SNA(↑)	SR(↑)	SPL(↑)	SNA(↑)	SR(↑)	SPL(↑)	SNA(↑)
Random	0.185	0.049	0.018	0.185	0.049	0.018	0.091	0.021	0.008	0.091	0.021	0.008
DF	0.720	0.547	0.411	0.111	0.172	0.084	0.412	0.232	0.238	0.180	0.139	0.107
AV-Nav	**0.903**	0.672	0.354	0.576	0.394	0.180	**0.700**	**0.227**	**0.252**	0.359	0.105	0.109
Ours+AV-Nav	0.898	**0.689**	**0.400**	**0.823**	**0.606**	**0.330**	0.652	**0.227**	0.217	**0.602**	**0.198**	**0.171**
AV-Wan	**0.918**	0.676	0.522	0.424	0.289	0.218	0.796	0.453	0.337	0.567	0.409	0.306
Ours+AV-Wan	0.904	**0.709**	**0.552**	**0.628**	**0.434**	**0.330**	**0.829**	**0.614**	**0.468**	**0.607**	**0.423**	**0.314**

We apply our method to AV-Nav and AV-Wan and obtain higher quantitative Results. The SPL and SNA show that our method improves the efficiency of the previous works, allowing the agent to choose a shorter Path (Higher SPL) and a faster path (Higher SNA) to reach the sound location.
Bold value indicate the best results achieved by different methods under each metric for every experiment, thereby clearly demonstrating that our approach has reached the state-of-the-art (SOTA) in the vast majority of cases.

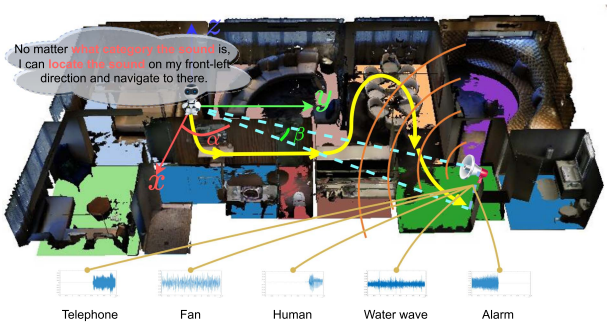

Fig. 1. Problem Setting. The robot should navigate to the sound source location with the visual-audio observation, no matter what category of sound is being played. In this example, the agent is in the bedroom initially and locates the sound in its front-left direction. α and β are the yaw and pitch angles of the sound source relative to the agent.

TABLE II
ABLATION STUDY FOR OUR METHOD ON AV-NAV

		Ablation	SR(↑)	SPL(↑)	SNA(↑)
Replica	AV-Nav	Ours w/o AC and LP	0.576	0.394	0.180
		Ours w/o AC	0.749	0.584	0.292
		Ours w/o LP	0.766	0.588	0.266
		Ours	**0.823**	**0.606**	**0.330**
	AV-Wan	Ours w/o AC and LP	0.424	0.289	0.218
		Ours w/o AC	0.432	0.319	0.292
		Ours w/o LP	0.606	0.349	0.256
		Ours	**0.628**	**0.434**	**0.330**
MP3D	AV-Nav	Ours w/o AC and LP	0.359	0.105	0.109
		Ours w/o AC	0.561	0.169	0.154
		Ours w/o LP	0.478	0.143	0.150
		Ours	**0.602**	**0.198**	**0.171**
	AV-Wan	Ours w/o AC and LP	0.567	0.409	0.306
		Ours w/o AC	0.556	0.399	0.305
		Ours w/o LP	0.599	0.391	0.295
		Ours	**0.607**	**0.423**	**0.314**

We apply our method to AV-Nav and perform an ablation study on two components of our method, audio classifier (AC) and location predictor (LP), on testing sets of replica and matterport3D datasets. Three metrics are compared, including SR, SPL, and SNA.
Bold value indicate the best results achieved by different methods under each metric for every experiment, thereby clearly demonstrating that our approach has reached the state-of-the-art (SOTA) in the vast majority of cases.

graphs. Where there are obstacles there are no nodes. Thus the action space $\mathcal{A}$ has only four actions: MoveForward, TurnLeft, TurnRight and Stop. The Soundspace removes episodes where the distance from the start position to the target position is less than 4 m and episodes where the shortest path is almost a straight line (ratio of geodesic to Euclidean distance less than 1.1).

Since we apply our method on AV-Nav [10] and AV-Wan [11], we follow the design of their reward function, in which the agent is given a $+10$ reward if the agent executes action *Stop* at the sound source location, $+1$ reward on AV-Nav or $+0.25$ reward on AV-Wan if the agent reduces the geodesic distance to the sound source location and an equivalent penalty if the agent increases the geodesic distance and -0.01 for time penalty.

We train all learnable models jointly with Proximal Policy Optimization (PPO) [41]. Each episode contains 150 steps, and the success criterion is met if the agent executes the action Stop at the sound position in 150 steps.

IV. EXPERIMENTAL RESULTS

A. Experiment Settings

Environments and Datasets: We use the same audio and visual dataset and train/val/test splits as AV-Nav [10] and AV-Wan [11] to demonstrate the improvement of our method. We use the same simulator, SoundSpaces [10], with two real-world 3D scene datasets, Replica and Matterport3D (MP3D), for training and testing our method along with train/val/test splits of 73/11/18 sound categories. Replica is a relatively small scene dataset with an average area of 47.24 m^2 and train/val/test splits of 9/4/5 scenes. Matterport3D has relatively large scenes with an average area of 517.34 m^2 and train/val/test splits of 57/10/12 scenes. We also follow basic configuration and hyper-parameters from AV-Nav and AV-Wan and only use depth maps as visual information.

Metrics: We evaluate our method on the following metrics:

1) Success Rate (SR): the fraction of successful episodes.
2) Success Weighted by Path Length (SPL) [42]: we weigh the success by the ratio of the execution path length to the shortest path length.
3) Success Weighted by Number of Actions (SNA) [11]: we weigh the success by the ratio of the executive action numbers to the minor action numbers.

We use the model with the highest SPL on the validation set for testing and reporting the table results.

Baselines: We compare our methods with the following baselines:

1) *Random:* An agent randomly selects an action in action space $\mathcal{A}$. The episode ends when executing Stop.

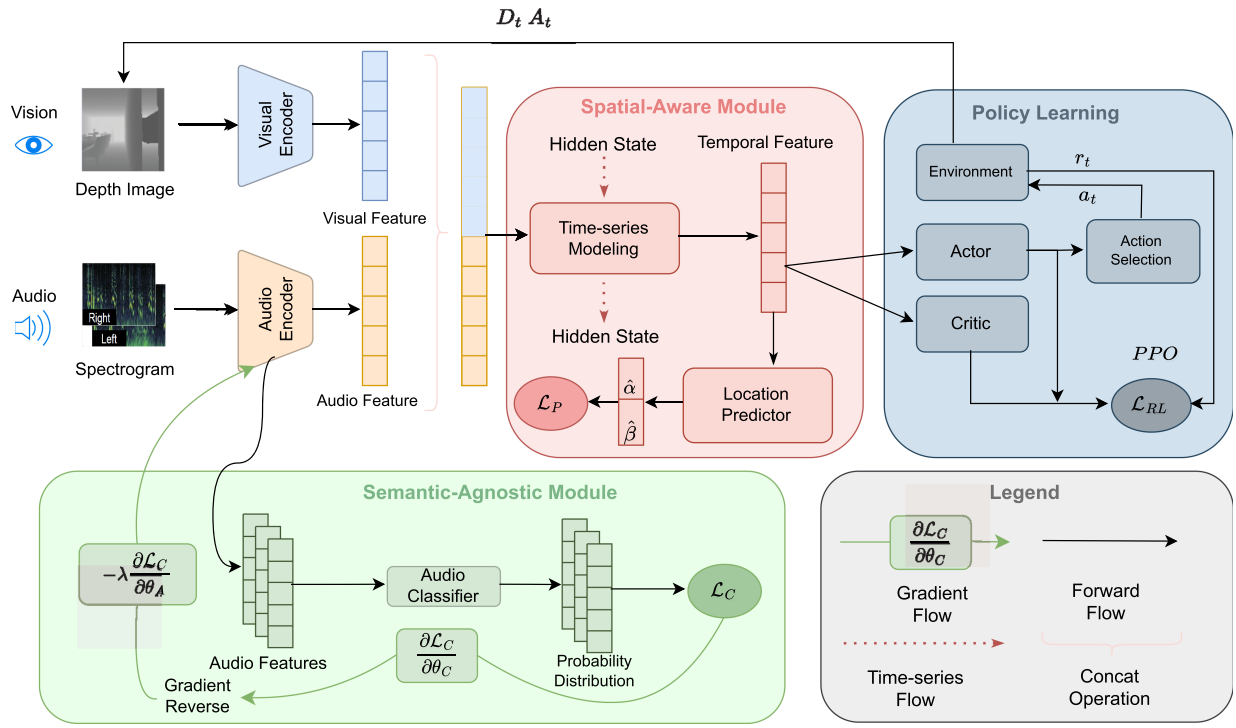

Fig. 2. Training Pipeline. At each time step t, our method uses depth images (D_t) and spectrograms (A_t) as inputs for navigation. During the training procedure, an Audio Classifier (AC, parameterized by θ_C) enforces the model to neglect semantic information via adversarial training supervised by $\mathcal{L}_C$. Concurrently, the temporal features (O_t) are given to a Location Predictor (LP) to pull out the sound source direction (α, β) supervised by $\mathcal{L}_P$. α and β are the yaw and pitch angles of the sound source relative to the agent. Action Selection samples from the probability distribution generated by Actor to obtain action a_t. After executing a_t in the environment, the environment returns a reward signal r_t. At the end of each RL epoch, we train the Audio Encoder (parameterized by θ_A), the Audio Classifier and the Location Predictor simultaneously.

2) *Direction Follower(DF) [11]:* This method pretrains a model to predict the direction of arrival (DoA). An agent sets an intermediate goal K meters away in the predicted direction and plans to navigate there. We set $K = 2$ in Replica and $K = 4$ in Matterport3D.
3) *AV-Nav [10]:* It is a state-of-the-art VAN method that makes decisions using visual-audio fusion features with temporal sequences.
4) *AV-Wan [11]:* It is a state-of-the-art VAN method that builds geometric and acoustic maps and uses them to predict an intermediate goal adaptively. AV-Wan uses the Dijkstra [43] shortest path algorithm to compute the path from the current node to the intermediate goal.

B. Quantitative Comparison

We apply our method on AV-Nav [10] and AV-Wan [11] and test baselines and our method referred by Ours+AV-Nav and Ours+AV-Wan on unheard sound categories in Table I.

Random performs poorly on both datasets, showing that the difficulty of the task and the robot is supposed to make good use of visual and audio cues. Direction Follower uses only audio information for decision making, while visual information is only used for path planning, so Direction Follower performs worse than the method that fuses information from both modalities to make decisions.

After applying our method, AV-Nav and AV-Wan achieve significant improvements on Replica and Matterport3D datasets on unheard sound categories, proving that our method works well for different backbone algorithms and datasets. In particular, on Replica, our method gains about 50% SPL improvement on the previous works. The results on AV-Nav and AV-Wan demonstrate the advantages of our method where we optimize the features and represent them in a more task-specific manner. We also test our method on heard sound categories, shown in Table I. The results show that our method improves performance slightly, showing that our method does not trade performance on the heard sound categories for generalizability by impairing it.

Considering that there exist domain gaps between the real world and the simulator, such as audio and depth noise, we add these two parts of noise to the environment to simulate the real world and demonstrate the robustness of our method following the setting of audio noise and depth noise from AV-Wan [11]. We conducted experiments on noise levels ranging from 20 to 50, with intervals of 10. Notice that, while AV-Wan [11] only use *telephone* in the noise experiments as the target sound, our work focuses on the generalization ability towards unheard sound categories, so we use all the sound categories in the testing set as target sounds instead. The results are shown in Table III. Note that even with different noise levels, our method still improves the performance of the previous works. With different levels of noise, the performance of our method shows no significant degradation and exhibits strong robustness. The robustness to noise can indicate that our method has the potential to be used in the real world.

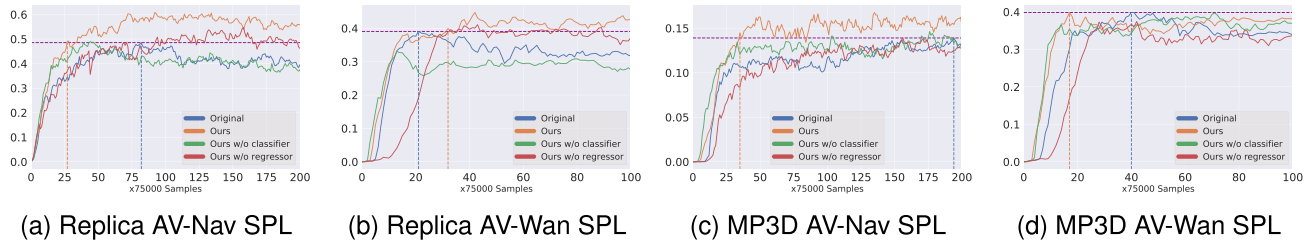

Fig. 3. Learning Curve on testing sets. We plot the testing results of the previous works and ours during training in both Replica and MatterPort3D environments with AV-Nav and AV-Wan as backbones, respectively. We plot a horizon dashed purple line across the highest SPL value of the previous works as a benchmark. We also draw vertical dashed lines for the previous works and ours in their corresponding colors, to indicate where their SPL values are greater than or equal to the benchmark for the first time. Our method can outperform the previous works with fewer training samples.

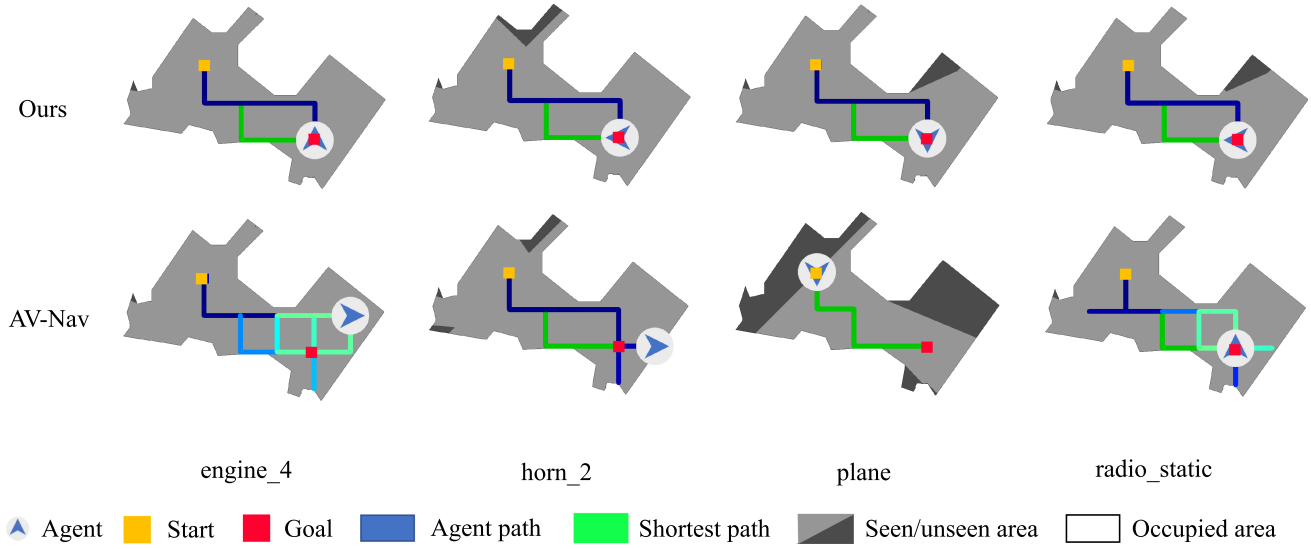

Fig. 4. Trajectory Visualization for different sound categories. We visualized agent trajectories using our method and AV-Nav, respectively with the same set of start and end position episodes in the same scene. In each episode, the agent needs to navigate from the yellow point to the red point. The name at the bottom represents the category of sound, which means that each column has a different sound. Agent path fades from dark blue to light blue as time goes by. Green is the shortest geodesic path in continuous space. We aim to show that our method yields the same trajectory for different sound categories, which shows that the features we learn are indeed semantic-agnostic. The first row shows our results, and the second row is the results from AV-Nav. AV-Nav may fail in some episodes, e.g., the first three columns, and run quite differently when navigating to different sounds, while our method navigates to the goal in all four episodes and keep trajectory consistent in these episodes.

C. Sample Efficiency and Learning Curve

To demonstrate our method's high sample efficiency, we show the learning curves on the testing set on the Replica and MatterPort3D with both AV-Nav and AV-Wan as backbones. Fig. 3 shows that our method can achieve higher performance than the final results of the previous works, with fewer samples than the previous works needs to converge. We compare the number of samples required by ours and the previous works, using the highest point of the previous works as a benchmark. In Fig. 3(a), (c), and (d), our methods require fewer samples, and the performance still grows as the samples grow. In Fig. 3(b), although there is no significant sample difference between ours and the previous works, our method is more stable in the later stages and the performance continues to grow.

D. Trajectory Visualizations

We visualize the trajectories using our method and AV-Nav under four categories of sounds, shown in Fig. 4. We refer to the same start agent position and the same sound source location within the same scene as the *same task*. To view the trajectory generation process, please watch the attached video. In the first line of Fig. 4, our method can come out of the trajectory equivalent to the shortest path in various sounds consistent with each other. In the second line, however, AV-Nav either fails to complete the task or the trajectory is very complex and inconsistent.

We also visualize the trajectories in different scenes, shown in Fig. 5. Our method can generate more efficient trajectories within different scenes than AV-Nav.

E. Ablation Studies and Analysis

Table II shows the ablation results of the audio classifier and the location predictor components of our method. Removing either the audio classifier or the location predictor leads to a reduction in performance. Notably, reducing the location predictor hurts the performance more than reducing the audio

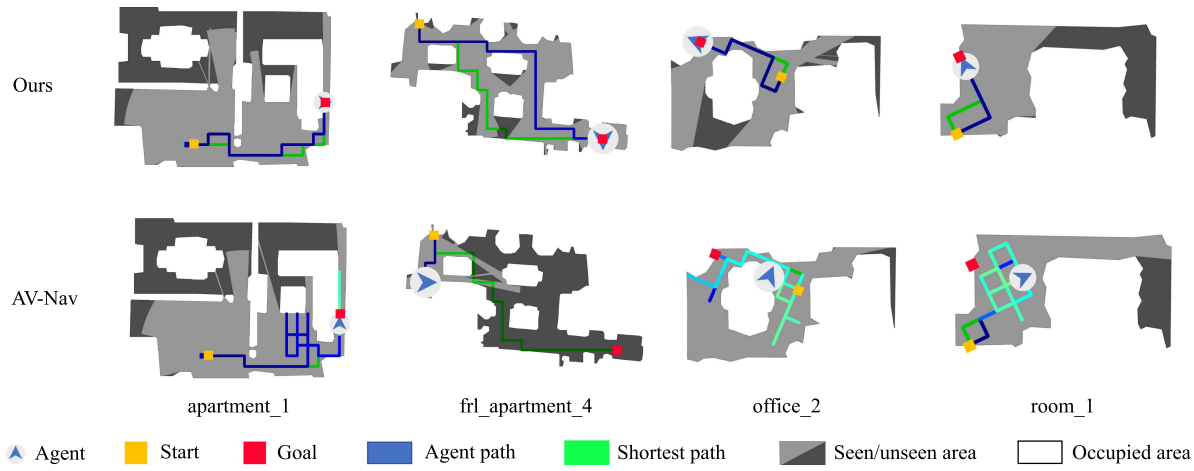

Fig. 5. Trajectory Visualization for Different Scenes. We visualized navigation trajectories using our method and AV-Nav in various scenes. The name at the bottom represents the scene. In each episode, the agent needs to navigate from the yellow point to the red point. Agent path fades from dark blue to light blue as time goes by. Green is the shortest geodesic path in continuous space. The first row shows our results, and the second row is the results from AV-Nav. AV-Nav may fail in some episodes, e.g., the second and third column, or take a complex route, e.g., the first, third, and fourth column. Our method finds a good path to the end point in all four episodes.

TABLE III
AUDIO NOISE EXPERIMENTS

	Replica										Mp3d									
Audio Noise Level (SNR)	50		40		30		20		Avg. SPL		50		40		30		20		Avg. SPL	
Depth Noise	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o	w	w/o
AV-Nav	0.327	0.363	0.360	0.362	0.266	0.300	0.243	0.305	0.299	0.333	0.102	0.103	0.101	0.111	0.104	0.107	0.118	0.124	0.106	0.111
Ours+AV-Nav	**0.398**	**0.401**	**0.432**	**0.461**	**0.452**	**0.472**	**0.405**	**0.442**	**0.422**	**0.444**	**0.114**	**0.143**	**0.131**	**0.133**	**0.149**	**0.146**	**0.139**	**0.140**	**0.133**	**0.141**
AV-Wan	0.291	0.275	0.286	0.275	**0.337**	**0.352**	0.314	0.307	0.307	0.302	0.279	0.279	0.314	0.314	0.331	0.331	**0.375**	**0.375**	0.325	0.325
Ours+AV-Wan	**0.367**	**0.368**	**0.354**	**0.353**	0.327	0.338	**0.347**	**0.350**	**0.349**	**0.352**	**0.297**	**0.297**	**0.377**	**0.377**	**0.342**	**0.342**	0.361	0.361	**0.344**	**0.344**

We show the SPL in the experiments with different levels of noise (following the noise settings from AV-Wan [11]). Our method still outperforms the previous works in most cases, and the performance does not show a large degradation compared to the noise-free experiments, showing our method's robustness to noise. We present the average SPL on differenet noise Levels in the additional columns. Bold value indicate the best results achieved by different methods under each metric for every experiment, thereby clearly demonstrating that our approach has reached the state-of-the-art (SOTA) in the vast majority of cases.

classifier does in Matterport3D. Compared to Replica, scenes in Matterport3D have bigger areas; thus, spatial information is more helpful in completing tasks in Matterport3D.

In addition, the audio classifier provides an adversarial training mechanism, which implicitly boosts the model's generalization by forcing the model to ignore the semantic information of the audio inputs. Meanwhile, the model can benefit from the auxiliary localization task's additional training signals and directly improve navigation performance.

V. CONCLUSION AND DISCUSSION

This work focuses on the generalization and sample efficiency problem for VAN tasks. The different properties of spatial and semantic information inspired us to reduce the generalization gap between unheard and heard sound categories and learn task-relevant representations fast. Therefore, we propose a plug-and-play method to narrow the performance gap on unheard and heard sound categories by neglecting semantic information while enhancing spatial information. Evaluations on Replica and Matterport3D show that our method significantly outperforms the baseline on the unheard sound categories and slightly improves the heard sound categories. Learning curves show that our method has better sample efficiency than baselines. We also conducted audio and depth noise experiments to demonstrate the robustness of our method to depth image noise and varying levels of audio noise. The results show that our method performs well even with noisy inputs.

In the future, we will further explore the methods to enhance the generalization in more challenging visual-audio navigation settings, e.g., real-world development and complex environments. 1) Real-world development (sim2real transfer) involves the challenging task of transferring reinforcement learning models trained in simulated environments to real robots. Due to the significant sim2real gap in both audio and visual modalities, conducting experiments in the real world remains difficult. To overcome this challenge, we must address the discrepancy between simulation and reality and improve the model's generalization ability. One potential solution is to apply bi-directional domain adaptation to align the feature distributions of simulation and reality during training. Additionally, exploring meta-reinforcement learning algorithms may enable the agent to efficiently mitigate domain drift during test time. 2) In complex environments, the agent must handle interference from multiple sound sources and uncertainty from moving sound. To tackle scenarios with multiple sound sources at similar volume levels, we can leverage semantic information and sound source separation algorithms [44], [45] to filter out the target sound source as input to the navigator. Moreover, we can augment the training process with a multi-agent game [46] to automatically generate diverse and challenging distracting or moving sources, further enhancing the robustness of the system.

REFERENCES

[1] Y. Zhu et al., "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 3357–3364.

[2] D. Mishkin, A. Dosovitskiy, and V. Koltun, "Benchmarking classic and learned navigation in complex 3D environments," 2019, *arXiv:1901.10915*.

[3] E. Wijmans et al., "Decentralized distributed PPO: Solving pointgoal navigation," 2019, *arXiv:1911.00357*.

[4] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 4247–4258.

[5] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "Clip on wheels: Zero-shot object navigation as object localization and exploration," 2022, *arXiv:2203.10421*.

[6] B. Liu, Z. Wang, and Z. Jin, "The integration processing of the visual and auditory information in videos of real-world events: An ERP study," *Neurosci. Lett.*, vol. 461, no. 1, pp. 7–11, 2009.

[7] X. Guo, X. Li, X. Ge, and S. Tong, "Audiovisual congruency and incongruency effects on auditory intensity discrimination," *Neurosci. Lett.*, vol. 584, pp. 241–246, 2015.

[8] M. Gori, G. Sandini, and D. Burr, "Development of visuo-auditory integration in space and time," *Front. Integrative Neurosci.*, vol. 6, 2012, Art. no. 77.

[9] N. R. Carlson, *Foundations of Behavioral Neuroscience*. London, U.K.: Pearson, 2013.

[10] C. Chen et al., "Soundspaces: Audio-visual navigation in 3D environments," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 17–36.

[11] C. Chen, S. Majumder, Z. Al-Halah, R. Gao, S. K. Ramakrishnan, and K. Grauman, "Learning to set waypoints for audio-visual navigation," in *Proc. Int. Conf. Learn. Representations*, 2020.

[12] C. Gan, Y. Zhang, J. Wu, B. Gong, and J. B. Tenenbaum, "Look, listen, and act: Towards audio-visual embodied navigation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 9701–9707.

[13] J. Straub et al., "The replica dataset: A digital replica of indoor spaces," 2019, *arXiv:1906.05797*.

[14] S. N. Graven and J. V. Browne, "Auditory development in the fetus and infant," *Newborn Infant Nurs. Rev.*, vol. 8, no. 4, pp. 187–193, 2008.

[15] C. Alain, Y. He, and C. Grady, "The contribution of the inferior parietal lobe to auditory spatial working memory," *J. Cogn. Neurosci.*, vol. 20, no. 2, pp. 285–295, 2008.

[16] S. R. Arnott, M. A. Binns, C. L. Grady, and C. Alain, "Assessing the auditory dual-pathway model in humans," *Neuroimage*, vol. 22, no. 1, pp. 401–408, 2004.

[17] W. A. Yost, "Auditory image perception and analysis: The basis for hearing," *Hear. Res.*, vol. 56, no. 1/2, pp. 8–18, 1991.

[18] H. E. Heffner and R. S. Heffner, "Role of primate auditory cortex in hearing," *Comp. Percep.*, vol. 2, pp. 279–310, 1990.

[19] M. Adriani et al., "Sound recognition and localization in man: Specialized cortical networks and effects of acute circumscribed lesions," *Exp. Brain Res.*, vol. 153, no. 4, pp. 591–604, 2003.

[20] I. C. Zündorf, J. Lewald, and H.-O. Karnath, "Testing the dual-pathway model for auditory processing in human cortex," *Neuroimage*, vol. 124, pp. 672–681, 2016.

[21] A. Luo, Y. Du, M. Tarr, J. Tenenbaum, A. Torralba, and C. Gan, "Learning neural acoustic fields," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 3165–3177.

[22] C. Cao, Z. Ren, C. Schissler, D. Manocha, and K. Zhou, "Interactive sound propagation with bidirectional path tracing," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–11, 2016.

[23] E. Veach and L. Guibas, "Bidirectional estimators for light transport," in *Proc. Photorealistic Rendering Techn.*, 1995, pp. 145–167.

[24] A. Chang et al., "Matterport3D: Learning from RGB-D data in indoor environments," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 667–676.

[25] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15516–15525.

[26] G. Tatiya et al., "Knowledge-driven scene priors for semantic audio-visual embodied navigation," 2022, *arXiv:2212.11345*.

[27] B. Kartal, P. Hernandez-Leal, and M. E. Taylor, "Terminal prediction as an auxiliary task for deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell. Interactive Digit. Entertainment*, 2019, pp. 38–44.

[28] P. Hernandez-Leal, B. Kartal, and M. E. Taylor, "Agent modeling as auxiliary task for deep reinforcement learning," in *Proc. AAAI Conf. Artif. Intell. Interactive Digit. Entertainment*, 2019, pp. 31–37.

[29] J. Foerster et al., "Stabilising experience replay for deep multi-agent reinforcement learning," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1146–1155.

[30] Z.-W. Hong, S.-Y. Su, T.-Y. Shann, Y.-H. Chang, and C.-Y. Lee, "A deep policy inference Q-network for multi-agent systems," in *Proc. 17th Int. Conf. Auton. Agents MultiAgent Syst.*, 2018, pp. 1388–1396.

[31] G. Liu et al., "Return-based contrastive representation learning for reinforcement learning," in *Proc. Int. Conf. Learn. Representations*, 2020.

[32] M. Jaderberg et al., "Reinforcement learning with unsupervised auxiliary tasks," in *Proc. Int. Conf. Learn. Representations*, 2017.

[33] P. Mirowski, "Learning to navigate," in *Proc. 1st Int. Workshop Multimodal Understanding Learn. Embodied Appl.*, 2019, p. 25.

[34] F. Zeng, C. Wang, and S. S. Ge, "A survey on visual navigation for artificial agents with deep reinforcement learning," *IEEE Access*, vol. 8, pp. 135426–135442, 2020.

[35] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *J. Acoustical Soc. Amer.*, vol. 87, no. 5, pp. 2188–2200, 1990.

[36] J. C. Middlebrooks and D. M. Green, "Sound localization by human listeners," *Annu. Rev. Psychol*, vol. 42, pp. 135–59, 1991.

[37] L. Picinali, A. Afonso, M. Denis, and B. F. Katz, "Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge," *Int. J. Hum.- Comput. Stud.*, vol. 72, no. 4, pp. 393–407, 2014.

[38] B. A. Morrongiello and A. Gotowiec, "Recent advances in the behavioral study of infant audition: The development of sound localization skills," *J. Speech- Lang. Pathol. Audiol.*, vol. 14, no. 4, pp. 51–63, 1990.

[39] B. A. Morrongiello, K. D. Fenwick, L. Hillier, and G. Chance, "Sound localization in newborn human infants," *Devlop. Psychobiol.: J. Int. Soc. Devlop. Psychobiol.*, vol. 27, no. 8, pp. 519–538, 1994.

[40] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 34th Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[41] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.

[42] P. Anderson et al., "On evaluation of embodied navigation agents," 2018, *arXiv:1807.06757*.

[43] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, Dec. 1959.

[44] S. Majumder and K. Grauman, "Active audio-visual separation of dynamic sound sources," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 551–569.

[45] E. Tzinis, S. Venkataramani, Z. Wang, C. Subakan, and P. Smaragdis, "Two-step sound source separation: Training on learned latent targets," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 31–35.

[46] F. Zhong, P. Sun, W. Luo, T. Yan, and Y. Wang, "Towards distraction-robust active visual tracking," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12782–12792.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.