Omnidirectional Information Gathering for Knowledge Transfer-based Audio-Visual Navigation

Jinyu Chen[1], Wenguan Wang[2*], Si Liu[1*], Hongsheng Li[3], Yi Yang[2]
[1] Institute of Artificial Intelligence, Beihang University [2] ReLER, CCAI, Zhejiang University [3] The Chinese University of Hong Kong
https://github.com/chenjinyubuaa/ORAN

Abstract

Audio-visual navigation is an audio-targeted wayfinding task where a robot agent is entailed to travel a never-before-seen 3D environment towards the sounding source. In this article, we present ORAN*, an omnidirectional audio-visual navigator based on cross-task navigation skill transfer. In particular,* ORAN *sharpens its two basic abilities for a such challenging task, namely wayfinding and audio-visual information gathering. First,* ORAN *is trained with a confidence-aware cross-task policy distillation (CCPD) strategy. CCPD transfers the fundamental, point-to-point wayfinding skill that is well trained on the large-scale PointGoal task to* ORAN*, so as to help* ORAN *to better master audio-visual navigation with far fewer training samples. To improve the efficiency of knowledge transfer and address the domain gap, CCPD is made to be adaptive to the decision confidence of the teacher policy. Second,* ORAN *is equipped with an omnidirectional information gathering (OIG) mechanism,* i.e.*, gleaning visual-acoustic observations from different directions before decision-making. As a result,* ORAN *yields more robust navigation behaviour. Taking CCPD and OIG together,* ORAN *significantly outperforms previous competitors. After the model ensemble, we got 1st in Soundspaces Challenge 2022, improving SPL and SR by 53% and 35% relatively.*

1. Introduction

Developing intelligent autonomous wayfinding agents that can robustly navigate in unexplored environments to reach target locations is one of the most classic and fundamental tasks in robotics and is widely viewed as a critical building block of embodied AI. To simulate different real-world application scenarios of wayfinding agents, various navigation tasks are proposed, where the target goal is appointed by, for example, GPS coordinate [42, 27, 26], semantic tag [57, 8, 9, 40], visual language instruction [23, 36, 4, 25, 3, 38, 30, 56, 37, 2, 55, 47, 49, 48, 15], image photo [10, 1, 34, 29]. Among these navigation tasks, in this article, we are particularly interested in audio-visual navigation (also named as AudioGoal) [12, 19], in which the navigation agent is entailed to find sound-emitting objects in visually-and-acoustically rich 3D environments. Audio plays an irreplaceable role here – it reveals not only the

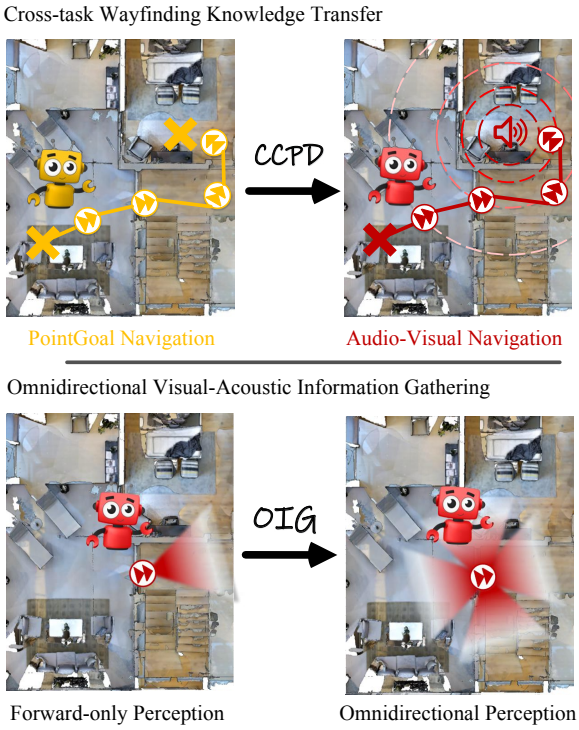

Figure 1: Top: Based on CCPD, our ORAN transfers wayfinding knowledge well-trained on PointGoal task to audio-visual navigation. Bottom: Based on OIG, ORAN collects visual and acoustic information from different directions for robust decision-making.

*Corresponding author: *Wenguan Wang, Si Liu.*

properties of the target object but also the layout of the unexplored areas. For example, when we hear a sound, we can know what makes the such sound and locate the sound source even before we observe it. Along this direction, existing AudioGoal solutions are typically reinforcement learning (RL) algorithms for direct audio-visual perception to low-level navigation action mapping [12, 11, 53]. Some methods further predict the next intermediate goal [13] on a top-down topology map to improve the agent's long-term stability. Methods like [53] instead consider the sound attacks to promote navigation robustness in complicated audio environments.

In this article, we develop ORAN, an omnidirectional audio-visual navigation agent based on cross-task wayfinding skill transfer. As shown in Figure 1, ORAN advances state-of-the-art technologies for AudioGoal in two aspects, namely wayfinding and visual-audio information gathering. Our technique innovations are born from two crucial insights. **First**, besides comprehending the received visual and audio signals, a successful AudioGoal agent needs to master some very basic wayfinding skills, such as precisely moving towards a short-term target, safely travelling without collision, and entering/leaving a room through the door. It is clear that these basic wayfinding skills are shared among different wayfinding tasks. Hence it is reasonable to assume that an AudioGoal agent can benefit from the knowledge of a high-performance navigator that is already well-trained on other navigation tasks, especially considering that the training samples of AudioGoal are relatively limited, while AudioGoal itself is a very challenging task. **Second**, it is natural for our human to turn around upon hearing a sound behind us, or turn our head to find out what we have heard [5]. In sharp contrast, existing AudioGoal agents only make use of visual-audio information perceived forward during navigation decision-making. It is clear that there exists a huge gap between such a simple, forward-only perception regime between the omnidirectional decision-making mode (particularly for those top-down map-based AudioGoal agents [13, 24]).

Our ORAN is elaborately designed to be fully aware of the aforementioned issues, so as to sharpen its wayfinding and visual-audio information-gathering abilities. **First**, ORAN is trained with a Confidence-aware Cross-task Policy Distillation (CCPD) strategy. CCPD allows ORAN to transfer the navigation knowledge learned on PointGoal to AudioGoal. Recently, PointGoal [41, 45] has seen significant advance – with millions of frames of experience and assistance of a GPS+Compass sensor, PointGoal agents can achieve nearly perfect navigation performance, given the coordinates of starting and target locations [50, 35]. During the training of AudioGoal, we consider the behaviours of a well-trained PointGoal agent, queried with the same starting and target waypoint locations, as informative demonstrations for ORAN. The reuse of the navigation knowledge relieves ORAN's burden of learning to both masters basic navigation skills and how to understand and plan with visual-audio perception. Moreover, to further better overcome the domain gap between the two tasks and improve the efficiency of such cross-task navigation knowledge transfer, we render larger weights to those more confident steps of the PointGoal policy during policy distillation. **Second**, ORAN is empowered with an omnidirectional information gathering (OIG) ability. OIG enables ORAN to make use of acoustic-visual information collected from different directions, instead of the only direction it is facing, to support its omnidirectional navigation decision-making.

We experimentally demonstrate that combining CCPD and OIG together makes our ORAN a powerful AudioGoal navigator, which sets state-of-the-art performance on Soundspaces Challenge dataset [12], *e.g.*, >10% absolute lifting in SPL on *unhead* sets. Moreover, after model assembling, ORAN yields further 12% absolute promotion in SPL. The fused model won 1st on the Soundspaces challenge 2022 [12] and improves SPL by 53% and SR by 35% relatively.

2. Related Work

Audio-visual Navigation. The audio-visual Navigation task, also referred to as AudioGoal navigation involves the exploration of unfamiliar surroundings, utilizing both visual and auditory cues in order to navigate towards a designated sound source. Various platforms are developed for simulating indoor audio-visual environments. For instance, VAR [19] employs the AI2-THOR [32] platform as a foundation to construct a simulated audio-visual navigation environment. Soundspaces [12], on the other hand, utilizes the Habitat [45, 41] simulator to create a photorealistic indoor environment that includes acoustic simulation. Additionally, Soundspaces2.0 [14] allows for the continuous visual and acoustic simulation of arbitrary mesh indoor data.

To address the AudioGoal task, the majority of studies employ RL for training deep learning policies. For instance, AV-Nav[12] employs the Proximal Policy Optimization (PPO) RL algorithm [42] to train an LSTM model that predicts low-level actions (*e.g.*, turning left, moving forward) based on raw RGB/depth images and audio spectrograms at the current step. SAVi [11] incorporates a goal descriptor network for sound source direction prediction and a transformer for sequence modelling, aimed at addressing the audio-visual navigation problem with a temporary sound source. Inspired by the study in other acoustic area [44, 21, 6, 17, 31], [53] proposes sound attackers as a means of enhancing the model's robustness in real-world scenarios. [46] introduces an agent capable of performing various navigational tasks, encompassing audio signals as well. Additionally, the incorporation of occupancy maps

into the model is introduced in the works of [19, 13]. The agent predicts intermediate goals on the map and utilizes a graph-based shortest path algorithm to reach the intended position. The implementation of these techniques reduces decision-making times, thus enhancing the agent's long-term navigation performance. The objective of this paper is to enhance the dependability and efficacy for the occupancy-map-based AudioGoal model.

PointGoal Navigation. PointGoal navigation is a classic problem in the robotics area. The objective of the task is for an agent to traverse from a randomized point of origin to a designated target location within an unfamiliar environment using GPS and visual data. The traditional way is to compose this task into several sub-tasks, *e.g.*, mapping, planning and control. Recently, [50] found that this task can be solved with RL based end-to-end model. After training with a large number of samples, the RNN-based navigation policy can achieve near-perfect performance. [52] proposes to apply the auxiliary losses to speed up the training process of the PointGoal policy. [28] attempts to apply the visual encoder of the PointGoal policy to other visual navigation tasks. [35] proposes to use the visual odometry module for PointGoal navigation in the presence of noise, which yields outstanding results. In this paper, we utilize the basic navigation knowledge in the PointGoal policy to improve the ability of the AudioGoal policy.

Policy Distillation. The knowledge distillation technique is widely studied in different deep learning area [16, 54, 20, 51], which aims to improve the performance of the student model by having it learn from the insights and knowledge acquired by the more complex and accurate teacher model. In RL, transferring knowledge from teacher policies has attracted great research interests. In particular, a student policy is trained to match the state-dependent probability distribution over actions provided by the teacher, while the student policy is usually of lighter architecture for faster inference. [22] devised an online policy distillation paradigm that enables student policies to have better adversarial robustness. Recently, [18] designed a teacher reward-based method for better leveraging imperfect teachers. [39] introduces frame importance analysis for the policy distillation training. In [33], they propose a dual policy distillation, in which two learners extract knowledge from each other to enhance their learning. In this paper, we apply policy distillation to transfer knowledge from PointGoal to AudioGoal. So far, cross-task policy distillation has been less studied in the field of embodied navigation.

3. Our Approach

We propose ORAN, a powerful AudioGoal navigation agent built with two essential technique contributions, namely i) confidence-aware cross-task policy distillation (CCPD, §3.2), and ii) omnidirectional visual-audio information gathering (OIG, §3.3), as shown in Figure 3. Before elaborating on the algorithm details, we first provide a overview of our task setup and basic architecture (§3.1).

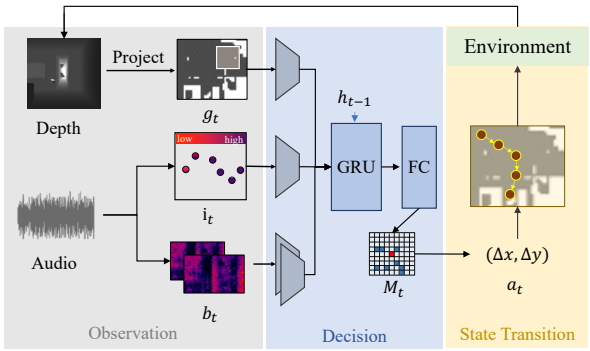

Figure 2: Illustration of the basic architecture. the navigation process can be expressed as the iterative execution of three steps, *i.e.*, observation, decision and state transition.

3.1. Problem Setup and Basic Architecture

Problem Setup. In the AudioGoal navigation task [12, 19], the agent is initially located at position p_0 in a 3D environment that is partially observable. The agent's objective is to navigate to a persistent sounding target located at p_τ. This task is typically formulated as a partially observable Markov decision process (POMDP) represented by a 7-tuple, $\{\mathcal{S}, \mathcal{A}, \Omega, O, \mathcal{T}, \mathcal{R}, \gamma\}$, where $\mathcal{S}$, $\mathcal{A}$, and Ω denote sets of states, actions, and observations, respectively. The observation function $o_t = O(s_t)$ maps the current state to an observation, where $o_t \in \Omega$ and $s_t \in \mathcal{S}$. The state transition function $s_{t+1} = \mathcal{T}(s_t, a_t)$ describes the evolution of the system, where $a_t \in \mathcal{A}$. The reward function $r_t = \mathcal{R}(s_t, a_t)$ assigns a scalar reward to the agent's action at time t, and γ is the discount factor used to weight future rewards. The navigation policy $\pi : \Omega \rightarrow \triangle^{|\mathcal{A}|}$ outputs the distribution of actions based on the current observation. The training target is to approach the optimal policy π^*:

$$\pi^* = argmax_\pi [\mathbb{E}_\pi (\Sigma_{t=1}^{|\tau|} \gamma^{t-1} r_t)]. \quad (1)$$

Basic Architecture. As depicted in Figure 2, the basic process of navigation can be formulated as the iterative execution of the following three steps [13]:

- *Observation* $O(s_t)$: The agent constructs top-down geometric and acoustic maps of the observed environment using egocentric visual-audio perceptions. At step t, the depth observation is back-projected to the 3D point cloud and then transformed into a 2D local map, which is used to update a *global geometry map* g_t. g_t has two channels, one for recording explored/unexplored areas and the other

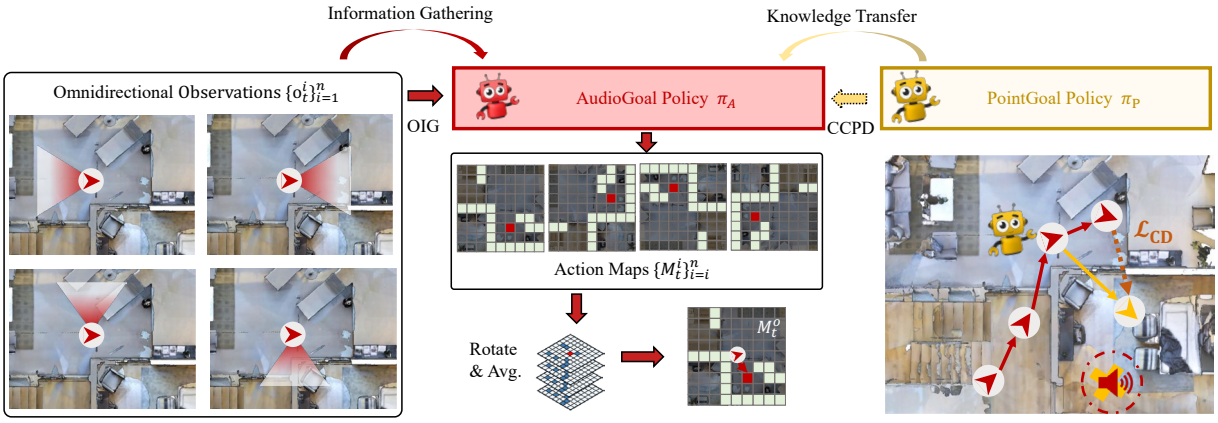

Figure 3: The overview of ORAN. There are two main components of the proposed ORAN, *i.e.* OIG (in left) and CCPD (in right). The CCPD transfers the navigation knowledge from a PointGoal agent during training, and OIG gathers omnidirectional information during inference. The OIG and CCPD together make the ORAN a powerful AudioGoal agent.

for occupied/free areas. Similarly, based on binaural audio perception b_t, a *global acoustic map* i_t is updated, which stores all sound intensities and their corresponding locations. The observation function can be formulated as:

$$o_t = O(s_t) = (g_t, i_t, b_t). \tag{2}$$

- *Decision* $\pi_A(o_t)$: A parameterized AudioGoal policy π_A predicts the next short-term navigation goal a_t based on multimodal cues and partial maps. We apply π_A as a GRU-based waypoint predictor to approximate π^*, which generates the next short-term navigation goal. π_A takes the current observation o_t as input and outputs a score map M_t of size $m \times m$, representing the probability distribution for the next waypoint:

$$M_t = \pi_A(o_t). \tag{3}$$

  M_t represents a set of $m \times m$ candidate waypoints and their associated action probabilities within a region centered around the agent. The waypoint predictor, denoted as π_A, is trained using reinforcement learning to predict actions at the waypoint level. Essentially, π_A serves as the AudioGoal navigation policy for action prediction. we sample the next waypoint a_t from M_t based on the probability predicted by π_A.
- *State Transition* $\mathcal{T}(s_t, a_t)$: The agent plans the shortest feasible path to the waypoint and executes the plan using low-level actions to obtain the new navigation state s_{t+1}. After selecting the waypoint a_t, the agent computes the shortest path from its current position to a_t using Dijkstra's algorithm. It then moves towards a_t by executing a sequence of low-level actions. This process is repeated until the agent either selects its standing location as the next waypoint or reaches a navigation step limit.

Based on this waypoint-based scheme, ORAN further improves its waypoint navigation policy through CCPD (§3.2), and changes its forward-only perception with 90-degree field of view to an omnidirectional visual-audio information gathering mode – OIG (§3.3).

3.2. Confidence-Aware Cross-task Policy Distillation

The PointGoal navigation policy is a valuable source of knowledge for indoor navigation. To transfer this knowledge, we propose using CCPD, which leverages a PointGoal policy (π_P) as a teacher to guide the training process of an AudioGoal policy (π_A). In each episode, the target p_τ for π_P and π_A is set to the same position, allowing them to share the same optimal policy π^*. Since π_P is trained using a large-scale training samples and achieves near-perfect performance, it provides a more accurate estimation of π^*. So we train π_A to imitate the actions of π_P, using policy distillation. The training target can be defined as follows:

$$\pi_A^* = argmin_{\pi_A} \mathbb{E}_\tau[\mathcal{D}(\pi_A, \pi_P)], \tag{4}$$

where $\mathcal{D}(\cdot,\cdot)$ is a distance metric function, and $\tau = \{o_1, o_2, \ldots, o_{|\tau|}\}$ are the observations from a sampled trajectory via a control policy. We utilize π_A as the control policy to implement student-forcing sampling, which helps to reduce the distribution distance between training and inference, as mentioned in multiple works [18, 33]. We set $\mathcal{D}(\cdot,\cdot)$ as the KL divergence function and the loss function of policy distillation can be defined as:

$$\mathcal{L}_D = \mathbb{E}_\tau[\mathcal{D}_{KL}(\pi_A, \pi_P)]. \tag{5}$$

However, the intrinsic differences between the AudioGoal and PointGoal tasks create a domain gap between the policies π_A and π_P. Applying distillation supervision $\mathcal{L}_D$ over all actions in a trajectory can introduce unnecessary bias due to this gap. To mitigate this side effect, we propose selecting a subset of steps for distillation based on the action confidence of π_P, *i.e.* the confidence-aware reweighting mechanism. We measure the confidence of π_P using the Shannon entropy, which is defined as:

$$\mathcal{H}[\pi_P(o_t)] = -\sum \pi_P(o_t) \log \pi_P(o_t). \tag{6}$$

Intuitively, the entropy of the action distribution increases as it becomes more even, indicating that the agent is less confident in choosing a particular action. Therefore, we choose the steps with lower entropy (*i.e.*, higher confidence) for distillation. We calculate the confidence factor of π_P as:

$$c_P(o_t) = \frac{1}{\mathcal{H}[\pi_P(o_t)]}, \tag{7}$$

The actions with higher $c_P(o_t)$ are crucial for navigation and less risk for π_A to follow. We rank the o_t in τ based on $c_P(o_t)$, and take the top k to compute the distillation loss. So the loss with the confidence-aware reweighting can be defined as:

$$\mathcal{L}_{CD} = \mathbb{E}_\tau\{\sum \mathbb{I}_k[c_P(o_t)]c_P(o_t)\mathcal{D}_{KL}[\pi_A(o_t), \pi_P(o_t)]\}. \tag{8}$$

The notation $\mathbb{I}_k(\cdot)$ denotes selecting the top k largest elements. In the training phase, we employ a combination of CCPD and PPO [42] to train the agent. With CCPD, the π_A can learn the wayfinding ability from π_P efficiently and performs better with less training samples, as shown in §4.

3.3. Omnidirectional Information Gathering

Another notable benefit of ORAN is its ability to facilitate an omnidirectional visual-acoustic information gathering process. As mentioned in §3.1, the decision model π_A presently relies solely on visual-audio information obtained from a single direction to set the subsequent intermediate goals through the self-centered action distribution map M_t. The $\mathcal{T}(s_t, a_t)$, which includes a non-parametric path planner, enables the agent to progress towards the next waypoint. Owing to the randomness in $\mathcal{T}(s_t, a_t)$, the agent's orientation on the subsequent waypoint is rendered indeterminate. So changing the agent's direction has limited impact on π_A while predicting the next waypoint, and the decision-making mode of π_A is omnidirectional. However, the forward-only perception contradicts the decision-making mode, making the intermediate goal unreliable for two reasons. Firstly, the agent's perception is limited in range, resulting in incomplete occupancy maps and audio information. This limitation impedes the agent's ability to understand its surrounding environment, thereby reducing the navigation's robustness. Secondly, in our experiments, we have observed that the predicted M_t exhibits a direction-relative bias. For instance, when the agent is uncertain about the direction to take, it may tend to gravitate towards a particular waypoint on M_t, potentially resulting in the agent becoming trapped in certain positions. To overcome these obstacles, we enable the agent to collect information and integrate action decisions from multiple directions, as depicted in Figure 3.

At each action step of π_A, we collect n observations $\{o_t^i\}_{i=1}^n$ from n directions in $\mathbf{\Omega} = \{\omega_i\}_{i=1}^n$ to obtain panoramic information. Next we input them to π_A to obtain the action maps $\{M_t^i\}_{i=1}^n$ in all directions :

$$M_t^i = \pi_A(o_t^i). \tag{9}$$

In order to collect the action distribution of various orientations, we rotate the action map to a consistent direction and then aggregate the distribution. Therefore, the resulting omnidirectional action map M_t^o can be expressed as follows:

$$M_t^o = \frac{1}{n}\sum_{i \le n} R(M_t^i, -\omega_i), \tag{10}$$

where $R(X, \theta)$ represents the rotation of the X matrix for θ degree. the next waypoint is sampled base on the distribution of M_t^o. To enhance the agent's ability to make accurate terminal judgments, we additionally employ a stop policy model to assist π_A in identifying the target position. With this OIG, the ORAN agent exhibits greater accuracy and efficiency in navigation without requiring finetuning.

4. Experiments

4.1. Experimental Setup

Environments. We utilize the Soundspaces to evaluate our approach which encompasses two separate environments, namely Matterport3D [7] and Replica [43]. To be consistent with prior work [12, 13], we divide the scenes into three sets, train/val/test, consisting of $9/4/5$ for the Replica and $73/11/18$ for the Matterport3D. We evaluate the performance of our agent in both environments under two different settings following [13]: heard and unheard. The unheard setting in Matterport3D is the most demanding, and it is also our primary focus.

Evaluation Metrics. We measure navigation performance using the following metrics: 1) Success Rate (SR): the ratio of agent trajectories where the agent stops at the goal position; 2) Success weighted by Path Length (SPL): Evaluating navigation efficiency involves factoring in the length of the agent's path and weighing its success accordingly; 3) SoftSPL: a version of SPL where binary success is replaced by progress towards the goal; 4) Success weighted by the Number of Actions (SNA): measures action efficiency via weighting success by the number of actions taken.

Model	*Replica*						*Matterport3D*					
	Heard			*Unheard*			*Heard*			*Unheard*		
	SNA	SR	SPL	SNA	SR	SPL	SNA	SR	SPL	SNA	SR	SPL
Random Agent [13]	1.8	18.5	4.9	1.8	18.5	4.9	0.8	9.1	2.1	0.8	9.1	2.1
Direction Follower [13]	41.1	72.0	54.7	8.4	17.2	11.1	23.8	41.2	32.3	10.7	18.0	13.9
Frontier Waypoints [13]	35.2	63.9	44.0	5.1	14.8	6.5	22.2	42.8	30.6	8.1	16.4	10.9
Supervised Waypoints [13]	48.5	88.1	59.1	10.1	43.1	14.1	16.2	36.2	21.0	2.9	8.8	4.1
Gan et al. [13]	47.9	83.1	57.6	5.7	15.7	7.5	17.1	37.9	22.8	3.6	10.2	5.0
AV-Nav [12]	52.7	94.5	78.2	16.7	50.9	34.7	32.6	71.3	55.1	12.8	40.1	25.9
AV-WaN [13]	70.7	98.7	86.6	27.1	52.8	34.7	54.8	93.6	72.3	30.6	56.7	40.9
ORAN (**ours**)	70.1	96.7	84.2	**36.5**	**60.9**	**46.7**	57.7	93.5	73.7	**35.3**	**59.4**	**50.8**

Table 1: **Comparisons on Soundspaces dataset**. Comparison with the state-of-the-art methods on the Soundspaces challenge dataset. The proposed ORAN boosts the performance in terms of all the key metrics on the unheard setting.

Name	CCPD	OIG	SR↑	SPL↑	SoftSPL↑
Baseline			53.1	38.6	45.0
#1	✓		**59.5**	45.1	49.0
#2	✓	✓	59.4	**50.8**	**55.7**

(a) **Detail analysis on essential components of ORAN**, *i.e.* CCPD and OIG on the Matterport3D unheard setting. The performance is gradually improved with the continuous addition of proposed modules.

Name	$\mathcal{L}_D$	$\mathcal{L}_{CD}$	SR↑	SPL↑	SoftSPL↑
Baseline			53.1	38.6	45.0
#1	✓		50.0	37.0	43.2
#2		✓	**59.5**	**45.1**	**64.0**

(b) **Ablation analysis on CCPD**. The ablation study on the effect of confidence-aware reweighting on the Matterport3D unheard setting. The Confidence-aware reweighting gains a large accuracy improvement.

Table 2: **Ablations.** The ablation study on the key component of ORAN, see §4.3 for details.

Implementation Details. We set up the simulator following [13]. The PointGoal navigation model is initialized with the pre-trained model in [50]. We substitute the action prediction layer and finetune the model on the training split to predict the action map as that of π_A. The learning rate is 2.5×10^{-4} and decreases linearly during training. We set the hyper-parameter of the PPO [42] following [13]. The loss weight for $\mathcal{L}_{CD}$ is 0.3. We select the top $k = 30$ most confident steps at each CCPD update step. Both our method and baseline employ the spectrogram data augmentation during the training phase of the unheard split. In OIG, we specify the $\mathbf{\Omega} = \{0°, 90°, 180°, 270°\}$. The actions for information collection of OIG are not considered in metric as they are irrelevant to the decision-making process.

Reproducibility. Our model is implemented in PyTorch, and trained on one NVIDIA TITAIN RTX GPU.

4.2. Comparison with State-of-the-art Methods

We compare our approach with existing methods, and the specifics of these methods are elaborated in the supplementary materials. Note that the metric used for ranking is the SPL on the unheard setting of Matterport3D in the Soundspaces 2022 challenge [12]. As shown in Table 1, our model attains comparable performance to the AV-WaN model [13] on the heard setting. Given the high SR observed in both environments on this setting, we primarily concentrate on comparing the performance on the unheard setting. In the Matterport3D environment, The ORAN model consistently outperforms all existing methods across all evaluation metrics. Specifically, ORAN achieves a SR of 59.4%, surpassing the second-best AV-WaN [13] approach by 2.7%. Additionally, ORAN demonstrates significant improvements in navigation efficiency, with a performance increase of 9.9% in SPL. In the Replica environment, ORAN outperforms previous methods by a wide margin, achieving results of (46.7%, 60.9%, 36.5%) in (SPL, SR, SNA) respectively, with a relative improvement of (35%, 15%, 35%) over AV-WaN [13]. These experimental results effectively demonstrate the efficacy of the ORAN model.

4.3. Ablation Study

In this section, a set of ablation studies are conducted in the matterport3D unheard setting to verify the proposed components, as shown in Table 2. Moreover, the design of the CCPD is also discussed in Table 3 and Table 4.

Ablations on CCPD. We first examine the influence of the CCPD. As shown in Table 2b, comparing with the base agent, #2 with the CCPD lifts SR and SPL from (53.1%, 38.6%) to (59.5%, 45.1%) on the unheard setting of Matterport3D. Additionally, as shown in Table 3, we train our agent to follow the oracle shortest path planner by randomly selecting 30 steps and computing the cross-entropy loss. The agent performs 3% lower in SR and 2% lower in SPL than trained with CCPD. This indicates that the knowledge transmitted via CCPD enhances fundamental navigational proficiency. Moreover, we also attempt to directly apply

Models	SR	SPL	SoftSPL
CCPD	**59.5**	**45.1**	**49.0**
Oracle Supervision	56.7	42.2	45.7
PointGoal + Pseudo-GPS	17.2	12.3	12.1
PointGoal	94.4	74.2	74.4

Table 3: **Other Analysis of CCPD**. The comparison of agents trained with CCPD, shortest path oracle supervision, and PointGoal agent navigating with pseudo-GPS predictor in the Matterport3D unheard setting.

k	SR↑	SPL↑	SoftSPL↑
10	57.7	37.1	41.8
30	**59.5**	**45.1**	**49.0**
50	56.1	39.6	44.1
100	51.2	37.3	42.5

Table 4: **Influence of Selected Steps.** The models are evaluated in the Matterport3D unheard setting. The model performs best with top 30 most confident steps.

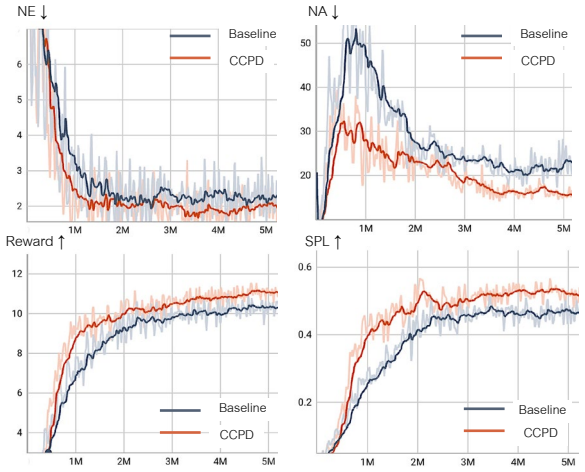

Figure 4: **Qualitative Analysis for CCPD.** The navigation metrics vary with the training steps. The model with CCPD exhibits faster convergence and superior performance.

the PointGoal agent for the AudioGoal task by introducing an audio-based direction predictor to provide pseudo-GPS for the PointGoal agent. From Table 3, we can observe that PointGoal + pseudo-GPS performs poorly in all metrics, which experimentally indicates the domain gap between PointGoal and AudioGoal, as another confirmation of the necessity of CCPD.

Ablations on OIG. Our proposed ORAN model incorporates the OIG technique, which allows the agent to utilize information from different directions, leading to a more robust navigation performance. As shown in Table 2a #2, the model with OIG achieves a significant improvement in both SPL and SoftSPL, from $(45.1\%, 49.0\%)$ to $(50.8\%, 55.7\%)$. This finding suggests that OIG enhances navigation efficiency in terms of distance traveled. We follow previous efforts [13] to view SPL as the primary ranking metric. It is worth mentioning that SPL and most metrics (*e.g.* NE, SR, and SoftSPL) are unaffected by the perception actions for OIG. The OIG also improves the PointGoal policy in Table 3, which achieves an SR of 95.1% and an SPL of 74.3%.

Confidence-Aware Reweighting. To confirm the essentiality of confidence-aware reweighting, we implement the basic policy distillation loss $\mathcal{L}_D$ described in Equation 5 in Table 2b #1. Comparing #2 to #1, the model trained with $\mathcal{L}_{CD}$ improves SR and SPL from $(50.0\%, 37.0\%)$ to $(59.5\%, 45.1\%)$. We can observe that the SR and SPL metrics show a decrease when trained without the reweighting, as compared to the baseline model. This outcome exemplifies the significance of employing confidence-based selection and reweighting during training.

Selected steps k in CCPD. In Table 4, we investigate the impact of the number of selected steps in CCPD, denoted by k, on the performance of the model. Specifically, we compare the performance of the model trained with the top $k = 10$, 30, 50, and 100 most confident steps. Notably, we observe that the model attains the best performance at $k = 30$, whereas the navigation performance decreases with an increase in k to 50 or 100. This implies that supplementary supervision does not yield any positive effects.

4.4. Qualitative Analysis

Qualitative Analysis for CCPD. In Figure 4, we investigate the influence of CCPD on the training phase of the AudioGoal agent by visualizing the curves of four navigation metrics — namely, NE (Navigation Error), NA (Navigation Actions), reward, and SPL — as they vary with the training steps. The model trained with CCPD achieves a lower level of NA and NE faster than the baseline model. During the initial training stage ($0 \rightarrow$ 1M steps), the reward and SPL of the model trained with CCPD grow faster and maintain a higher level throughout the remainder of the training phase. This demonstrates that CCPD, as a powerful knowledge transfer algorithm, helps the agent to better master AudioGoal navigation with far fewer training samples. Figure 5 depicts the comparison of the model trained with CCPD (in the first row) or not (in the second row) of the same episodes. We can observe that the model trained with CCPD performs better navigation robustness and efficiency, especially on the episodes with longer trajectories.

Qualitative Analysis for OIG. To analyze how the omnidirectional information benefits the navigation, we visualize

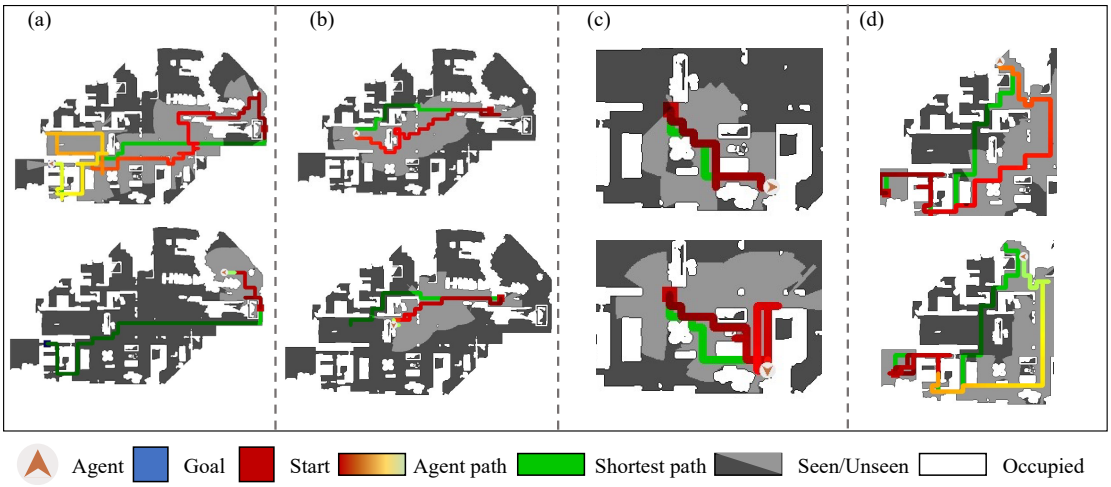

Figure 5: **Qualitative Analysis for CCPD.** The navigation trajectories of the agent trained with CCPD (in the first row) or not (in the second row). Each column is of the same test episode. The agent with CCPD shows more robust navigation performance, especially on long trajectories. The color of the agent's path turns from red to green gradually to indicate the trajectory length.

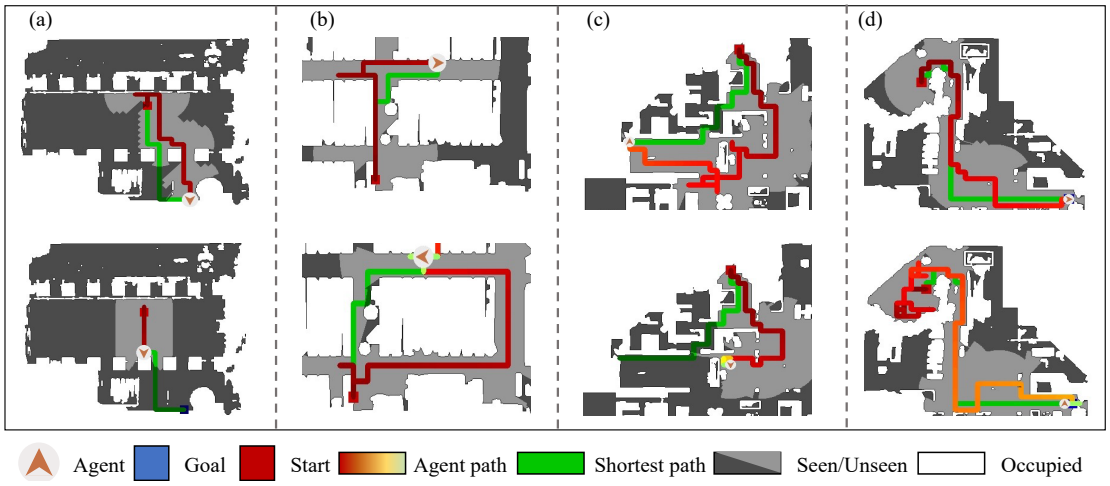

Figure 6: **Qualitative Analysis for OIG.** The navigation trajectories of the same agent with OIG (in the first row) or not (in the second row). Each column corresponds to the same test episode. The gradual transition of the path color from red to green denotes the trajectory length.

the same agent's trajectories equipped with OIG or not. We list here some key observations drawn from Figure 6: First, the agent with OIG can distinguish whether arrive at the sounding target more precisely, as in episode (a) and avoid hanging around the target, as in episode (b). Second, OIG gets the agent out of the deadlock, as in episode (c). Third, the agent with OIG is more likely to find the right direction towards the target and reduce detours, as in episodes (b) and (d). These phenomena intuitively show the help of omnidirectional information to the agent's navigation.

4.5. Model Ensemble

We apply post-fusion on models with diverse network architectures and training strategies to push the limits of the AudioGoal task. At each time step, we feed the current observation into each of these models and then average the

#	OIG	*Fused Models*				*Matterport3D Unheard*			
		A:40.9	B:43.5	C:42.4	D:43.1	SPL↑	SR↑	SNA↑	SoftSPL↑
1		✓				40.9	56.7	30.6	46.3
2		✓	✓			47.5	66.1	35.3	51.7
3		✓	✓	✓		48.0	64.7	35.9	53.7
4		✓	✓	✓	✓	50.0	66.1	37.2	54.7
5	✓	✓	✓	✓	✓	**62.6**	**77.7**	**42.2**	**67.2**

Table 5: **Model ensemble**. Fusing model with different accuracy can increase the navigation accuracy. OIG also works on the Fusion model.

resulting action maps from all models to obtain the final action distribution. Using a grid search strategy, we select a combination of trained models that yield promising results. As shown in Table 5, the selected models achieved SPL percentages of 40.9%, 43.5%, 42.4%, and 43.1%, respectively. Further details about these models are provided in the supplemental material. Notably, we observe that as we increase the number of models from #1 to #4, the gain in performance obtained from the model ensemble decreases. By employing the post-fusion strategy, our agent achieves an improvement of 6.5% in SPL accuracy over the best single model (from 43.5% to 50.0%). Additionally, compared to model #4, model #5 incorporating OIG technology yields significant enhancements in accuracy, with absolute improvements of (12.6%, 11.6%, 5.0%, and 12.5%) observed in SPL, SR, SNA, and SoftSPL metrics, respectively. These results highlight the inadequacy of forward-only perception in providing crucial information required for the AudioGoal wayfinding process.

5. Conclusion

We present an omnidirectional audio-visual navigator based on cross-task navigation skill transfer, named ORAN, that advances state-of-the-art technologies for AudioGoal in two aspects. Firstly, we transfer the navigation knowledge from the PointGoal policy via a cross-task policy distillation to sharpen the basic point-to-point wayfinding ability. We further consider the action confidence of the PointGoal agent to guide the distillation to overcome the domain gap between the two tasks. Secondly, we equip the AudioGoal agent with the ability to gather information from different directions, instead of the only direction it is facing, to better support the omnidirectional decision-making. Moreover, after model assembling, ORAN yields 12% absolute promotion in SPL, which makes the champion solution of the Soundspaces challenge in 2022.

Acknowledgements. This work is supported in part by the National Key R&D Program of China under Grant 2022ZD0115502, the National Natural Science Foundation of China under Grant 62122010, the Fundamental Research Funds for the Central Universities, the Fundamental Research Funds for the Central Universities (No. 226-2022-00051), and CCF-Tencent Open Fund.

References

[1] Ziad Al-Halah, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Zero experience required: Plug & play modular transfer learning for semantic visual navigation. In *CVPR*, 2022. 1

[2] Dong An, Yuankai Qi, Yan Huang, Qi Wu, Liang Wang, and Tieniu Tan. Neighbor-view enhanced model for vision and language navigation. In *ACM MM*, 2021. 1

[3] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. Bevbert: Topo-metric map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385*, 2022. 1

[4] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018. 1

[5] Richard Brown. *Consciousness inside and out: Phenomenology, neuroscience, and the nature of experience*. Springer Science & Business Media, 2013. 2

[6] Nicholas Carlini and David A. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *IEEE Security and Privacy Workshops*, 2018. 2

[7] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 5

[8] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Russ R. Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *NeurIPS*, 2020. 1

[9] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological SLAM for visual navigation. In *CVPR*, 2020. 1

[10] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, 2021. 1

[11] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *CVPR*, 2021. 2

[12] 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.

1, 2, 3, 5, 6

[13] Changan Chen, Sagnik Majumder, Al-Halah Ziad, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *ICLR*, 2021. 2, 3, 5, 6, 7

[14] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. In *NeurIPS 2022*, 2022. 2

[15] Jinyu Chen, Chen Gao, Erli Meng, Qiong Zhang, and Si Liu. Reinforced structured state-evolution for vision-language navigation. In *CVPR*, 2022. 1

[16] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *CVPR*, 2019. 3

[17] Moustapha Cissé, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples. In *NeurIPS*, 2017. 2

[18] Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. Distilling policy distillation. In *AISTATS*, 2019. 3, 4

[19] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. In *NeurIPS*, 2020. 1, 2, 3

[20] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, pages 9737–9746, 2022. 3

[21] Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. Sirenattack: Generating adversarial audio for end-to-end acoustic systems. In Hung-Min Sun, Shiuh-Pyng Shieh, Guofei Gu, and Giuseppe Ateniese, editors, *Asia Conference on Computer and Communications Security*, 2020. 2

[22] Marc Fischer, Matthew Mirman, Steven Stalder, and Martin T. Vechev. Online robustness training for deep reinforcement learning. *CoRR*, abs/1911.00887, 2019. 3

[23] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *NeurIPS*, 2018. 1

[24] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B. Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *ICRA*, 2020. 2

[25] Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *CVPR*, 2021. 1

[26] Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. In *ICRA*. IEEE, 2022. 1

[27] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *CVPR*, 2022. 1

[28] Daniel Gordon, Abhishek Kadian, Devi Parikh, Judy Hoffman, and Dhruv Batra. Splitnet: Sim2sim and task2task transfer for embodied visual navigation. In *ICCV*, 2019. 3

[29] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *CVPR*, 2017. 1

[30] Y Hong, Q Wu, Y Qi, C Rodriguez-Opazo, and S Gould. A recurrent vision-and-language bert for navigation. arxiv 2021. *arXiv preprint arXiv:2011.13922*, 2021. 1

[31] Dan Iter, Jade Huang, and Mike Jermann. Generating adversarial examples for speech recognition. *Stanford Technical Report*, 2017. 2

[32] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv*, 2017. 2

[33] Kwei-Herng Lai, Daochen Zha, Yuening Li, and Xia Hu. Dual policy distillation. In *IJCAI*, 2020. 3, 4

[34] Yunlian Lyu, Yimin Shi, and Xianggang Zhang. Improving target-driven visual navigation with attention on 3d spatial relationships. *Neural Processing Letters*, 54(5), 2022. 1

[35] Ruslan Partsey, Erik Wijmans, Naoki Yokoyama, Oles Dobosevych, Dhruv Batra, and Oleksandr Maksymets. Is mapping necessary for realistic pointgoal navigation? In *CVPR*, 2022. 2, 3

[36] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, 2020. 1

[37] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop: history-and-order aware pre-training for vision-and-language navigation. In *CVPR*, 2022. 1

[38] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[39] Xinghua Qu, Yew Soon Ong, Abhishek Gupta, Pengfei Wei, Zhu Sun, and Zejun Ma. Importance prioritized policy distillation. In *ACM KDD*, 2022. 3

[40] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. In *ICLR*, 2018. 1

[41] 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. 2

[42] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 1, 2, 5, 6

[43] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 5

[44] Vinod Subramanian, Arjun Pankajakshan, Emmanouil Benetos, Ning Xu, SKoT McDonald, and Mark B. Sandler. A study on the transferability of adversarial attacks in sound

event classification. In *IEEE ICASSP*, 2020. 2

[45] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, 2021. 2

[46] Hanqing Wang, Wei Liang, Luc V Gool, and Wenguan Wang. Towards versatile embodied navigation. *NeurIPS*, 2022. 2

[47] Hanqing Wang, Wei Liang, Jianbing Shen, Luc Van Gool, and Wenguan Wang. Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In *CVPR*, 2022. 1

[48] Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. Structured scene memory for vision-language navigation. In *CVPR*, 2021. 1

[49] Hanqing Wang, Wenguan Wang, Tianmin Shu, Wei Liang, and Jianbing Shen. Active visual information gathering for vision-language navigation. In *ECCV*, 2020. 1

[50] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *ICLR*, 2019. 2, 3, 6

[51] Huanlai Xing, Zhiwen Xiao, Dawei Zhan, Shouxi Luo, Penglin Dai, and Ke Li. Selfmatch: Robust semisupervised time-series classification with self-distillation. *International Journal of Intelligent Systems*, 2022. 3

[52] Joel Ye, Dhruv Batra, Erik Wijmans, and Abhishek Das. Auxiliary tasks speed up learning point goal navigation. In *Conference on Robot Learning*, pages 498–516. PMLR, 2021. 3

[53] Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, and Xiaohong Liu. Sound adversarial audio-visual navigation. In *ICLR*, 2022. 2

[54] Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, pages 11953–11962, 2022. 3

[55] Yusheng Zhao, Jinyu Chen, Chen Gao, Wenguan Wang, Lirong Yang, Haibing Ren, Huaxia Xia, and Si Liu. Target-driven structured transformer planner for vision-language navigation. In *ACM MM*, 2022. 1

[56] Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. Diagnosing vision-and-language navigation: What really matters. *arXiv preprint arXiv:2103.16561*, 2021. 1

[57] Yuke Zhu, Daniel Gordon, Eric Kolve, Dieter Fox, Li Fei-Fei, Abhinav Gupta, Roozbeh Mottaghi, and Ali Farhadi. Visual semantic planning using deep successor representations. In *ICCV*, 2017. 1