LETTER **Communicated by Xiaolong Liu**

Echo-Enhanced Embodied Visual Navigation

Yinfeng Yu
yyf17@mails.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

Lele Cao
lele.cao@eqtpartners.com
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, and Motherbrain, EQT, Stockholm 11153, Sweden

Fuchun Sun
fcsun@mail.tsinghua.edu.cn
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Chao Yang
yangchao@pjlab.org.cn
Shanghai AI Laboratory, Shanghai 200232, China

Huicheng Lai
lai@xju.edu.cn
College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China

Wenbing Huang
hwenbing@126.com
Institute for AI Industry Research, Tsinghua University, Beijing 100084, China

Visual navigation involves a movable robotic agent striving to reach a point goal (target location) using vision sensory input. While navigation with ideal visibility has seen plenty of success, it becomes challenging in suboptimal visual conditions like poor illumination, where traditional approaches suffer from severe performance degradation. We propose E3VN (echo-enhanced embodied visual navigation) to effectively perceive the surroundings even under poor visibility to mitigate this

Fuchun Sun is the corresponding author. Yinfeng Yu and Lele Cao are equal contributors.

Neural Computation **35, 958–976** (2023)
https://doi.org/10.1162/neco_a_01579

problem. This is made possible by adopting an echoer that actively perceives the environment via auditory signals. E3VN models the robot agent as playing a cooperative Markov game with that echoer. The action policies of robot and echoer are jointly optimized to maximize the reward in a two-stream actor-critic architecture. During optimization, the reward is also adaptively decomposed into the robot and echoer parts. Our experiments and ablation studies show that E3VN is consistently effective and robust in point goal navigation tasks, especially under nonideal visibility.

1 Introduction

In robotic navigation tasks, the autonomous subjects (i.e., robots) interact with their environment in a continuous cycle of action and perception. Specifically, the robot needs to reason wisely based on all available senses, such as visual, auditory, proprioceptive, and tactile. The goal is to choose a sequence of appropriate actions to maximize the quality and speed of task completion.

For example, service robots may need to navigate autonomously to find and fetch objects for users. Traditionally, 3D reconstruction algorithms like SLAM (simultaneous localization and mapping; Chaplot et al., 2020; Karkus et al., 2021) are adopted to build a map that is used in path planning (Gupta et al., 2017). In contrast, recent work directly learns navigation strategies from egocentric observations (Mirowski et al., 2017). A prevalent stream in this category is PointGoal navigation, where the target position is revealed to the robot in the form of a displacement vector relative to the robot location (Savva et al., 2019). Most of the recent advances in solving PointGoal navigation tasks utilize only the visual sensory input (hence the term *visual navigation*), as illustrated in Figure 1A. There are some recent advances in visual navigation research: Gordon et al. (2019) investigates the transferability between different simulators and tasks types; Ye et al. (2020) and Wijmans et al. (2020) propose to speed up learning via auxiliary tasks; Ramakrishnan et al. (2020) encourage exploration with occupancy anticipation; and Morad et al. (2021) adopted autocurriculum learning.

While the visual sensory input dominates the robotic navigation tasks, the situation in the biological world is slightly different: several animal species (e.g., bats, dolphins and whales) and even people with impaired vision have echolocation capability (Christensen et al., 2020; Tracy and Kottege, 2021), where they use sound to perceive spatial layout (Purushwalkam et al., 2021) and locate objects (Yu, Huang et al., 2022; Gan et al., 2022; Yu, Cao et al., 2022) in the wild. Inspired by echolocation, VisualEchoes (Gao et al., 2020) uses echo and visual input simultaneously to obtain an improved visual representation using self-supervised pretraining tasks. The learned visual representation is expected to benefit from the echo input and accomplish many downstream tasks better.

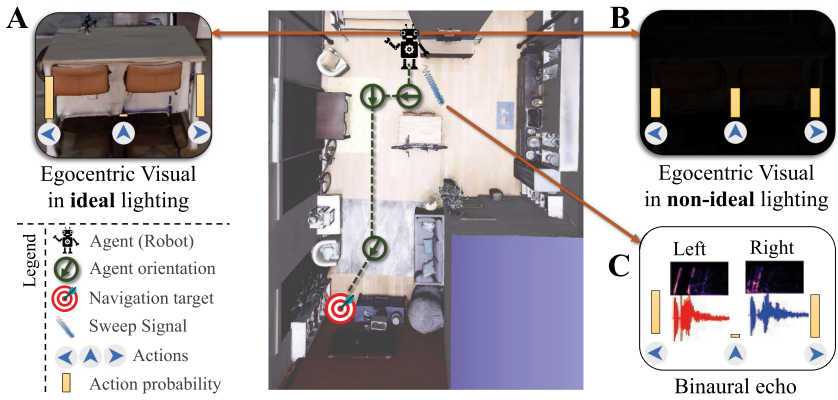


Figure 1: The illustration of echo-enhanced embodied 3D visual navigation. (A) Visual navigation under excellent lighting conditions. (B) Under poor lighting conditions, the performance of traditional visual navigation is degraded. (C) We propose to equip the robot with a sound generator, the environmental echo of which is used to maintain good navigation performance in poor lighting conditions.

Practically, the quality of the signal captured by visual sensors is vulnerable to environmental turbulence, typically interfering and degrading the quality of visual sensory input. One commonly seen scenario is poor lighting conditions, where the vision sensor produces images with low brightness and contrast, as exemplified in Figure 1B. In that case, how can we guarantee the successful completion of visual navigation tasks? As a result, reliable visual navigation under nonideal visibility circumstances is regarded as an important research problem. This work is dedicated to addressing this problem by introducing acoustic input from an echo generator mounted on the robot. The echo generator, that is, the echoer, is typically more cost-effective than some active sensors such as LiDAR. The echoer proactively emits sweep signals following a strategy (continuously learned under the condition of poor visibility), while the "ear" of the robot receives the echo of the emitted signal, as demonstrated in Figure 1C. An optimal navigation strategy can be learned using both the received echo and the egocentric visual input. By comparing the navigation performance under nonideal and ideal lighting conditions, we empirically show in section 4 that poor lighting conditions cause significant navigation difficulties. Therefore, we propose an end-to-end learning approach termed echo-enhanced embodied visual navigation (E3VN) to improve visual navigation performance under poor lighting conditions. The main contributions of this work are summarized as follows.

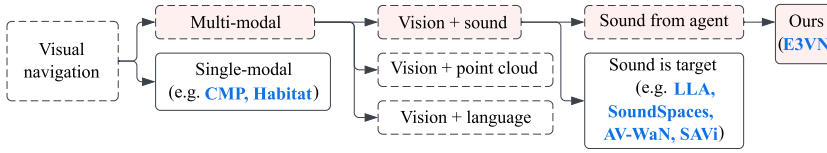


Figure 2: The landscape of the related work addressing visual navigation. We focus on multimodal (vision + sound) setups where sound is actively emitted from the agent.

- In addition to the visual sensor, we employ an echoer whose signal category, volume, and direction are treated as learnable actions.
- The robot and echoer policies are jointly optimized in a two-stream actor-critic paradigm, during which the overall reward is parametrically decomposed into the robot and echoer parts.
- Our comprehensive experiments and ablation studies validate the superior effectiveness and robustness of E3VN.

2 Related Work

According to the adopted type(s) of sensor input, visual navigation methods are either single modal and multimodal, as illustrated in Figure 2.

Single-modal visual navigation uses only data collected from visual sensors as the input of navigation tasks. The single-modal visual input still dominates the visual navigation research, such as CMP (cognitive mapping and planning; Gupta et al., 2017) and Habitat (Savva et al., 2019).

Multimodal navigation uses different modes of sensory input to navigate the robot. According to the combination of sensors used, multimodal navigation embodies several main categories: visual+language (Irshad et al., 2021; Wang et al., 2021; Kurita and Cho, 2021), visual+point cloud (Teng et al., 2019), and visual+sound (Dean et al., 2020), among others. Among these categories, visual+sound is the main focus (see the boxes with red background in Figure 2) of this letter.

Visual+sound navigation requires both visual and sound sensors. There are largely two forms of utilizing sound information: (1) the sound source acts as the navigation target, which comprises work such as LLA (Look Listen & Act; Gan et al., 2020); SoundSpaces (Chen et al., 2020); AV-WaN (Chen, Majumder et al., 2021); and SAVi (Chen, Al-Halah et al., 2021); and (2) the robotic agent actively emits sound as an active perception, which is a rarely seen (e.g., Gao et al., 2020) setup we try to address in this work. However, the majority of the

work in both forms merely utilizes sound in pretraining to enhance the existing visual representation.

It is commonly believed that pretraining is beneficial to representation learning (Beery et al., 2020; Chen et al., 2017; Vaswani et al., 2017; Fan et al., 2021; Qin et al., 2021) and spatial reasoning in navigation (Chen, Chen et al., 2021; Hong et al., 2021). Therefore, work like VisualEchoes (Gao et al., 2020) emerged to pretrain visual encoders using both visual and echo input in a supervised fashion. Concretely, it performs supervised learning using echo and vision as input before using the learned visual encoder for navigation. This approach has shown that visual representations learned in this way are helpful for tasks that require spatial reasoning. Compared to this pretraining paradigm, the main advantage of our work lies in the direct use of echo input during navigation.

Our work belongs to multimodal navigation with some unique innovations. First, unlike the work that uses sound sources as the navigation target (Chen et al., 2020), the sound in our work is emitted from a moving robot. Second, unlike the common setup (e.g., Savva et al., 2019) with only one agent (the robot), our work treats the robot and echoer as two cooperative agents, where the echoer is a trainable agent whose actions include the category, direction, and volume of the sound. Third, different from VisualEchoes (Gao et al., 2020) that use echo only in pretraining the visual encoder, our work uses echo as a real-time input throughout the entire navigation procedure. Fourth, the robot and echoer jointly optimize through a learnable reward distribution strategy. In a nutshell, our research aims to use the active perception of vision and echo input to enhance the performance and robustness of point navigation tasks in 3D scenes.

3 The Proposed Approach: E3VN

We propose a novel approach, E3VN, to tackle the echo enhanced embodied visual navigation tasks. E3VN models the robot agent as playing a Markov game with an echoer agent. The overall algorithm is illustrated in Figure 3. E3VN has four main modules: the robot agent, the ehoer agent, the reward assignment module, and the critic.

3.1 Problem Definition and Notations. We denote the robot and echoer with superscript ω and v, respectively. They play a game denoted as $\mathcal{M} = (\mathcal{S}, (\mathcal{A}^\omega, \mathcal{A}^v), \mathcal{P}, (\mathcal{R}^\omega, \mathcal{R}^v))$, where $\mathcal{S}$ is the state set, $\mathcal{A}^\omega$ is the robot action set, $\mathcal{A}^v$ is the echoer action set, and $\mathcal{P}: \mathcal{S} \times \mathcal{A}^\omega \times \mathcal{A}^v \to \mathcal{S}$ is a joint state transition function. The robot and echoer reward functions $\mathcal{R}^\omega : \mathcal{S} \times \mathcal{A}^\omega \times \mathcal{A}^v \times \mathcal{S} \to \mathbb{R}$ and $\mathcal{R}^v : \mathcal{S} \times \mathcal{A}^\omega \times \mathcal{A}^v \times \mathcal{S} \to \mathbb{R}$ depend on the current state, the next state, and the actions taken by robot and echoer, respectively. Each player wishes to maximize their discounted accumulative rewards. We use r to denote the reward given by the environment at every time step in an episode.

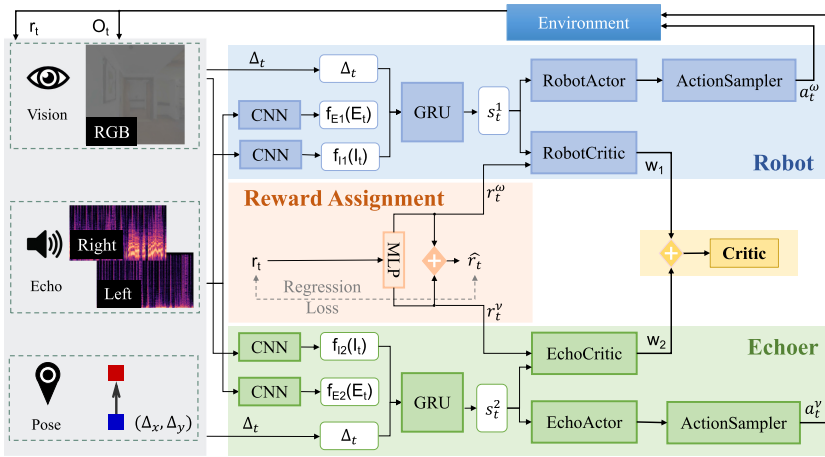
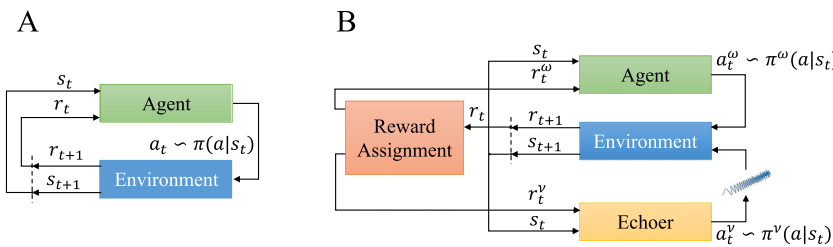


Figure 3: E3VN architecture: the robot and echoer first learn to encode observations as s_t^1 and s_t^2 respectively, which are fed to actor-critic networks to predict the next actions a_t^ω and a_t^ν. The reward assignment module decomposes the reward into the robot (r_t^ω) and echoer (r_t^ν) parts.



The standard MDP. E3VN (ours).

Figure 4: A comparison between the standard MDP and our E3VN.

As illustrated in Figure 4A, visual navigation can be modeled as a standard MDP (Markov decision process): $\pi^\star = \arg\max_{\pi \in \Pi} G(\pi)$. Seen from Figure 4B, E3VN is modeled as a multiagent (Sunehag et al., 2018; Rashid et al., 2018) problem involving two collaborating players sharing the same goal:

$$\pi^\star = \underset{\pi^\omega \in \Pi^\omega,\, \pi^\nu \in \Pi^\nu}{\arg\max}\; G(\pi^\omega,\, \pi^\nu,\, r) = \left(\underset{\pi^\omega \in \Pi^\omega}{\arg\max}\, G(\pi^\omega,\, r),\; \underset{\pi^\nu \in \Pi^\nu}{\arg\max}\, G(\pi^\nu,\, r) \right),$$

$$\text{s.t.} \qquad G(\pi^\omega, \pi^\nu, r) = w^\omega G(\pi^\omega, r) + w^\nu G(\pi^\nu, r),$$

$$G(\pi^{\omega}, r) = \sum_{t=0} \gamma^t r_t \rho^{\omega}, \quad G(\pi^{v}, r) = \sum_{t=0} \gamma^t r_t \rho^{v},$$
$$\rho^{\omega} = (1-\rho)/2 + \rho \cdot \rho_R^{\omega}, \quad \rho^{v} = (1-\rho)/2 + \rho \cdot \rho_R^{v}, \quad \rho_R^{\omega} + \rho_R^{v} = 1,$$
$$w^{\omega} > 0, \quad w^{v} > 0, \quad 0 \le \rho \le 1, \quad 0 \le \rho_R^{\omega} \le 1, \quad 0 \le \rho_R^{v} \le 1, \quad (3.1)$$

where $G(\pi^{\omega},\ \pi^{v},\ r)$ is the expected joint rewards for the agent and echoer as a whole. $G(\pi^{\omega},\ r)$ and $G(\pi^{v},\ r)$ are the discounted and cumulative rewards for the agent and echoer, respectively. w^{ω} and w^{v} denote the constant cumulative rewards balance factors. ρ^{ω} and ρ^{v} are the immediate reward contribution. ρ_R^{ω} and ρ_R^{v} represent the trainable reward decomposition weights. ρ is a constant (throughout the training) reward allocation parameter. The theoretical derivation of equation 3.1 is detailed in appendix A.

3.2 Action Spaces and Reward. We adopt the same robot action space as SoundSpaces (Chen et al., 2020): *MoveForward*, *TurnLeft*, *TurnRight*, and *Stop*; and appendix C has more details. In SoundSpaces, the echoer emits a chosen sound from the current robot position; the emitted omnidirectional audio is convolved with the corresponding binaural RIR (room impulse responses) to generate a binaural response that is "heard" by the robot. In this sense, the echoer's sound input is informative about the reflections on the surface of objects, making it physically admissible and realistic. Moreover, the echo contains the geometry and material information of the object being sensed. The echoer has a hybrid action space $\mathcal{A}^{v} = \mathcal{A}^{v,\text{cat}} \times \mathcal{A}^{v,\text{vol}} \times \mathcal{A}^{v,\text{dir}}$, which is the Cartesian product of three subspaces: category $\mathcal{A}^{v,\text{cat}}$, volume $\mathcal{A}^{v,\text{vol}}$, and direction $\mathcal{A}^{v,\text{dir}}$.

The reward is calculated based on two factors: (1) how far the robot is from the navigation target and (2) whether it succeeds in reaching it. Specifically, if the robot successfully reaches the target and executes the *Stop* action, it is rewarded with $+10$, plus an additional bonus of 0.25 to reward a shorter Manhattan distance to the target. To encourage faster navigation, we impose a time penalty of -0.01 on each action performed.

3.3 Joint Optimization of Robot and Echoer. At each time step t, the agents (robot and echoer) observe a state $O_t = (I_t, E_t, \Delta_t)$ where I is the egocentric visual input (i.e., the RGB image); E is the received echo in the form of a binaural audio waveform represented as a two-channel spectrogram; $\Delta = (\Delta x,\ \Delta y)$ is a relative displacement vector from the agent to the goal in the 2D ground plane of the scene by pose sensor.

E3VN, as shown in Figure 3, starts with encoding the visual and echo input using a convolution neural network (CNN), respectively. The CNNs generate visual vector $f_{I1}(I_t)$ and echo vector $f_{E1}(E_t)$. Then we concatenate the two vectors together with Δ to obtain the global observation embedding $e^1 = [f_{I1}(I_t),\ f_{E1}(E_t),\ \Delta_t]$. We transform the observation embeddings to state representations using a gated recurrent unit (GRU),

$s_t^1 = \text{GRU}(e_t^1,\ h_{t-1}^1)$. For echoer, we adopt a similar procedure to obtain s_t^2. The state representations, s_t^1 and s_t^2 are respectively fed to an actor-critic network to predict the action distribution (i.e., $\pi_\theta^\omega(a_t^\omega|s_t^1,\ h_{t-1}^1)$ for the robot and $\pi_\theta^\nu(a_t^\nu|s_t^2,\ h_{t-1}^2)$ for the echoer) and state value (i.e., $V_\theta^\omega(s_t^1,\ h_{t-1}^1)$ for the robot and $V_\theta^\nu(s_t^2,\ h_{t-1}^2)$ for the echoer). The actors and critics are approximated by single linear layer neural networks.

Finally, two action samplers sample the next actions, a_t^ω and a_t^ν, from the action distributions. The overall critic is a linear sum of RobotCritic and EchoCritic, as shown in Figure 3. $\mathcal{C}^j$ corresponds to four different loss components by substituting the superscript "j" with "ν, cat," "ν, vol," "ν, dir," and "ω," respectively.

$$\mathcal{C}^j = \frac{1}{2}\sum(V^j(s) - \hat{V}_{\theta^j}(s))^2 - \sum[\hat{A}^j \log(\pi_{\theta^j}(a \mid s)) + \beta \cdot H(\pi_{\theta^j}(a \mid s))], \quad (3.2)$$

where $\mathrm{V}^j(s) = \max_{a \in \mathbb{A}^j} \mathbb{E}[r_t + \gamma \cdot \mathrm{V}^j(s_{t+1}) \mid s_t = s]$ and $\hat{\mathrm{V}}_{\theta^j}(s)$ is the state value for the target network. Notation $\hat{\mathrm{A}}_t^j = \sum_{i=t}^{T-1} \gamma^{i+2-t} \cdot \delta_i^j \hat{\mathrm{A}}_t^j$ is the advantage for a given length-T trajectory, where $\delta_t^j = r_t + \gamma \cdot \mathrm{V}^j(s_{t+1}) - \mathrm{V}^j(s_t)$. The overall loss $\mathcal{L}$ to optimize is formulated based on loss for each actor-critic branch:

$$\begin{aligned} \mathcal{L} &= w_1\mathcal{L}^\nu + w_2\mathcal{C}^\omega + w_3\mathcal{L}^r\text{, where} \\ \mathcal{L}^\nu &= \frac{1}{3}(\mathcal{C}^{\nu,\text{cat}} + \mathcal{C}^{\nu,\text{vol}} + \mathcal{C}^{\nu,\text{dir}})\text{, and} \\ \mathcal{L}^r &= \sum(r - r^\omega - r^\nu)^2, \end{aligned} \quad (3.3)$$

where r^ω and r^ν are the predicted reward (for the tth time step) of the robot and echoer, respectively. r is the reward obtained from the environment. $\mathcal{L}^r$ is the regression loss for the reward assignment module. The weights $w_1,\ w_2,\ w_3$ should add up to 1.0 exactly: $w_1 + w_2 + w_3 = 1.0$. The overall loss $\mathcal{L}$ is minimized following proximal policy optimization (PPO; Schulman et al., 2017). The entire procedure is illustrated in algorithm 1 in the form of pseudocode.

3.4 Reward Assignment. The distribution of rewards is a combination of trainable and fixed weighting:

$$\rho^\omega = \frac{1-\rho}{2} + \rho_R^\omega \cdot \rho, \quad \rho^\nu = \frac{1-\rho}{2} + \rho_R^\nu \cdot \rho, \quad \mathcal{L}^\rho = \sum(1 - \rho_R^\omega - \rho_R^\nu)^2, \quad (3.4)$$

where ρ is a constant weight parameter; ρ_R^ω and ρ_R^ν are reward weights predicted by neural networks; and ρ^ω and ρ^ν are immediate reward weights.

Algorithm 1: Echo-Enhanced Embodied Visual Navigation.

Data: Environment $\mathcal{E}$, initial policies π_{θ_0}, # rollout episodes M, # updates N.

Result: π_{θ_N}

for i=*1, 2, ...* N **do**

  $\{(o_t,\ h_{t-1},\ a_t,\ r_t)_i\}_{t=1}^{T} \leftarrow \text{roll}(\mathcal{E}, \pi_{\theta_{i-1}}, T)$, // Run $\pi_{\theta_{i-1}}$ for M episodes;

  Compute advantage estimates $\hat{A}_1, \cdots, \hat{A}_T$;

  $\theta_i \leftarrow \theta_{i-1}$, // Optimize $\mathcal{L}$ in equation 3.3 w.r.t. θ ;

end

$\mathcal{L}^{\rho}$ is equivalent to $\mathcal{L}^{r}$ in equation 3.3. Reward r^{ω} and r^{v} can be respectively calculated with $r^{\omega} = r \cdot \rho^{\omega}$ and $r^{v} = r \cdot \rho^{v}$.

4 Experiments and Results

We evaluate the performance of E3VN on the PointGoal navigation (Savva et al., 2019) tasks, where the robot moves inside one of the two 3D environments (Chen et al., 2020): Replica (Straub et al., 2019) and Matterport3D (Chang et al., 2017). The scene map is unknown to the robot. Hence, the robot has to incrementally gather observations to understand the scene. The original environments only provide RIRs for generating sounds when an agent is facing 0, 90, 180, or 270 degrees; to obtain a better granularity, we augment (see appendix C.2) the original RIRs to $\{0, 15, \ldots, 345\}$.

The baseline methods selected to benchmark E3VN include (1) Random action sampling, (2) PointGoal (Savva et al., 2019), (3) VisualEchoes (Gao et al., 2020), and (4) PointGoal+: same as Savva et al. (2019) except zero-mean and unit-variance normalization of RGB images.

As of evaluation metrics, we calculate SPL (success weighted by path length; Anderson et al., 2018), SSPL (soft SPL), SR (success rate), $\mathbf{R}_{\text{mean}}$ (average episode reward), DTG (distance to goal), and NDTG (normalized DTG). Appendix D provides the definition of these metrics, which are commonly adopted. All reported metrics are averaged over five runs. Since every metric value has a standard deviation (STDEV) lower than 0.01, for conciseness, we do not show the value of STDEV.

In order to test the model's performance in nonideal illumination conditions, we simulate the nonideal input I^{l} from the raw visual input I:

$$I_t^l = f(M(M^{-1}(f^{-1}(I_t \cdot \alpha)) + n_s + n_c)), \tag{4.1}$$

where I_t^l is the simulated visual input for the tth time steps, and I_t is the corresponding raw visual input captured by the sensor in Habitat Simulator

Table 1: Echo Boosts Other Modalities.

Vision	Echoer	SPL (↑)	SR (↑)	R_{mean} (↑)	DTG (↓)
RGB	✓	**0.481**	**0.562**	**6.7**	**4.29**
RGB	✗	0.440	0.542	6.4	4.42
Depth	✓	**0.511**	**0.637**	**8.3**	**3.90**
Depth	✗	0.484	0.582	7.3	4.11
RGBD	✓	**0.528**	**0.647**	**8.8**	**3.65**
RGBD	✗	0.511	0.643	8.5	3.73

Notes: The best-performing results obtained using the same vision type are emphasized in bold. This note applies to the balance of the tables in this article.

(Savva et al., 2019). The exposure time α takes a value between 0 and 1. The independent noise n_c is sampled from a zero-mean gaussian with a variance of 0.049. For simplicity, we set the illumination-dependent noise n_s to zero. The combination of α, n_s, and n_c describes different low light conditions. $M(\cdot)$ and $M^{-1}(\cdot)$ denote Bayer pattern and inverse Bayer pattern, respectively. $f(\cdot)$ is the camera response function (Grossberg and Nayar, 2004), and $f^{-1}(\cdot)$ is the inverse of $f(\cdot)$. Equation 4.1 is approximated via a pretraining paradigm following Wang et al. (2019). We use L.L.1 (best lighting) to L.L.5 (worst lighting) to denote different simulated low-light intensities. More details are in appendix E.1.

4.1 Echoer Is a Performance Booster. The echoer emits a wavelet signal with linearly adjustable frequency at every time step, enabling a more flexible perception of the external environment. In principle, echo input may be a supplement to any other modality, such as RGB, depth, and RGBD. Table 1 shows the capability of echo to supplement other modalities under L.L.1 on data set Replica. For conciseness, in the upcoming experiments, we report the results only in low-light situations where the visual input is RGB images.

4.2 On Optimal Reward Assignment. The optimal parameters (ρ, ρ^{ω}, and ρ^{ν}) of reward assignment module (see section 3.4) is searched experimentally in this section. The experiments were carried out under L.L.5 on the Replica data set. From Table 2, we observe that the robot has the best navigation ability when $\rho = 0.4$, where the learned ρ^{ω} and ρ^{ν} are, respectively, 0.69 and 0.31. This experiment also shows that a combination of fixed and trainable allocation works better than a fixed equal allocation, $\rho^{\omega} = \rho^{\nu} = 0.5$. Appendix F.2 has more details.

4.3 On Echoer Action Space: Category, Direction, and Volume. In the action space of echoer, the relative contribution of the sweep signal's

Table 2: Optimal Reward Assignment.

ρ	ρ^{ω}	ρ^{v}	SPL (↑)	SR (↑)
0.0	0.50	0.50	0.338	0.363
0.2	0.60	0.40	0.368	0.417
0.4	**0.69**	**0.31**	**0.407**	**0.474**
0.6	0.21	0.79	0.406	0.446
0.8	0.72	0.28	0.327	0.373
1.0	0.33	0.67	0.302	0.331

Table 3: The Contribution of Echo Category (C), Direction (D), and Volume (V).

C	V	D	SPL (↑)	SSPL (↑)	SR (↑)	R_{mean} (↑)	DTG (↓)	NDTG (↓)
✓	✗	✗	**0.384**	**0.537**	**0.425**	**4.8**	**4.58**	**0.437**
✗	✓	✗	0.265	0.518	0.291	3.4	4.70	0.466
✗	✗	✓	0.301	0.515	0.329	3.4	4.83	0.471
✗	✓	✓	0.346	0.506	0.377	3.8	4.93	0.480
✓	✗	✓	0.346	0.495	0.378	3.6	5.05	0.494
✓	✓	✗	0.287	0.518	0.321	3.7	4.72	0.460
✓	✓	✓	0.338	0.527	0.363	4.0	4.83	0.468

Table 4: The Impact of Different Echo Volumes.

Echo volume	SPL (↑)	SR (↑)
0.25	0.314	0.351
0.5	0.340	0.388
0.75	0.269	0.295
1.0	**0.341**	**0.389**

category (C), direction (D), and volume (V) might be different. With that in mind, we measure the performance under different combinations of C, D, and V. Table 3 illustrates the results for L.L.5 and $\rho = 0$ on Replica, where sound category (C) seems to be the main factor. The impact of scanning direction (D) has a moderate influence. The sound volume (V) has a negligible impact, coinciding with the assumption that a stronger signal is generally appreciated; this is verified in Table 4, where the echo category and direction are set to 5 ms and the same as the facing direction of the robot camera, respectively. Based on the results in Tables 3 and 4, we believe it is sufficient to fix the echo volume ($a^{v,\text{vol}} = 1.0$) while aligning the sound direction with that of the camera ($a^{v,\text{dir}} = 0$). We notice that enabling the direction and volume seems to degrade the performance. This is a consequence of a static environment in each episode, meaning there are no other moving distracting objects, and the forward-facing (i.e., same as the moving

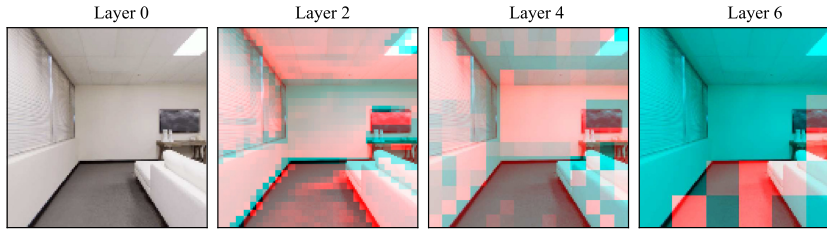


Figure 5: Visualization of the learned visual features that are overlaid on RGB images.

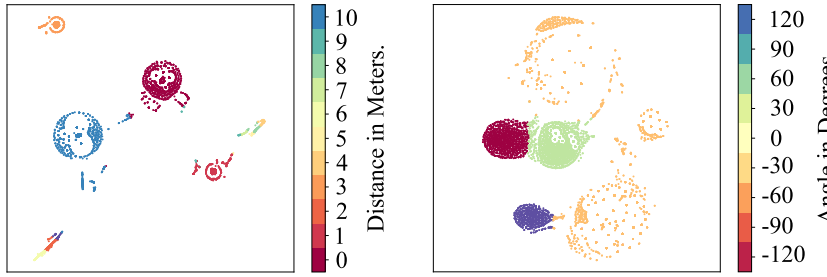


Figure 6: Visualization of the learned echo features. The dimension is reduced using TSNE. The horizontal and vertical axes are the two primary dimensions of the reduced audio features. The ground truth is represented by different colors: (Left) The discrete distance (from 0 to 10) between the agent and the target. (Right) The orientation (−120 to 120 degrees) between the agent and the target.

direction) signal tends to be the most informative. Likewise, the signal with a high volume is always preferred to ensure a stronger echo unless there is a requirement for energy saving. In theory, adding direction and volume enables the agent to explore more action possibilities, yet potentially makes policy learning more challenging and unstable.

4.4 Visualization of Learned Features. To qualitatively examine the disengagement of the learned visual and echo features, we (1) overlay the visual features from different encoder layers (see Figure 5) on top of the original RGB image and (2) apply TSNE (Flexa et al., 2021) on the learned echo features (see Figure 6). We observe from Figure 5 that the visual encoder learns to pay more attention to the walkable area, which is more evident with a deeper encoder. Figure 6 demonstrates that the learned echo features are naturally correlated with the distance and angle to the goal.

4.5 Robustness to Lighting Conditions. In Figures 7A and 7B, we compare our method with the selected baselines under different lighting

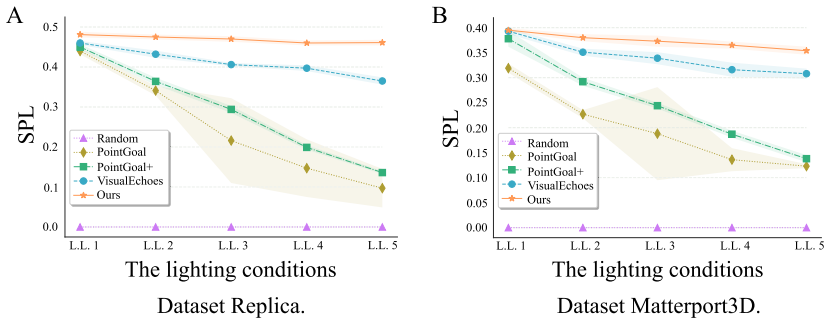


Figure 7: Navigation performance comparison to baselines under different lighting conditions.

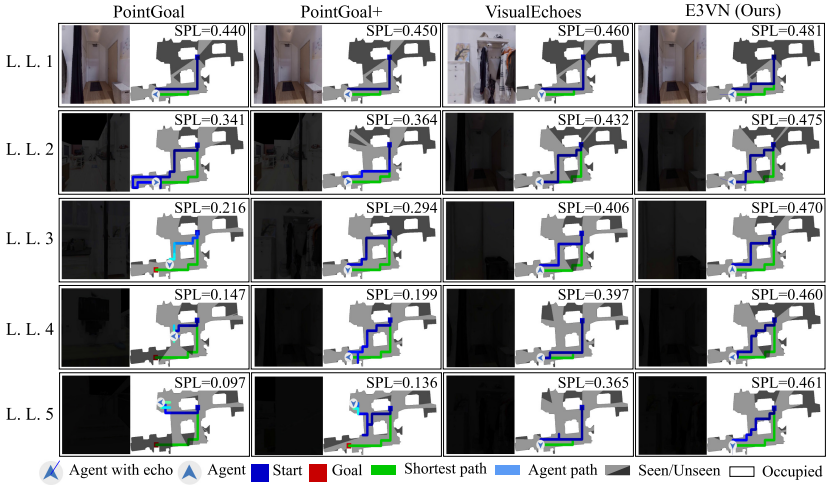


Figure 8: Demonstration of E3VN robustness: the navigation trajectories obtained by the end of one episode using different methods (columns) and lighting conditions (rows) on the Replica data set.

conditions simulated using equation 4.1. E3VN achieves the best performance under different low-light conditions, illustrating its robustness in low-light scenarios. (For more comparison using metrics other than SPL, see appendix F.) Figure 8 (for Replica) and Figure 9 (for Matterport3D) demonstrate the robot trajectories obtained using different approaches under different simulated lighting conditions. As the environment deteriorates from good to poor lighting conditions, the performance of all algorithms declines to varying degrees, while E3VN maintains a reasonable performance.

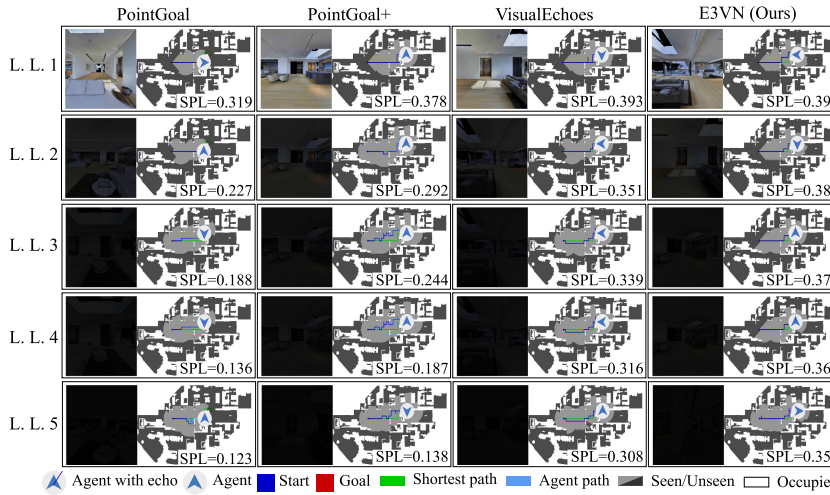


Figure 9: Demonstration of E3VN robustness: the navigation trajectories obtained by the end of one episode using different methods (columns) and lighting conditions (rows) on the Matterport3D data set.

E3VN outperforms others in normal lighting L.L.1, which coincides with the result in Table 1. The echoer emits a wavelet signal with linearly adjustable frequency at every time step, enabling a more flexible perception of the external environment. As we know, a visual signal is obtained from a passive sensor, and the emitted wavelength of the depth sensor is typically fixed. In principle, echo input can be used as an effective supplement to depth and RGB for both low light and normal lighting conditions.

4.6 On Relative Modality Impact. Because of the ever-changing environmental context and target location, we expect that the relative impact of echo and visual input on the agent's decision at a different time can vary. To quantify visual (echo) impact, we replace the visual (echo) input with random noise; the visual (echo) impact score is the absolute difference (normalized) between the logarithmic action probabilities and the semicorrupted model and the intact one. Figure 10 shows the impact scores on the egocentric robot view at different time steps. In ideal lighting conditions (the top row), the relative impact of vision and echo varies according to the surroundings. But in poor lighting conditions (the bottom row), the echo dominates the actions performed.

4.7 Ablation Study of the Learned Policies. Are the learned policies of robot (a^{ω}) and echoer (a^{ν}) better than a random strategy? To answer that

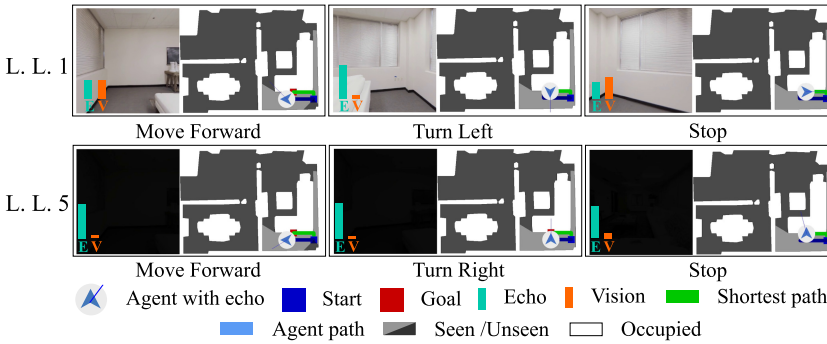


Figure 10: Relative visual and echo impact score for one episode under L.L.1 (top) and L.L.5 (bottom). Columns correspond to three sampled time steps. The green and orange bars represent the importance of echo and vision, respectively.

Table 5: Ablation Study of the Learned Robot and Echoer Policies.

Echo policy (a^v)	Robot policy (a^ω)	SPL (↑)	SSPL (↑)	SR (↑)	R_{mean} (↑)
Learned	Random	0.000	0.022	0.000	0.0
Random	Learned	0.417	0.474	0.540	6.2
Learned	Learned	**0.481**	**0.533**	**0.562**	**6.6**

question, we replace the learned policy with a random one for the trained robot and/or echoer. From Table 5, we can see that the best strategy is obtained only when both learned policies are used.

4.8 Single Agent versus Two Agents. One might argue that treating echoer and robot as a single agent might avoid the need for reward assignments with dual encoders. Notice that the echoer and the robot have different action spaces; they tend to focus on different aspects embodied in the state. As a result, the agents might act more effectively and efficiently if they can learn the state representation separately. On the other point, treating them as a single agent will lead to a joint action space $\mathbb{R}^{108}$ that is much more complex than either echoer ($\mathbb{R}^{27}$) or robot ($\mathbb{R}^4$). Therefore, the joint agent will likely increase the performance of policy learning and optimization. To validate that assumption, we carry out an additional experimental comparison between single and multiple agents. The results in Table 6 show that two-agent architecture performs much better than the single-agent counterpart from the aspect of many evaluation metrics.

Table 6: Comparison between Two Architectures (on Replica Data Set and under Perfect Vision): One Agent versus Two Agents.

Architecture	SPL (↑)	SSPL (↑)	SR (↑)	R_{mean} (↑)	DTG (↓)	NDTG (↓)
One agent	0.473	0.525	0.589	7.3	4.12	0.363
Two agents	**0.483**	**0.532**	**0.630**	**8.3**	**3.80**	**0.321**

5 Conclusion

This letter proposes a novel end-to-end point goal navigation approach, E3VN, to enhance performance when facing scenarios with poor visibility. E3VN models the robot as playing a Markov game with an echoer that actively emits sound and receives environmental echo simultaneously. The echoer can change the sound category, volume, and direction. Using both visual and echo input, the robot and echoer policies are jointly optimized by maximizing the reward obtained from the environment. Throughout the training, the overall reward is decomposed into the robot and echoer parts, which are also tuned and optimized. We conduct experiments on widely adopted navigation tasks. The performance of our approach surpasses all state-of-the-art baselines in different lighting conditions. E3VN is so robust to various visual conditions that it maintains more than 95% of the navigation performance while other methods degrade to a barely usable state.

Acknowledgments

This work is funded by Sino-German Collaborative Research Project Crossmodal Learning with identification number NSFC62061136001/DFG SFB/TRR169.

References

Anderson, P., Chang, A., Chaplot, D. S., Dosovitskiy, A., Gupta, S., Koltun, V., . . . Zamir, A. R. (2018). *On evaluation of embodied navigation agents*. arXiv:1807.06757.

Beery, S., Wu, G., Rathod, V., Votel, R., & Huang, J. (2020). Context R-CNN: Long term temporal context for per-camera object detection. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13072–13082).

Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., . . . Zhang, Y. (2017). Matterport3D: Learning from RGB-D data in indoor environments. In *Proceedings of the International Conference on 3D Vision*.

Chaplot, D. S., Gandhi, D., Gupta, S., Gupta, A., & Salakhutdinov, R. (2020). Learning to explore using active neural SLAM. In *Proceedings of the 8th International Conference on Learning Representations*.

Chen, C., Al-Halah, Z., & Grauman, K. (2021). Semantic audio-visual navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15516–15525).

Chen, C., Jain, U., Schissler, C., Gari, S. V. A., Al-Halah, Z., Ithapu, V. K., . . . Grauman, K. (2020). Soundspaces: Audio-visual navigation in 3D environments. In *Proceedings of the European Conference on Computer Vision* (pp. 17–36).

Chen, C., Majumder, S., Al-Halah, Z., Gao, R., Ramakrishnan, S. K., & Grauman, K. (2021). Learning to set waypoints for audio-visual navigation. In *Proceedings of the 9th International Conference on Learning Representations*.

Chen, K., Chen, J. K., Chuang, J., Vázquez, M., & Savarese, S. (2021). Topological planning with transformers for vision-and-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 11276–11286).

Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). *Rethinking atrous convolution for semantic image segmentation*. arXiv:1706.05587.

Christensen, J. H., Hornauer, S., & Stella, X. Y. (2020). Batvision: Learning to see 3D spatial layout with two ears. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation* (pp. 1581–1587).

Dean, V., Tulsiani, S., & Gupta, A. (2020). See, hear, explore: Curiosity via audio-visual association. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems, 33*. Curran.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., . . . Joulin, A. (2021). Beyond English-centric multilingual machine translation. *Journal of Machine Learning Research*, *22*(107), 1–48.

Flexa, C., Gomes, W. C., Moreira, I., Alves, R., & Sales, C. (2021). Polygonal coordinate system: Visualizing high-dimensional data using geometric DR, and a deterministic version of t-SNE. *Expert Syst. Appl.*, *175*, 114741. 10.1016/j.eswa.2021.114741

Gan, C., Gu, Y., Zhou, S., Schwartz, J., Alter, S., Traer, J., . . . Torralba, A. (2022). Finding fallen objects via asynchronous audio-visual integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10523–10533).

Gan, C., Zhang, Y., Wu, J., Gong, B., & Tenenbaum, J. B. (2020). Look, listen, and act: Towards audio-visual embodied navigation. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation* (pp. 9701–9707).

Gao, R., Chen, C., Al-Halah, Z., Schissler, C., & Grauman, K. (2020). VisualEchoes: Spatial image representation learning through echolocation. In *Proceedings of the 16th European ECCV Conference*, Lecture Notes in Computer Science 12354 (pp. 658–676).

Gordon, D., Kadian, A., Parikh, D., Hoffman, J., & Batra, D. (2019). SplitNet: Sim2Sim and Task2Task transfer for embodied visual navigation. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision* (pp. 1022–1031).

Grossberg, M. D., & Nayar, S. K. (2004). Modeling the space of camera response functions. *IEEE Trans. Pattern Anal. Mach. Intell.*, *26*(10), 1272–1282. 10.1109/TPAMI.2004.88, PubMed: 15641715

Gupta, S., Davidson, J., Levine, S., Sukthankar, R., & Malik, J. (2017). Cognitive mapping and planning for visual navigation. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7272–7281).

Hong, Y., Wu, Q., Qi, Y., Opazo, C. R., & Gould, S. (2021). VLN BERT: A recurrent vision-and-language BERT for navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1643–1653).

Irshad, M. Z., Ma, C., & Kira, Z. (2021). Hierarchical cross-modal agent for robotics vision-and-language navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation* (pp. 13238–13246).

Karkus, P., Cai, S., & Hsu, D. (2021). Differentiable SLAM-net: Learning particle SLAM for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2815–2825).

Kurita, S., & Cho, K. (2021). Generative language-grounded policy in vision-and-language navigation with Bayes' rule. In *Proceedings of the 9th International Conference on Learning Representations*.

Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A., Banino, A., . . . Hadsell, R. (2017). Learning to navigate in complex environments. In *Proceedings of the 5th International Conference on Learning Representations*.

Morad, S. D., Mecca, R., Poudel, R. P. K., Liwicki, S., & Cipolla, R. (2021). Embodied visual navigation with automatic curriculum learning in real environments. *IEEE Robotics Autom. Lett.*, *6*(2), 683–690. 10.1109/LRA.2020.3048662

Purushwalkam, S., Gari, S. V. A., Ithapu, V. K., Schissler, C., Robinson, P., Gupta, A., & Grauman, K. (2021). Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1183–1192).

Qin, L., Li, Z., Che, W., Ni, M., & Liu, T. (2021). Co-GAT: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence* (pp. 13709–13717).

Ramakrishnan, S. K., Al-Halah, Z., & Grauman, K. (2020). Occupancy anticipation for efficient exploration and navigation. In *Proceedings of the 16th European Conference on Computer Vision*, Lecture Notes in Computer Science 12350 (pp. 400–418).

Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J. N., & Whiteson, S. (2018). QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning* (pp. 4292–4301).

Savva, M., Malik, J., Parikh, D., Batra, D., Kadian, A., Maksymets, O., . . . Koltun, V. (2019). Habitat: A platform for embodied AI research. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision* (pp. 9338–9346).

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. arXiv:1707.06347.

Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J. J., . . . Newcombe, R. (2019). *The replica dataset: A digital replica of indoor spaces*. arXiv: 1906.05797.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V. F., Jaderberg, M., . . . Graepel, T. (2018). Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 2085–2087).

Teng, X., Guo, D., Guo, Y., Zhou, X., & Liu, Z. (2019). CloudNavi: Toward ubiquitous indoor navigation service with 3D point clouds. *ACM Transactions on Sensor Networks*, *15*(1), 1–28. 10.1145/3216722

Tracy, E., & Kottege, N. (2021). CatChatter: Acoustic perception for mobile robots. *IEEE Robotics and Automation Letters*, *6*(4),7209–7216.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. In I. Guyon, Y. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems*, *30* (pp. 5998–6008). Curran.

Wang, H., Wang, W., Liang, W., Xiong, C., & Shen, J. (2021). Structured scene memory for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8455–8464).

Wang, Y., Cao, Y., Zha, Z., Zhang, J., Xiong, Z., Zhang, W., & Wu, F. (2019). Progressive Retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement. In *Proceedings of the 27th ACM International Conference on Multimedia*.

Wijmans, E., Kadian, A., Morcos, A., Lee, S., Essa, I., Parikh, D., . . . Batra, D. (2020). DD-PPO: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *Proceedings of the 8th International Conference on Learning Representations*.

Ye, J., Batra, D., Wijmans, E., & Das, A. (2020). *Auxiliary tasks speed up learning point-goal navigation*. arXiv:2007.04561.

Yu, Y., Cao, L., Sun, F., Liu, X., & Wang, L. (2022). *Pay self-attention to audio-visual navigation.* arXiv:2210.01353.

Yu, Y., Huang, W., Sun, F., Chen, C., Wang, Y., & Liu, X. (2022). Sound adversarial audio-visual navigation. In *Proceedings of Tenth International Conference on Learning Representations*.

Received October 22, 2022; accepted January 2, 2023.