AV-GeN: Audio-Visual Navigation Framework with Generalisable Audio Representations

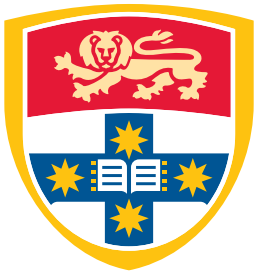
SHUNQI MAO
SID: 470079780

Supervisor: A/Prof. Weidong (Tom) Cai

This thesis is submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Computer Science and Technology (Advanced) (Honours)

School of Computer Science
The University of Sydney
Australia

29 May 2022



THE UNIVERSITY OF SYDNEY

Student Plagiarism: Compliance Statement

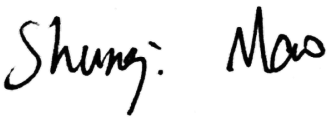
I certify that:

I have read and understood the University of Sydney Student Plagiarism: Coursework Policy and Procedure;

I understand that failure to comply with the Student Plagiarism: Coursework Policy and Procedure can lead to the University commencing proceedings against me for potential student misconduct under Chapter 8 of the University of Sydney By-Law 1999 (as amended);

This Work is substantially my own, and to the extent that any part of this Work is not my own I have indicated that it is not my own by Acknowledging the Source of that part or those parts of the Work.

Name: Shunqi Mao

Signature:  **Date**: 21 May 2022

Abstract

Imagine you lost your phone at your friend's house. You can hear it ringing but cannot see it, you need to explore the unfamiliar environment to find its position. This scenario is formalised as a task called "Audio-Visual Navigation" (AVN), where individuals need to navigate to a constantly sounding object based on current vision and auditory senses. With the popularisation of embodied AI, researchers believe that developing intelligent agents capable of completing the AVN task is significant for developing general-purpose AI systems and implies a variety of applications. While several deep-learning-based methods have been proposed for more efficient path planning and obstacle avoidance, navigation with unfamiliar sounds remains a critical challenge. None have explored the approaches of how models can effectively learn audio representations that generalise to various unheard sound classes.

In this thesis, a novel Audio Visual Generalisable Navigation (AV-GeN) framework is proposed to overcome the generalisation problem. A contrastive learning-based method is designed to regularise the audio encoder, where the sound-agnostic latent representations with goal-driven features can be learnt from the audio signals of various classes. In addition, we propose that sound signals treated as the navigation target can be augmented to reduce overfitting, and we designed a sound augmentation pipeline with two augmentation strategies to enrich the training sounds.

We evaluate our proposed AV-GeN framework on the Matterport3D dataset with the SoundSpaces simulator. The experiment results demonstrate that the proposed framework achieves significantly better performance compared to the state-of-the-art AVN frameworks for navigating toward unfamiliar audio goals. Moreover, the AV-GeN framework has been submitted to the SoundSpaces Challenge, the most popular AVN challenge, and achieved the top-1 performance on the public leaderboard. Our method has been summarised in a paper and submitted to the CVPR Embodied AI 2022 Workshop.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Associate Professor Weidong (Tom) Cai, who has supported me with detailed guidance and enthusiastic encouragement in my research study. I would also like to recognize the invaluable assistance of my mentor Chaoyi Zhang, who has been motivating and guiding me with insightful opinions and great help. I would also like to express my gratitude to Heng Wang for her help with this project. In addition, I wish to thank my peer research students. Communicating with them makes me feel like being of a research community. They keep me on track with the cutting edge topics and prompt me to work harder.

Finally, I appreciate my parents for always supporting and encouraging me throughout my daily life both emotionally and economically. Without their support, it would not have been possible for me to struggle through the difficulties encountered during the honours project.

CONTENTS

List of Figures

List of Tables

CHAPTER 1

Introduction

1.1 Embodied AI

The evolutionary breakthroughs in deep learning have brought great success to various AI fields such as computer vision and natural language processing. Benefiting from the images, videos, and text data collected over the internet, critical progress has been made in building powerful AI models that analyse the randomised internet data. However, these AI systems are usually defined with fixed inputs and outputs, and they do not map the ways how human learns from sequential experience obtained by moving, seeing, and interacting with the environment.

With the recent development in learning-based AI algorithms, researchers began to shift their attention from "Internet AI" that learns from data collected over the internet, towards "embodied AI" where the intelligent agent with virtual embodiment learns to act by interacting with the environment [24]. Specifically, embodied AI is defined in recent research as intelligent agents that operate and learn in realistic 3D simulations of environments to complete specific tasks, based on constant feedback loops made up of the planned actions from the agent and received egocentric perceptions from the environment. Overall, the embodiment of AI has been widely considered a necessary goal for building machines with "true intelligence" [58], as they learn in the same way that humans learn.

Due to the importance of embodied AI, a set of tasks such as navigation, exploration, and embodied question answering have been designed to facilitate the development of embodied agents. These tasks assess the comprehensive capability of the embodied agents to accomplish them with human-level competence. Currently, research on embodied AI majorly focuses on applying the traditional AI concepts, including language processing, reasoning, and vision, to solve these embodied tasks where intelligent agents interact with the environment to complete complex tasks. Taking the embodied question answering task as an example, an agent might be spawned at a random position in a house, and given the

question "What colour is the car in the garage?". The agent will have to *interpret* the question, *explore* the environment, *move* to the garage, *reason* with the visual observations, and finally produce a *language* output. With the successful development of embodied AI algorithms, it can be a prospect that the robotics applications equipped with embodied AI models can be developed to help humans with critical tasks in various aspects. In this thesis, we will be focusing on an interesting and crucial embodied AI task - audio-visual navigation.

1.2 Audio-Visual Navigation

1.2.1 Importance of Audio-Visual Navigation

Among embodied AI tasks, *navigation* is of particular importance, with applications in various scenarios such as search, rescue and service robotics [12]. The embodied agent needs to navigate to the goal positions in 3D environments defined as a coordinate, an object, or specific areas. Much relevant research focused on leveraging visual cues so that the agents can form spatial understandings of the environment. However, it should be noticed that humans have the remarkable ability to address the correspondences between different signal modalities. Therefore, mobile robots with human-level competency should also be capable of effectively exploiting the multi-modality signals in the environment.

Among the human senses, vision and acoustic sensory signals are better interpreted in computer science compared to other sensories, including taste, smell and tactile impression. The recent proposed audio-visual navigation (AVN) problem [12] defines a multi-modality learning navigation task bridging the two signal modalities, where the embodied agent needs to navigate to the position of a constantly sound-making object providing audio and visual sensory inputs, as shown in Figure 1.1 (the sound-making object will be referred to as the 'audio goal'). The agent is required to interpret the received sound signals at each step, and decide the appropriate action for the next move. Compared to the object-goal navigation tasks where the agent searches for target objects in the environment based only on vision perceptions, AVN defines agents with significantly higher intelligence. Over the past few years, several methods have been proposed for the efficient completion of AVN with preciously designed room-geometry mapping and path planning. Chen *et al*. proposed a straightforward framework 'AV-NAV' [12], for the navigation task by generating step-by-step actions from multi-modality inputs. In [29], the authors decomposed AVN into predicting the relative position of the sound source and navigating to the target position given

visual inputs. The AV-WaN [13] further enabled agents to navigate through intermediate waypoints with an occupancy map. These methods have established meaningful frontiers for the development of AVN.

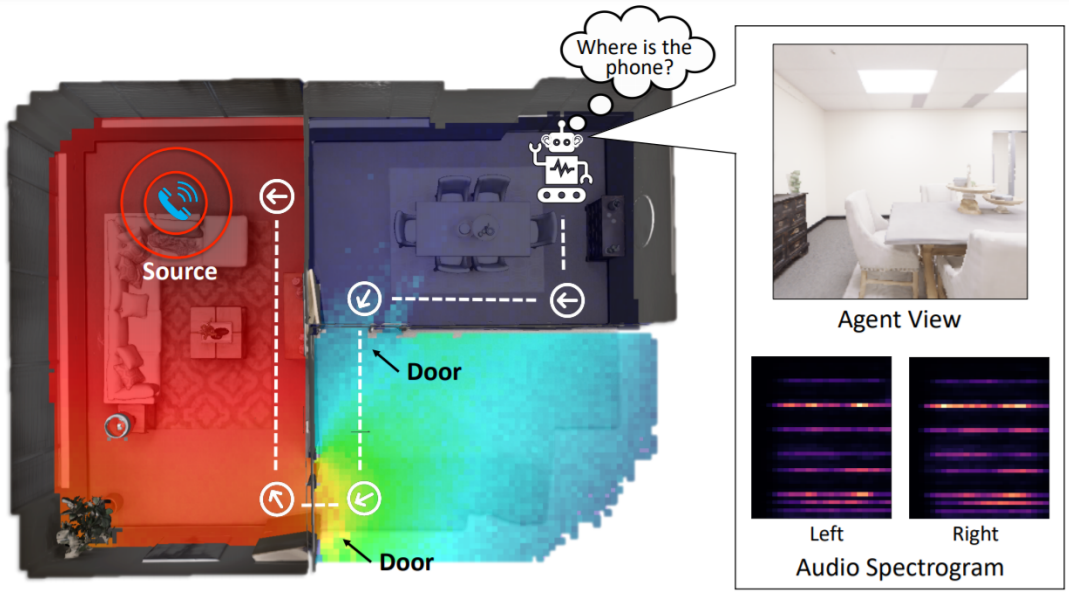


FIGURE 1.1. The graphical illustration of the audio-visual navigation (AVN) problem, the agent needs to navigate to the position of the sounding object in indoor environments. Image adapted from [12].

1.2.2 Challenges of Audio-Visual Navigation

While the existing methods [12, 29, 13] attempt to build AVN frameworks with delicate path plannings, a substantial gap lies in the navigation performance between receiving familiar sounds and unheard sounds as audio goals. For example, the agent trained with telephone sound as navigation goals could navigate to the learned telephone sounds positions efficiently, but it would still be challenging for the agent to navigate to other sound classes, such as people speaking. Existing works typically suffer from a performance decrease of 50% when generalising to various unheard sounds. For example, the state-of-the-art AVN framework proposed by Chen et al. achieves 72.3% SPL (a quantitative measure) when training and evaluating with one identical audio goal, but only achieves 36.2% SPL on disjoint set training and testing sounds. Such a performance gap indicates that current frameworks severely overfit the training sounds, where the agent learns to interpret the audio goals by memorising specific training sounds, instead of exploiting meaningful patterns in the audio signals. Therefore, learning effective position-oriented audio representation from limited types of training sound remains a critical challenge.

1.3 Contributions

Modern AVN frameworks are easily subject to specific sounds learned during training and generalise poorly to unheard sounds. In this thesis, we designed the Audio Visual Generalisable Navigation (AV-GeN) framework, which significantly outperforms existing AVN methods for navigating toward unfamiliar audio goals.

To reduce the generalisation errors, We regularise the audio network to learn sound-agnostic features by a novel Audio Feature Similarity Optimisation (AFSO) method, which maximises the feature similarity between pairs of audio observations that imply equivalent source-receiver spatial relationships but with distinct types of sound. Meanwhile, AFSO minimises the feature similarity between audio observation pairs that imply different relative goal position information. The overfitting issue is thus alleviated with our contrastive learning formulation of the similarity optimisation, as the learned representation focuses more on the audio-goal position features instead of memorising specific sounds.

Moreover, we propose a set of novel techniques to augment the input sounds. Different from existing audio augmentation methods where related labels need to be modified or preserved according to the augmentation method, the modification of original sounds in our setting hardly affects the audio-goal information. Therefore, we design several unique strategies to augment sound sources that significantly enrich the sound distributions and avoid overfitting.

We examine the effectiveness of our framework by extensive evaluations on Matterport3D [9], a large scale dataset of 3D indoor environments. We demonstrate that our AV-GeN framework significantly outperforms the existing SOTA method by 12% in SPL, indicating that our proposed methods can effectively improve the generalisation of AVN with unfamiliar audio goals.

We have submitted part of our methods to the SoundSpaces Challenge 2022, a public challenge organised by the proposer team of the AVN task. By the time of the thesis submission, our AV-GeN framework achieved the top-rank performance on the public leaderboard. Moreover, part of this work has been submitted to the CVPR Embodied AI 2022 Workshop. The challenge and workshop information can be found at the official websites: `https://soundspaces.org/challenge` and `https://embodied-ai.org/`.

To summarise, our contributions are three-fold:

- We propose the Audio Feature Similarity Optimisation (AFSO) method, which significantly improves the generalisation of learning audio representations.
- We propose the Source Sound Augmentation method, which effectively eliminates overfitting sounds heard during training.
- We designed a novel AVN framework, AV-GeN, based on the proposed methods and a baseline framework AV-WaN [13]. We demonstrate that our framework is superior to existing baselines, and it achieves state-of-the-art performance on the SoundSpaces AVN challenge.

1.4 Thesis Outline

To begin, we introduce the relevant background on research related to embodied AI, including literature on embodied AI simulators, tasks, and reinforcement learning methods. We then cover critical works on navigation, which are most closely related to AVN, followed by the close literature on AVN (Chapter 2).

In Chapter 3, we first show an overview of the architecture of the proposed AV-GeN framework. Afterwards, we explain each component in the framework, including environment, input-processing, visual and auditory perception models, and path planning. We detailedly explain the novel AFSO and sound augmentation methods we designed as our main contributions.

In Chapter 4, we demonstrate that the proposed AV-GeN framework significantly outperforms the existing methods through experiments and results. In addition, we analyse the experiment findings and discuss the current limitations with future refinements of the framework in different aspects in Chapter 5. Finally, We finish this thesis by summarising the crucial findings (Chapter 6).

CHAPTER 2

Literature Review

Audio-visual navigation stands as a sub-topic in embodied AI, which combines the navigation problem with audio-visual cross-modality learning. In this chapter, we review the broad literature related to audio-visual navigation. To begin with, we briefly introduce the critical developments in embodied AI and reinforcement learning, followed by various research on in-door navigation with different task settings. Meanwhile, we present several insightful problems and ideas proposed to bridge the visual and audio modalities. Finally, we discuss in detail the existing methods for the audio-visual navigation problem.

2.1 Embodied AI in 3D Environments

Unlike mainstream AI problems that focus on learning from labelled data collected over the internet, embodied AI enables intelligent agents to learn through interacting with the nearby environments [24]. Here we will first introduce the development of embodied AI simulators that enables higher levels of interactions with the environment. Then we shortly present the major embodied AI tasks.

2.1.1 Embodied AI Simulators

The earliest large-scale environment simulator was the DeepMind Lab proposed by Beattie et al. [7]. It was designed for trivial navigation tasks, thus the only interaction the agent can make is to move in the environment. Following DeepMind, several simulators are proposed with more flexible scalability on the objects in the environment, such as AI2-THOR [45], CHALET [78], VirtualHome [59], and VRKitchen [31]. In addition to movement, these simulators allow for a higher level of interactions with the object, where the agent can interact with specific objects and change their states. However, to enable flexible modifications of the objects in the scene, these simulators only provide game-based scene constructions with insufficient realism. To train intelligent agents capable of acting in real-life environments, Savva

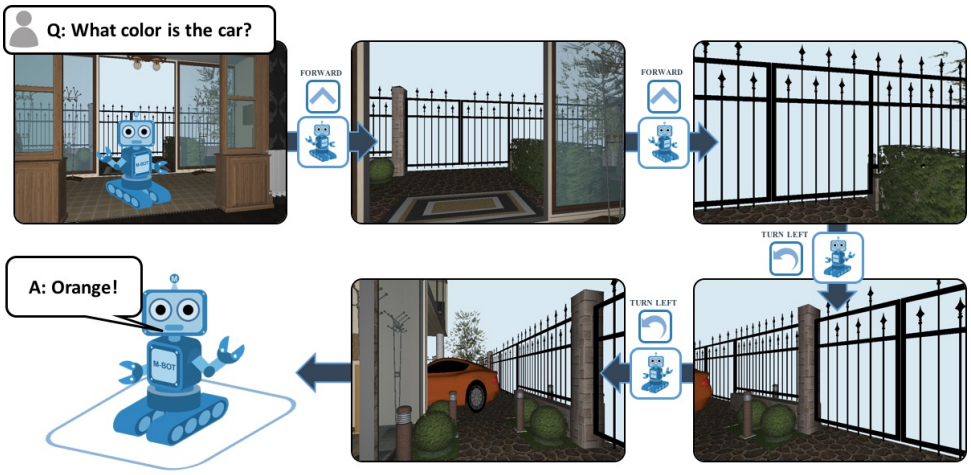
et al. [64] proposed Habitat-Sim, where the environment was constructed based on scans of real-life indoor scenes such as offices and apartments. Li et al. [49] further combined the interactability and realism of the previous simulators and proposed the iGibson simulator, where the agent can not only move and see but also perform rudimentary physic interactions with highly-realistic objects. On the other hand, the recently proposed ThreeDWorld simulator [28] focuses more on complex physics capabilities. Unlike previous simulators, where interactions were typically defined as object displacement caused by collision, or state changes such as turning off the lights, ThreeDWorld focuses on realistic physical interactions for various material types including cloths, liquid, and deformable objects [28].

In this thesis, we will be working with the SoundSpaces simulator [12], which is an adaption of the Habitat simulator that allows for flexible audio rendering for specific source-receiver position pairs. The agent can learn from realistic visual and audio signals that enable generalisation on real-world applications.

2.1.2 Embodied AI Tasks

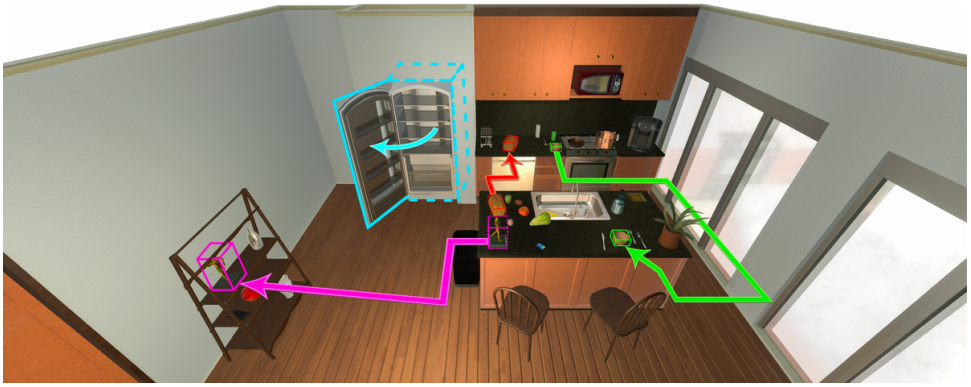
With the advancement of the simulators, several embodied AI tasks are proposed towards the goal of developing more powerful AIs. The most popular embodied AI task is navigation, where an agent navigates in the 3D environments to specific goals. Besides the simple position-goal navigation task where the navigation goal is defined as coordinates, many variants of the navigation problem are proposed in the literature. For instance, object navigation [10] requires the agent to search for specific object categories in the environment, and the semantic navigation task requires the agent to act according to the given natural language instructions such as ‘move through the door’. Moreover, these tasks are sometimes combined to train agents capable of dealing with more complex task settings. For example, in semantic-object-navigation, the agent needs to find the target objects with the help of natural language instructions.

Another fundamental embodied AI task similar to navigation is visual exploration [57]. In visual exploration, the agent needs to explore the novel environment as efficiently as possible. The performance of the exploration agent is usually measured by the number of different objects visited by the intelligent agent. Quite often, the visual exploration problem is combined with a downstream navigation task such that the agent first explores the environment and then starts to complete the navigation task. With such settings, the effectiveness of the exploration can be measured by the navigation performance. Noticeably, A recent study has extended the exploration task to the audio-visual setting [22].



(A) Embodied Question Answering [21].

(B) FurnMove [40].



(C) Rearrangement [74].

FIGURE 2.1. Examples of embodied AI tasks in the literature. The visualisations of diverse navigation tasks will be demonstrated later in Section 2.2. Images adapted from [21, 40, 74].

A more complex embodied AI task is embodied question answering [21]. It integrates a wide range of AI capabilities including visual recognition, natural language processing, question answering, reasoning, planning, and navigation [24]. As shown in Figure 2.1a, the agent will be given problems such as 'What colour is the car?', and it needs to explore or interact with the environment efficiently to make an answer.

In addition, it is recently proposed [28, 24, 6] that AI should learn the physics phenomenons in embodied environments. Compared to previous physics prediction models trained with collected video frames, AI systems should instead learn the physical outcome through self-driven interactions such as touching the objects with their arms. While little work exists on this problem, we consider it a cutting-edge topic in embodied AI that remains to be explored.

Finally, several novel and challenging problems are recently proposed in the embodied AI literature. For example, in the two-stage embodied re-arrangement [74] depicted in Figure 2.1c, the agent explores and memorises the environment in the first stage, then certain objects in the environment are re-arranged to different positions, and the agent needs to locate all re-arranged objects in the second stage. In FurnMove [40, 41] shown in Figure 2.1b, multiple agents are required to collaboratively move furniture to target positions. These problems incorporate the concept of exploration and navigation with more advanced abilities such as cooperation and scene understanding, to build agents with higher levels of intelligence. However, these tasks are less studied in the literature for their novelty and complexity.

2.1.3 Reinforcement Learning

Reinforcement learning (RL) is a machine learning training method that serves as a critical foundation for embodied AI. It enables the intelligent agent to be trained to take appropriate actions based on perceptions of the environment. Therefore, diverse RL methods are generally used in all embodied AI research. Unlike supervised learning where explicit labels are given as supervision, RL trains the agent to maximise the manually defined cumulative reward. Early RL research focused on estimating the rewards on discrete states through modelling the Markov decision process. To enable the reward approximation on continuous states and actions, the function approximation method is introduced where a function is learned to estimate the reward regarding input states.

With the evolutionary breakthroughs in deep learning (DL), the deep Q-learning network (DQN) [53] was proposed where the authors used a deep neural network for function approximation on the discrete input and action space of the Atari game. Silver et al. [67] proposed the deterministic policy gradient

(DPG) algorithm, which can estimate rewards for both continuous states and actions. Lillicrap et al. [50] further combined DQN and DPG, and proposed the deep deterministic policy gradient (DDPG), as shown in Figure 2.2. The DPG, known as the actor, was used to plan for the actions, given states. Meanwhile, the critic head is trained to estimate the reward of states, then compute back-propagatable values to train the actor with gradient ascent. In this way, RL rewards the desired behaviours and punishes undesired ones in a learnable end-to-end style. In 2017, Schulman et al. [66] proposed proximal policy optimisation (PPO), which improves the actor-critic RL strategy with an update constraint that ensures the updated states lie within the trust region [65] in a data-efficient style.

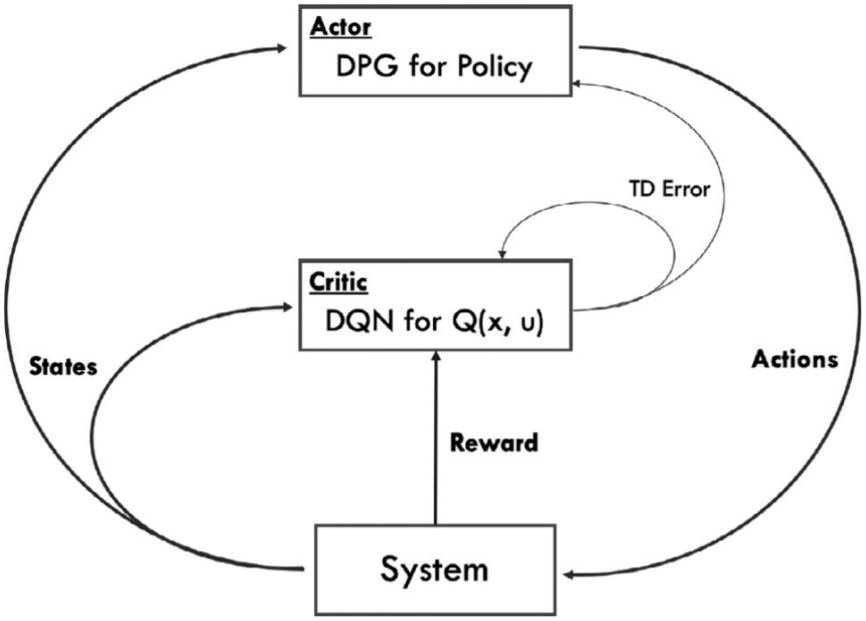


FIGURE 2.2. The architecture of the deep deterministic policy gradient algorithm. Image adapted from [54].

As our method does not involve modifying the RL algorithm, more RL literature will not be discussed in-depth. We directly adapt the PPO algorithm following [12] to train the AVN agent.

2.2 Goal-Oriented Navigation

Goal-Oriented Navigation is one of the most cutting-edge topics in the field of embodied AI. It examines the intelligent agents on exploring, moving and planning, which are the most crucial abilities of embodied AI. The agents are required to navigate to goal positions based on their sensory inputs, including but not limited to visual perceptions, where the goals are also defined differently to examine different

intelligent aspects of the agent. By starting from a random position and stopping at the target position, the agent successfully finishes a *navigation episode*. With the advancement of embodied AI simulators, researchers start to explore end-to-end learning strategies to train navigation models that can leverage visual and other cues on unseen environments [82, 33, 34]. While different navigation tasks exist in the literature, modern navigation frameworks typically follow the encoder-policy structure, where the observations at each step are encoded with deep neural networks, and the learned feature is fed to the policy network for action planning. In this section, we explain various navigation tasks in the literature and discuss in detail the research directions for improving such navigation frameworks from two perspectives, navigation map constructions and navigation memory mechanisms.

2.2.1 Various Navigation Tasks

In this section, we introduce the existing goal-oriented navigation tasks with different goal definitions and their variants. Except for the erratic goal-indicative audio signal received at each step, AVN is just similar to the other navigation tasks, where the agent needs to approach the goal position while avoiding hitting obstacles. Therefore, it is crucial to learn the cross-domain navigation tasks and methods, as their approaches could be very inspiring for developing an AVN framework.

To start with, the most basic navigation task is point-goal navigation [1] shown in Figure 2.3a, where the goal position is defined with a displacement vector. The agent only needs to interpret the visual observations effectively and avoid hitting obstacles to be competent in the task. Chaplot et al. [10] soon proposed object-goal navigation to further challenge the agent in terms of visual object understanding. As presented in the top-left corner of Figure 2.3b, the goal is defined as objects of a specific semantic class, and the agents are required to locate one of the semantic objects in the environment by navigating to it. The agent not only needs to understand the semantic meaning of the visual inputs but also reasoning on where the target object is more likely to present to complete the task more efficiently. Image-goal navigation [83] demonstrated in Figure 2.3c slightly increases the difficulty of object-goal navigation. In detail, the navigation goal is given as an input image, and the agent is required to navigate to the position where a visionary scene same as the input image can be observed by the agent. Compared to object-goal, image-goal navigation agents need to interpret the complex target scene rather than a simple object class.

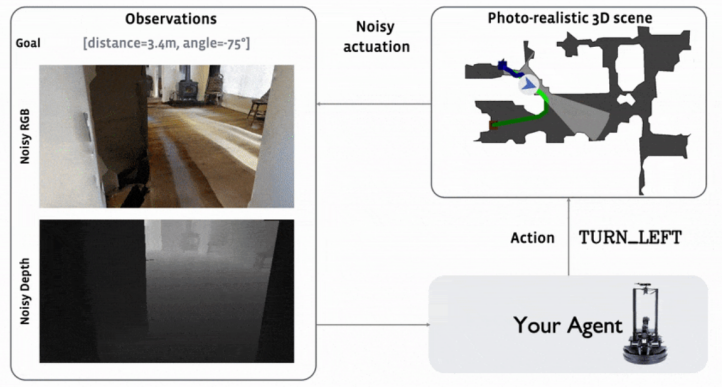
While the above tasks involve only visionary components and could be summarised as visual navigation tasks, the vision-language navigation (VLN) [3] task introduces a new modality, the natural language,

as the key component of the navigation task. As shown in Figure 2.3d, the navigation goal is defined as language-based instructions that keep changing during the navigation process. In contrast to previous visual navigation tasks, the VLN agents need to interpret the language-based goal indicator at each step to plan for the next steps. The additional language modality and dynamic goal-indicator have made the VLN task one of the most challenging navigation tasks in the literature.

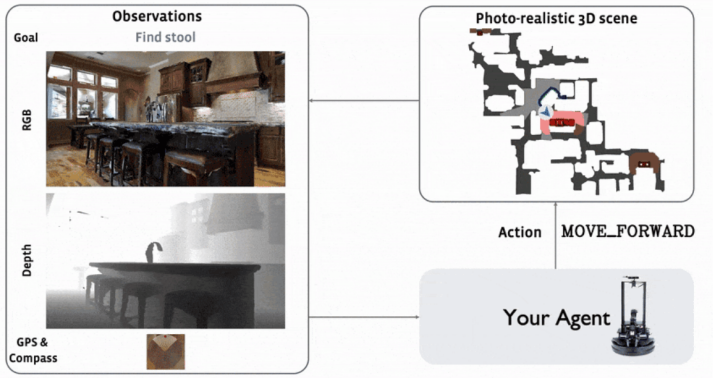
In 2020, Chen et al. [12] proposed audio-visual navigation (AVN), where the goal is defined as the position of a constantly-sounding object. The agent will receive both visual and auditory observations at each step and navigate to the position where the sound is emitted. Compared to VLN, the audio signals used as goal indicators are a stronger and more natural sensory modality. Meanwhile, the erratic goal indicator in AVN varies at each step with a fine scale (i.e., the goal instructions in VLN might be identical in certain regions of the environment, but the audio signal indicating the goal positions will be different at each navigable position). These characteristics have made the AVN task more complex than VLN and produced embodied agents with a level of intelligence more similar to humans. Moreover, semantic audio-visual navigation (SAVN) [11] combines AVN with object-goal navigation, where the sounding positions will be limited to locations with semantic implications corresponding to the type of the source sound (e.g., the emitting positions of the water flushing sound will only be in the toilet or around a tap). Meanwhile, the sound signals will not be emitted at each step but only observed occasionally, so the SAVN agents have to memorise the audio signals for long-term path planning. The graphical illustrations of the AVN and SAVN tasks are demonstrated in Figure 2.3e and Figure 2.3f respectively. It can be observed that the AVN tasks are similar to the VLN task in the sense that the goal definitions of both tasks vary depending on the position of the agent. Therefore, the VLN frameworks could be valuable references for developing novel AVN frameworks

To the best of our knowledge, the above tasks have included all possible goal definitions in the field of goal-oriented embodied navigation. Additionally, some variants of the navigation tasks have been proposed that modify the environments based on existing tasks. We do not regard these tasks as novel goal-oriented navigation tasks, as they do not alter the definition of the navigation goal, and the agent can still be capable of completing the original task with potentially better performance. A list of such navigation variants is presented in Figure 2.4.

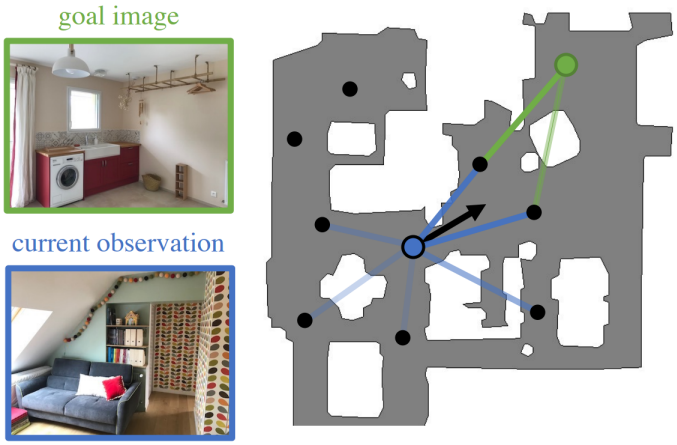
To begin with, interactive point-goal navigation [77] is a simple variant of point-goal navigation, where the agent is allowed, or even encouraged to collide and interact with light objects in the environment. The agent can then navigate more efficiently by pushing away trivial obstacles along the path. On the



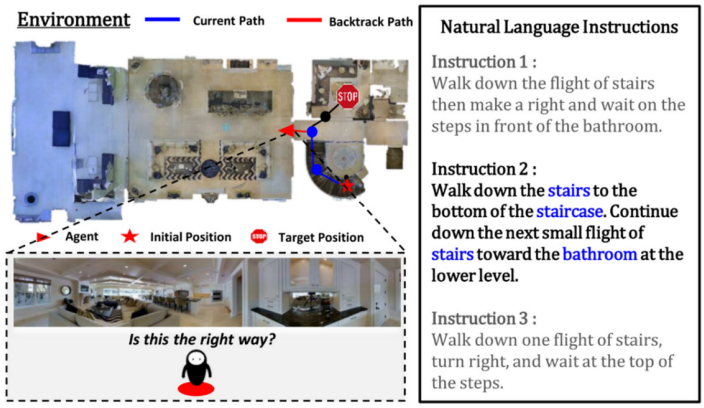
(A) Point Goal Navigation [1].



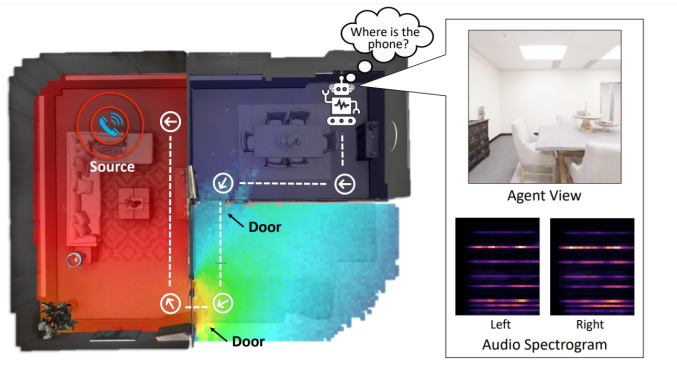
(B) Object Goal Navigation [10].



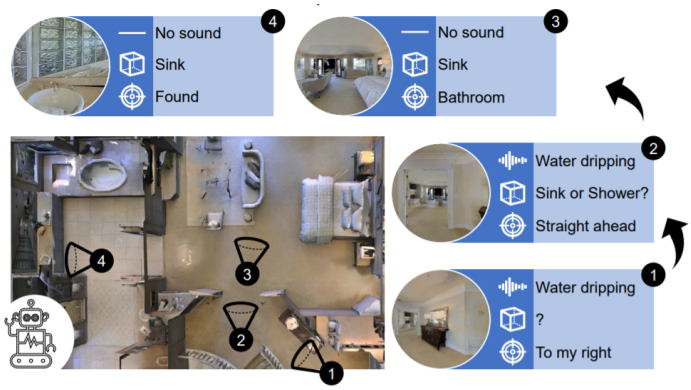
(C) Image Goal Navigation [83].



(D) Vision-Language Navigation [3].



(E) Audio-Visual Navigation [12].



(F) Semantic Audio Visual Navigation [11].

FIGURE 2.3. Various goal-oriented navigation tasks in the literature. Images adapted from [64, 52, 39, 12, 11].

other hand, the social point-goal navigation [61] demands a stronger obstacle avoidance capacity, where multiple pedestrians will be moving around in the environment. The embodied agent had to deliberately avoid collisions or proximity to pedestrians to complete the navigation task successfully. These tasks not only formulate more realistic navigation assumptions, but also implicitly improve the diversity of visually presented scenes by introducing movable objects in the environment.

An interesting variant of the VLN task, the vision-dialog navigation (VDN) [69], is depicted in Figure 2.4c. Different from the original VLN task where the agent accepts the language instructions passively from the environment, the VDN agents are required to communicate with the environment by generating language-based sentences to retrieve goal-relevant information. It facilitates the language-processing ability by both interpreting the instructions and generating responses.

Recently, sound adversarial audio-visual navigation [80] (SA-AVN) has been proposed as a variant of the AVN task. As shown in Figure 2.4d, in addition to the agent that learns to approach the audio goal, the authors train a distracting agent responsible for moving and the environment and playing distracting sounds to attack the navigation agent. The authors find that the AVN agent trained with the adversarial distracting attacks is more capable of navigation with the interference of distracting sounds, which will be a common factor in real-world scenarios.

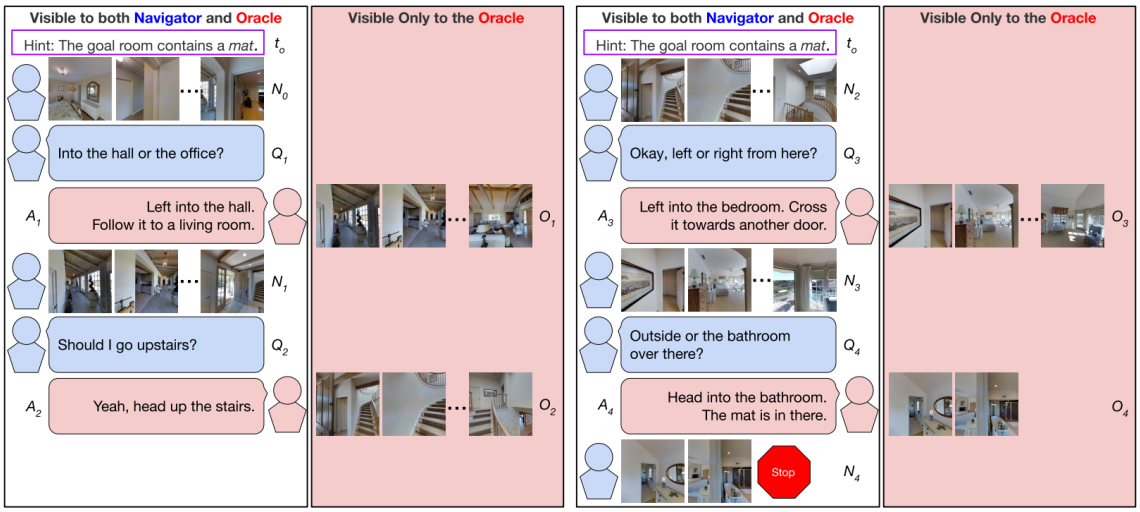
2.2.2 Navigation Map Construction

In this section, we discuss the methods related to navigation map construction, which is an important branch of research on improving goal-oriented navigation. Many researchers have tried to map the explored environment to facilitate navigation [36]. An early attempt at environment mapping on visual navigation was proposed by [63], the authors built a semi-parametric topological map on the connectivity of the rooms based on the semantic labelling of the rooms. Before navigation, the agent needs to explore the rooms at the beginning to learn the room connectivity and navigable pathways. Specifically, the topological maps must be constructed first and used as a prior input to the navigation, making the proposed framework less generalisable and efficient.

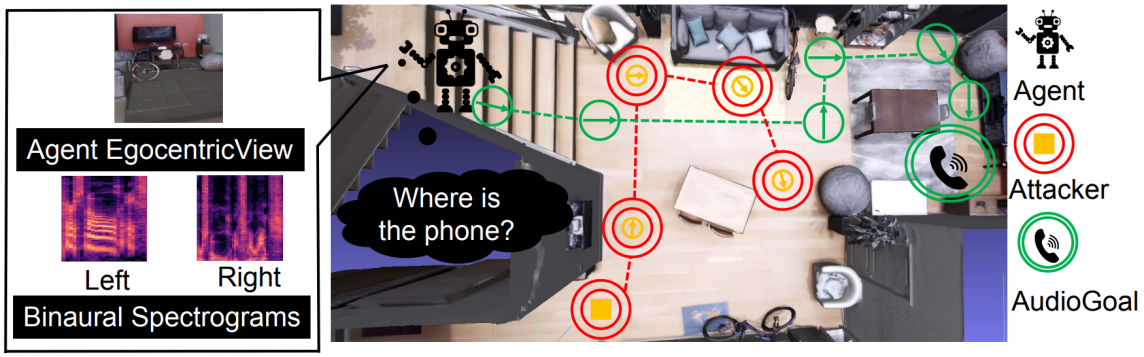
To tackle the issue that the graph mappings cannot be constructed dynamically during navigation, Wu et al. [75] developed a more flexible framework for object goal navigation. Instead of pre-compute a topological map with rooms as the node, the proposed model predicts semantic concepts on scene series on the fly. The probability of each edge on the relational graph is updated at each step to construct the

(A) Interactive Point Goal Navigation [77].

(B) Social Point Goal Navigation [61].



(C) Vision-Dialog Navigation [69].



(D) Sound Adversarial Audio-Visual Navigation [80].

FIGURE 2.4. Variants of standard navigation tasks. Images adapted from [49, 69, 80].

new semantic topological map. An LSTM [37] is used to memorise past decisions and derive a subgoal for navigation. However, the agent still needs to fully explore the unseen scenes in the environment during navigation for high-confidence room orientations.

Instead of room-based topological maps, some methods tried to construct observation-based maps stepwise with graphs. Chen et al. [15] built an observation map with graphs for the VLN problem, in which the agent needs to execute the instructions as given by a natural language sentence, as shown in Figure 2.3d. Exploring the environment is thus less crucial, as the agent can receive explicit commands such as "turn left" as extra guidance. The authors encoded the step observation with a latent vector as a node in the graph. Meanwhile, the categories and length of the nodes are determined by the ground truth odometry. Recently, several similar methods have been proposed at around the same time to improve the above framework on VLN. Wang et al. [72] proposed the structured scene memory module, which is designed to encode the experienced perceptions accurately. They argued that their graph representation serves as a structured scene representation, which addresses and disentangles visual and geometric cues in the environment [72]. Deng et al. [23] proposed an evolving graphical planner where the model can perform global planning with the dynamically constructed graph representation. Compared to [72], a Graph Convolutional Network (GCN) [44] was employed to encode the global state from the graph directly instead of iteratively feeding nodes through a GRU. Zhu et al. [81] further extended the graph-based mapping methods to scenario-oriented object navigation, where the target is defined by a language-based scenario description.

While these graph-based methods are slightly different in terms of planning, they all follow the paradigm where the agent plans for the next actions globally with constructed graphs. Compared to room-based planning, the graph-based maps qualify representations of each step with a learnable encoding at an arbitrary scale. Furthermore, by modelling the navigation problems with a graph, the agent can perform reasoning and action planning globally by traversing the constructed graph, which benefits the planning substantially compared to local action planning.

Besides the graph-based mapping, some have tried to build deterministic maps of the geometrical environment, which only encodes simple information such as the occupancy (traversability). For example, an egocentric geometrical map was constructed in [13] using projections from the point cloud. Then the map is used for both action planning and obstacle avoidance. Wani et al. [73] focused on evaluating the effectiveness of such shallow maps. In addition to the commonly used geometrical modelling of whether a location is occupied by obstacles or not, the authors proposed several other mappings by incorporating

the semantic meaning of the mapped objects. Purushwalkam et al. [60] further modelled the map construction as the task definition in the audio-visual embodied environment. While they implemented the policy following the exploration rather than navigation tasks, the observed representation is converted to egocentric semantic maps of the floorplan using upsampling and convolutional layers.

In our AV-GeN framework, we construct a geometrical occupancy map and an acoustic audio intensity map following [13]. While we implement the geometrical map for ease of deterministic action planning similar to the above methods, the audio intensity map will serve as an indicator for the position of the audio goal. Compared to the graph-based or geometry-based maps, the intensity map we construct will be encoded as an indicator of the goal instead of a tool for path planning.

2.2.3 Memory Mechanism

Another method to improve the navigation agent is to enable the model to effectively leverage experience from previous observations. Unlike map-based methods that encode past observations explicitly, these methods aim at finding out the helpful hints from memory implicitly. The AV-Nav [12] and AV-WaN [13] methods on the AVN task used a Gated Recurrent Unit (GRU) [20] that encodes the knowledge from memory in a hidden state. While this method is simple and straightforward, the memory stored in a fix-length vector can be limited. With the sequences of memory growing long, the GRU might forget some of the past knowledge, leading to degradations in performance.

To effectively leverage navigation memories, Fang et al. [25] proposed a general framework for visual navigation tasks called the scene-memory transformer. While most components of the framework are designed trivially with simple convolutional networks, they replaced the GRU with a transformer [71] that is widely believed to be effective in learning representations from sequences of data. Specifically, with the features extracted from the sequences of past observations, the transformer generates a global descriptor on the current state based on cross-attention on past latent representations. The proposed method outperforms many object-goal navigation methods, demonstrating that memorising past observations is of vital importance for navigation tasks. Similarly, Chen et al. [16] applied the transformer-based memory to the VLN task. In addition to the memories, the authors proposed to encode the multi-modality input with cross-modal attention, instead of simply concatenating the features from different modalities. Meanwhile, they stored the step memory using a learned compact encoding instead of the complete step-wise feature vector. The coarse-scale encoding substantially reduces the complexity of the feature space that the transformer needs to learn, as a result, the framework can be trained more efficiently.

Such transformer-based memory is also used in the AVN literature. In 2021, Chen et al. [11] tried to apply the scene-memory transformer on a variation of the AVN task called "semantic audio-visual navigation". In this task (as shown in Figure 2.3f), the sound will not be heard constantly at each step but are sporadic or short in duration, and the agent must be equipped with a solid memory module to enable the memorisation of observed sounds. To solve the semantic AVN task, the authors implemented the scene-memory transformer and added a GRU as the goal descriptor network that aggregates the step-wise acoustic observations. This method outperforms other navigation frameworks adapted from the AVN literature, as it enables the agent to memorise the periodic sounds at previous steps. Although the authors claimed that the memory mechanism contributes significantly to the success of the semantic AVN task, we have found that the memory-based method does not provide a substantial performance gain on the classical AVN task, where the memory is less important.

While the transformer-based memory mechanism can effectively leverage the past observations, the navigation planning is still limited to a local action space. Chen et al. [17] alleviate this issue by combining graph-based global planning with transformer memory for the VLN task. It implements both graph-based visual memory and transformer-based navigation memory. In this way, the local representation learned with the transformer can be further used to query for appropriate global action on the graph. Moreover, the local observation is encoded with fine-grained representations to compensate for the information loss in the coarse-grained graph-based encoding.

To conclude, implicit memory encoding is a meaningful direction to be further explored. In the proposed AV-GeN framework, we also implement a GRU-based observation encoder to utilise implicit memories following [12, 13]. We do not adapt the explicit memory mechanisms as we have found that memorising past observations is of less importance when the goal-indicative audio signal is given at a step-wise base, meanwhile, the crucial navigation hints are already memorised with our geometrical and acoustic maps.

2.3 Audio-Visual Multi-Modality Learning

Audio-Visual navigation combines goal-oriented navigation with audio-visual multi-modality learning. In this section, we will first briefly introduce the relevant literature on non-embodied research for learning representations from the audio-visual dual-modality through their correspondence. Afterwards, we will discuss the research on audio-visual learning methods in the embodied AI background. Finally, we focus on the goal-oriented navigation frameworks designed for the core task of this thesis, the AVN task.

2.3.1 Audio-Visual Representation Learning

Audio-visual multi-modality representation learning has been an active field of study in deep learning. Researchers aim to learn meaningful and robust representations from audio and visual data under different task settings. Current AVN methods simply trained the encoders for each modality separately and concatenated the multi-modality features. Such methods provide a simple and effective solution to the AVN task, but they can hardly exploit the mutual correspondence between audio-visual modalities. The following paragraphs briefly review the existing audio-visual representation learning methods with insightful explorations on processing multi-modality signals.

A large proportion of the early audio-visual research focuses on learning representations separately for dual-modality by performing some proxy tasks. For example, Arandjelović and Zisserman [4] proposed a representation learning method by addressing the correspondence between audio and visual signals. The method optimises the feature extractors of both modalities by maximising the similarities between two modalities, which are estimated by a discriminator network. Gan et al. [27] proposed a self-supervised audio-visual representation learning method in an embodied environment. They clustered acoustic events into several clusters and used a CNN on the visual signals to predict the observed scenario into acoustic event categories. Several methods addressing similar issues can be observed in Table 2.1. These methods train decent feature extractors for visual and audio modalities from different types of input data. However, these methods typically require the sound source to be present in the visual observation. The correlation between the audio and visual modalities is linked by the semantic meaning of visually presented objects and sound-making objects. To be more specific, these method explores the audio information from the visual modality. As the sound sources have no visual presence in the AVN task, these ideas are less useful in the AVN settings but could be helpful in the semantic AVN task where the audio source can be identified visually.

Without the visual presence of the audio source, Chen et al. [14] tried to exploit the audio-related information from the spatial layout and geometry of the nearby environment instead of the sound-making object. The authors proposed that with the visual observations on the room geometry, AI models can perform audio dereverberation where the computer tries to reconstruct the high-quality original sound based on the sound heard remotely, with higher accuracy. As it is known that the quality of the sound is damaged through propagations and reflections on the obstacles or walls, the author believed the visual understanding of the nearby environment would assist with the dereverberation. By adding a vision branch to the ordinary dereverberation deep model as input, the authors observed that the performance of

Method	Novelty	Weaknesses
[4, 5]	learn by addressing audio-visual correspondence	require large amounts of data to converge
[48]	learns audio-visual synchronisation	sound source must appear in the video
[55]	uses video label	requires prior knowledge
[38]	decouples modalities into distinct components	clusters need to be pre-defined
[27]	learns to predict auditory events	requires pre-trained audio feature extractor
[14]	learns audio dereverberation	requires data collection
[32, 56]	learn to synthesis binaural audio	sound source must present in the video
[30]	leverages audio echoes	learned audio representation is subject to a specific sound

TABLE 2.1. Methods on audio-visual representation learning on non-embodied datasets.

the audio dereverberation rises substantially, with the visual information on the interior environment. To summarise, this research has indicated that other than semantic objects, spatial clues can also effectively contribute to the acoustic analyse models.

Furthermore, some recent studies have combined the audio-related hints in semantic objects and spatial information to facilitate learning the coherence between room geometry and audio signals. Garg et al. [32] proposed that given a visual observation where sound-making objects are present, realistic binaural audio signals could be generated with deep models. By applying a series of auxiliary losses including spectrogram/audio-wave regression loss, RIR prediction loss, spatial coherence loss, and consecutive frame consistency loss, the authors showed that meaningful binaural sounds could be synthesised with basic feed-forward architectures. At the same time, Parida et al. [56] proposed a similar binaural sound synthesis framework, where the authors designed an audio-visual cross attention module to facilitate the multi-modality feature fusion.

On the contrary, some methods have attempted to explore spatial information from the auditory modality. For example, the publishers of the AVN task have argued that in the point-goal navigation task, if the agent is given the audio signal emitted from the goal position, then it can perform better than the plain point-goal agent [12]. While actually, the superiority of the audio-point-goal agent over the point-goal agent vanishes if the agent can only hear unfamiliar sounds, such a finding has proven that there exist implicit spatial clues in the received audio waves. Gao et al. [30] tried to treat audio signals as a form of

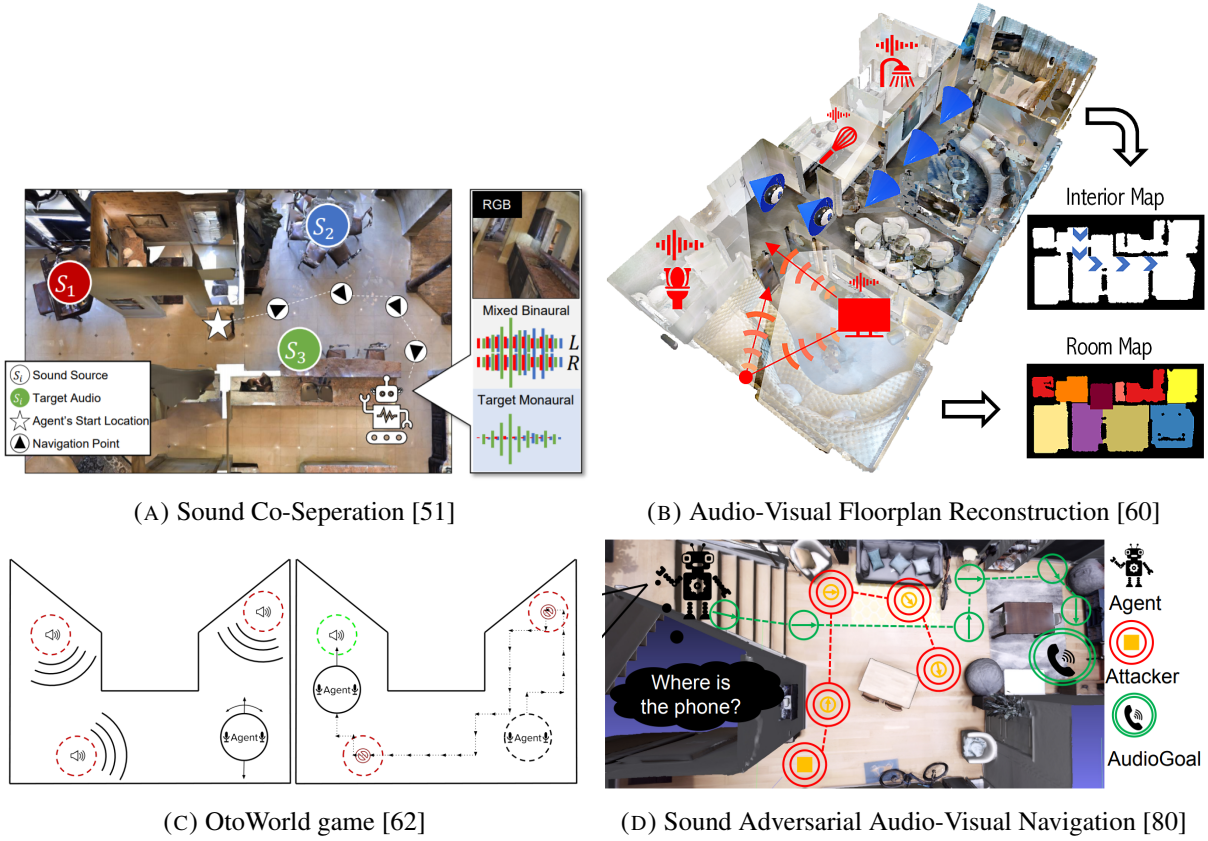
supervision to facilitate visual representation learning. They argued that similar to animals, intelligent agents should be capable of performing echolocation based on binaural echoes. By caching observations from the SoundSpaces simulator, the authors supervised the visual feature extractor by a correspondence prediction task, where the model needs to predict whether the echoes heard corresponds to the current visual observation. The authors showed that the visual feature extractor trained with the correspondence prediction task could perform well at vision tasks such as depth predictions.

Although the above methods can be inspiring for our future works on leveraging the multi-modality signals, so far our method only explores the single-modality learning strategies on the audio signals with contrastive learning. And we would like to treat "exploring the room geometry from the audio signals" as promising future work.

2.3.2 Embodied Audio-Visual Learning

Some other audio-visual learning works are designed with embodied learning methods, hence are more related to the AVN literature. These works design tasks where the agents can explore the environment to learn more solid understandings of received audio signals. By training in the embodied environments, agents learn better representations of comprehending the relationships between audio signals and spatial hints related to the room geometry. Some agents can even be more competent in completing the AVN task by learning in an audio-visual embodied environment. In this section, we will introduce the existing embodied audio-visual learning methods.

Majumder et al. [51] highlight the idea that intelligent agents should learn the audio representations by intuitively exploring audio observations at different positions. Based on the SAVN environment [11], they proposed the sound co-separation task, where the agent can move around arbitrarily in the environment with multiple sound sources, and the agent needs to perform the sound separation task [35] during exploration, as shown in Figure 2.5a. In detail, the agent heard a mixed audio signal as input at each step, and it is required to separate the sounds of the target object from the mixed audio signal received. For better sound separation performance, the agent is encouraged to walk around in the environment and stop at the position where it believes the optimal sound separation quality could be obtained. The proposed solution first predicts the target binaural sound, based on which the monaural signal target is predicted. Finally, the monaural predictions are refined through an additional network and fed to the policy network for policy planning. It is worth noticing that the authors found that the agent learned with this complex co-separation task is also capable of completing the semantic AVN task.



(A) Sound Co-Seperation [51]

(B) Audio-Visual Floorplan Reconstruction [60]

(C) OtoWorld game [62]

(D) Sound Adversarial Audio-Visual Navigation [80]

FIGURE 2.5. Embodied Audio-Visual Learning tasks. Images adapted from [51, 60, 62, 80].

Similarly, Purushwalkam et al. [60] proposed audio-visual floorplan reconstruction. The agent is required to explore the environment and reconstruct a semantic floorplan describing the room geometry and room usage. As could be observed in Figure 2.5b, the sound signals will be emitted at different rooms corresponding to the room usage, e.g., the water flushing sound will be emitted from the bathroom. This task helps the agent better learn the semantic meanings of the sounds by stressing the agent to exploit the correspondence between room semantics and sounds that are emitted from the room.

Some other works directly reformulate the AVN problem to train cleverer agents with more challenging task settings. Ranadive et al. [62] proposed a complex variant of the AVN task called the 'OtoWorld game' which combines the sound co-separation task. As presented in Figure 2.5c In an environment where multiple sound-making objects are placed at different locations, the agent is required to navigate to all the sources successively to shut them down. The agent is evaluated on both the audio-visual navigation ability and sound separation ability for successful completion of the task. The navigation model follows a simple reinforcement learning structure where the observed spatial features are extracted with

classical signal processing methods instead of learning-based methods. Although the overall framework is quite preliminary, it is inspiring to find that the agent learns more robust representations when learning to navigate with multiple successive goals compared to a single audio goal.

To improve the audio representation learning in the AVN task, Younes et al. [79] proposed to train the AVN agent with the sound source moving in the environment. Such a task definition effective reduce the overfitting caused by the fixed start-target position pairs defined in the dataset episodes. Furthermore, in the aforementioned adversarial audio-visual navigation [80], a moveable agent is trained to play distractor sounds to interlope the navigation agent. The authors demonstrated that agents trained with such settings could not only learn more generalisable representations but also grow more robust against distracting attacks.

To conclude, the attempts to improve audio-visual learning with embodied task settings have lightened an inspiring direction for developing more powerful general-purpose AI systems. While some research has shown that AVN agents can be better trained in more complex environments, the implementation of the additional task goals, constraints, or other reformulations typically requires more resources or modifications to the environment, which might not always be available in real-life deployment. Compared to [79, 80] which enhance the AVN agent, our method distinguishes itself by improving the generalisation of agents from available signals that can be obtained directly from a basic AVN task setting. Moreover, the proposed AFSO and sound augmentation method can be always deployed not only to our AV-GeN framework but to any AVN framework to obtain a significant performance gain.

2.3.3 Audio-Visual Navigation

As several variants of the AVN task such as [11, 60, 80, 79] have been discussed in the previous section, we focus on methods designed for the standard AVN task here.

The AVN problem was proposed by Chen et al. [12] in 2020. Though one of the main contributions is that they released the SoundSpaces audio-visual simulation environment, they also proposed an end-to-end framework that trains an agent that is competent in solving the AVN task. As shown in Figure 2.6, the model first converts the dual-modality inputs to feature vectors using different CNNs. Then the feature vectors are concatenated as the descriptor of the current observation. Next, the observation descriptor is fed into a GRU [20], to produce the state descriptor with the global knowledge of the prior navigation states. Eventually, the network is optimised using Actor-Critic [46] with PPO [66] RL

algorithm introduced in previous sections, where the whole model is trained to maximise the 'goodness' of the state as estimated by the critic. The actor finally produces a probabilistic distribution from which the next action is sampled. The authors named the method 'AV-NAV' and showed that this framework could be applied to solve the AVN task. Due to its simplicity and effectiveness, the AV-NAV has been used as a baseline for developing and comparing novel AVN frameworks.

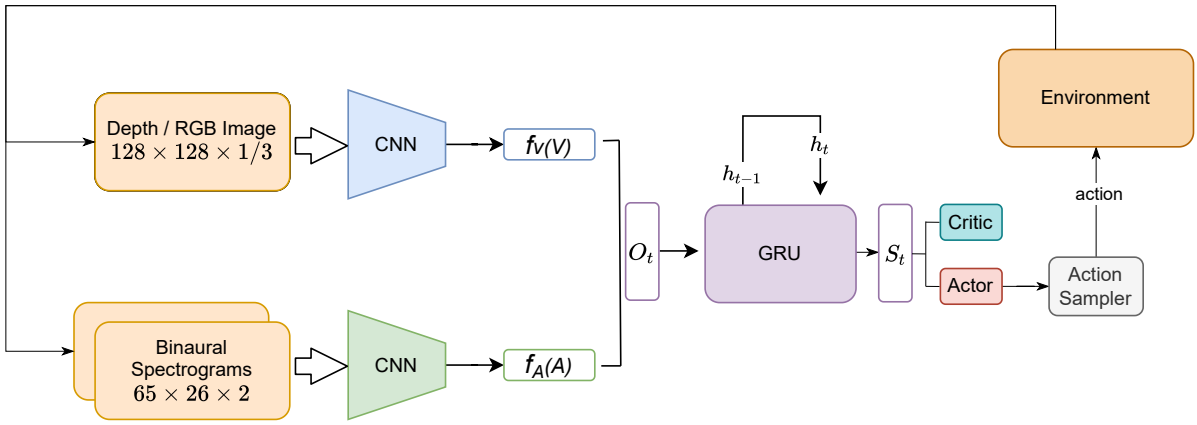


FIGURE 2.6. The graphical illustration of the audio-visual navigation (AV-NAV) framework [12].

Besides showing the effectiveness of the proposed framework, the authors have also proven that in addition to serving as a goal position indicator, the audio signals can also reveal vivid spatial hints on the room geometry of the surroundings, by showing the superiority of the audio-point-goal agent over the point-goal agent (the audio-point-goal agent will receive a displacement vector indicating the goal position in addition to the audio signal). Such hints can substantially improve the agent at obstacle avoidance and path planning. However, such spatial hints can only be learned effectively for familiar sounds, and hardly generalise to unknown sounds. In fact, under the settings where point-goal sensors are provided, the agent no longer benefits from receiving audio signals if the sound is never heard during training, which contradicts the claimed opinion and lays a research question to be studied.

Other than the AVN task defined on the SoundSpaces dataset, Gan et al. [29] also proposed an AVN framework. The authors solved the navigation problem by explicitly constructing an occupancy map of the room. In addition to the dynamic map construction as proposed in WaN, they also developed an explore-and-navigation pipeline, where the agent is allowed to explore the nearby environment to construct the occupancy map before navigation. Moreover, they trained the policy network with an auxiliary task where the agent needs to predict the audio goal position from the binaural spectrograms.

However, they implement their method only on game-based simulations and cannot be compared with the AVN methods on the SoundSpaces simulator or generalised to real-life scenarios.

Inspired by [29], in 2021, the SoundSpaces group proposed the Waypoint Navigation (AV-WaN) method [13] with two novel modules to enhance the trivial framework in the AV-NAV paper. First, the authors proposed that two top-down geometric maps can be maintained egocentrically, to memorise the occupancy and audio intensity of the navigation nodes. Similar to [29], the occupancy map is created based on the visually observed depth signal; A 3D point cloud is built according to the depth image, and the anterior occupancy map is then constructed by projecting the point cloud to the horizontal axis. The global mapping of the environment is obtained by accumulating the anterior maps at different positions. Similarly, the acoustic map is constructed by recording the audio intensity at each step. The global state descriptor is then obtained by concatenating the feature extracted from the audio spectrogram, the occupancy map, and the acoustic map respectively.

Moreover, the authors proposed that with the help of the occupancy map, the agent can move to arbitrary navigable positions indicated by the map in a deterministic way. In detail, the model predicts a nearby waypoint on the map as the intermediate target, and the agent can navigate to the predicted waypoint using Dijkstra's algorithm. In this way, the agent can effectively avoid colliding with the surrounding obstacles, thus outperforming the AV-Nav method significantly.

The proposed geometric map is an explicit way of leveraging experience from past observations as introduced in Section 2.2.2. Meanwhile, the model can avoid nearby obstacles effectively by navigating to the intermediate waypoint planned globally. These two advantages improve the robustness of the navigation model and make the WaN method the state-of-the-art audio AVN method. However, although both the AV-NAV and the AV-WaN methods demonstrate decent performance on navigation with novel scenes and training sound, the agents can hardly generalise in cases where the emitted sound is never heard during training. While methods like [79] claimed that performance on unheard sound could be improved by training with moving distracting sounds, it involves heavy addition of extra supervised sounds during training episodes, hence cannot be compared equitably with the AV-NAV and the AV-WaN methods.

Based on designs of AV-WaN, we implement our AV-GeN framework, which significantly improves the navigation performance by enhancing the generalisation of the agent on unheard audio goals. Meanwhile, different from [79] which collects more supervision with complicated reformulations of the task,

the proposed AV-GeN framework learns from the original AVN task setting without altering the simulations. Furthermore, the proposed AFSO and sound augmentation methods do not rely on the implementation of the AV-GeN framework, they can also be decoupled from the AV-GeN framework and integrated with other AVN frameworks to enhance their generalisation ability.

CHAPTER 3

Methods

In this chapter, we present our Audio Visual Generalisable Navigation (AV-GeN) framework. The overview of the AV-GeN architecture is demonstrated in Figure 3.1. Firstly, the source sound for each navigation episode will be augmented with the sound augmentation module. Afterwards, the visual and audio inputs are converted into navigation features with vision mapping, acoustic mapping, and audio encoding modules using CNN-based encoders. The input features will then be concatenated and fed into a GRU [20] to leverage navigation information from past steps and produce the navigation state descriptor S_t, which will be optimised using the actor-critic RL algorithm. Based on S_t The actor head will predict a waypoint that will work as an intermediate navigation goal. The path planning module will generate a sequence of low-level actions to reach the planned waypoint without hitting obstacles based on a maintained occupancy map.

It can be observed that the AV-GeN framework can be divided into six components (modules) with different functionalities. In the following sections, we will introduce the six components and the input processing in the AV-GeN framework. While several modules are implemented based on the AV-WaN framework [13] and will only be introduced briefly, we will discuss in detail the audio encoding module, where we design the AFSO method, as well as the sound augmentation module we proposed. AFSO and sound augmentation can be easily implemented as modules and inserted into AVN frameworks, and they significantly improve the generalisation of the AV-GeN framework for learning audio representations.

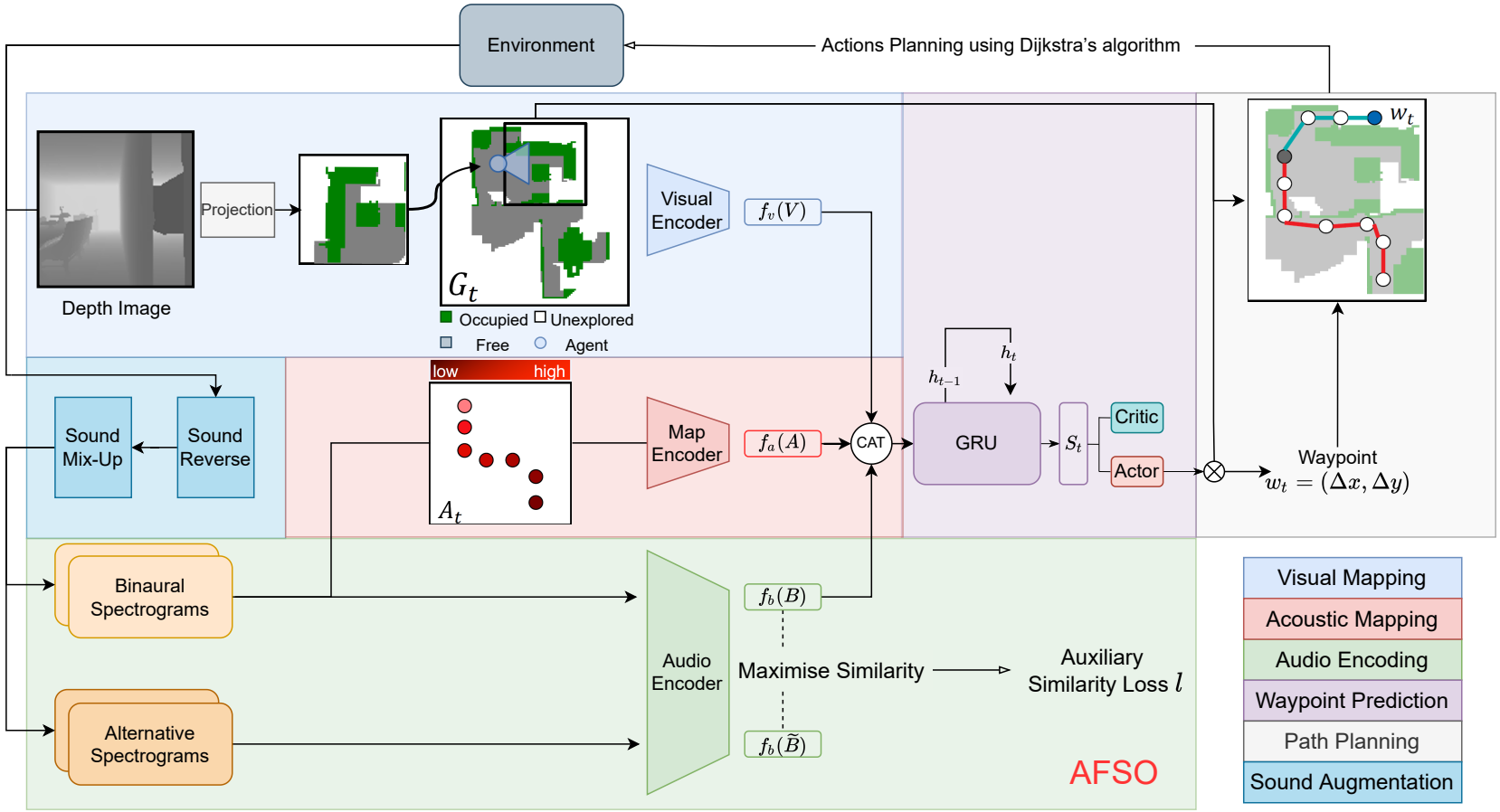


FIGURE 3.1. The architecture of the proposed AV-GeN framework. Firstly, the source sound for the navigation will be augmented with the sound augmentation module. Afterwards, the multi-modality inputs are converted into navigation features with vision mapping, acoustic mapping, and audio encoding modules. The input features will then be concatenated for predictions of a waypoint as an intermediate navigation goal. Finally, the agent will make a sequence of low-level actions to reach the planned waypoint.

3.1 Input Processing

In this section, we briefly introduce the preliminaries on how we obtain the multi-modality input from the simulator. Overall, the simulator works as a graph of navigable source and listener positions, where the agent could move around in the environment along the valid paths between immediate neighbour positions. While the visual observations at these navigation nodes can be rendered based on manually configured camera settings, the acoustics of the indoor scenes are simulated according to the precomputed room impulse responses (RIR). RIR encodes the remote acoustic information based on room geometry, relative positions and reflective materials, and it serves as a transfer function between the sound source and the microphone [47]. Let $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^N$ denote the set of N possible sound-emitting positions, and let $\mathcal{L} = \{(x_l^s, y_l^s)\}_{i=1}^N$ denote the set of N possible positions of the listener (agent). Consequently, for each possible pair of source and listener positions $\mathcal{S} \times \mathcal{L}$, the audio observation can be rendered as:

$$O_{s,l}^{left},\ O_{s,l}^{right} = f(A, R_{s,l}^{left}),\ f(A, R_{s,l}^{right}), \tag{3.1}$$

where $[O_{s,l}^{left},\ O_{s,l}^{right}]$ denotes the binaural (i.e., left and right ears) audio observation when the listener is at position l and the sound source is placed at position s. $f()$ denotes the array convolution operator. A denotes the monaural source audio waveform. A graphical illustration of the relationship between RIR, source sound (original sound), and the observed audio signal is shown in Figure 3.2. By convolving the source audio waveform with the RIR, the binaural auditory observation is rendered.

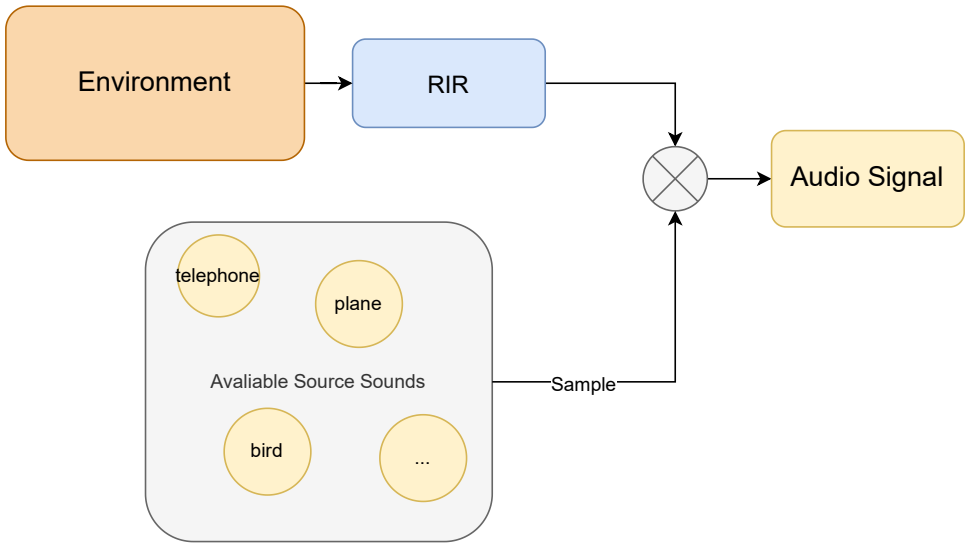


FIGURE 3.2. The graphical illustration of the relationship between RIR, source sound, and observed audio signal.

The agent receives inputs from both visual and audio modalities at each step. The visionary signal is given as a single-channel depth image of height and width 128 by 128. The auditory input is of the binaural audio waveform, represented as an array of the shape $(2, sr)$, where sr denotes the sampling rate. While the three-dimensional visual inputs can be effectively processed by a CNN, learning auditory features from lengthy wave signals is challenging for modern deep neural networks. To allow feature extraction using CNN, the binaural sound waves are converted to spectrograms using Short-Time Fourier Transform (STFT). Following the SoundSpaces team [12], the STFT was computed with a hop length of 160 samples and a windowed signal length of 512 samples. The first 1000 milliseconds of the binaural audio are input to the STFT algorithm, resulting in spectrograms of size and 257×101 for each binaural channel. Finally, the generated spectrograms are downsampled by a factor of 4 on both axes, followed by a logarithm transformation applied to the spectrograms to enhance the contrast [12].

3.2 Visual Mapping

Following [13], we convert the visual input signals from depth image to a top-down occupancy map G of size 200×200 as our visual perceptions. First, the depth image is back-projected into the coordinate by computing the local 3D point cloud based on the intrinsic parameters of the agent camera. Afterwards, the 3D point cloud is projected to a 2D top-down egocentric map G_{local} which describes the local occupancy in front of the agent. While local maps with arbitrary size could be constructed technically, we only maintain a map G_{local} of size (3×3) meters following [13], where modern depth sensors in the real world are believed to be reliable. In detail, the G_{local} is modelled with two channels, one describing the occupancy/free status of the environment and one for the explored/unexplored status. All cells are initialised to occupied and unexplored as default states, the cells will be updated to occupied if a 3D point at a height value within the range [0.2, 1.5] meters lies in the cell, and cells that contain any 3D points will be updated to the explored state. Finally, the global geometric map at the current step G_t will be maintained in an allocentric way by rotating G_{local} to the same orientation as G_t and averaged with the maintained map at the previous step G_{t-1}. The visual features at the current step will be then extracted with a CNN-based encoder f_v from G_t.

With the geometrical mapping, the complex visual inputs are simplified into map-based representations that are smaller in dimensions while effectively preserving all information necessary for navigation and obstacle avoidance. As a result, the geometric map representation effectively improves the generalisation of the agent in unseen scenes and novel environments. Meanwhile, the globally maintained map

explicitly leverages the past visual observations by recording them on the map, thus enabling accurate long-term path planning.

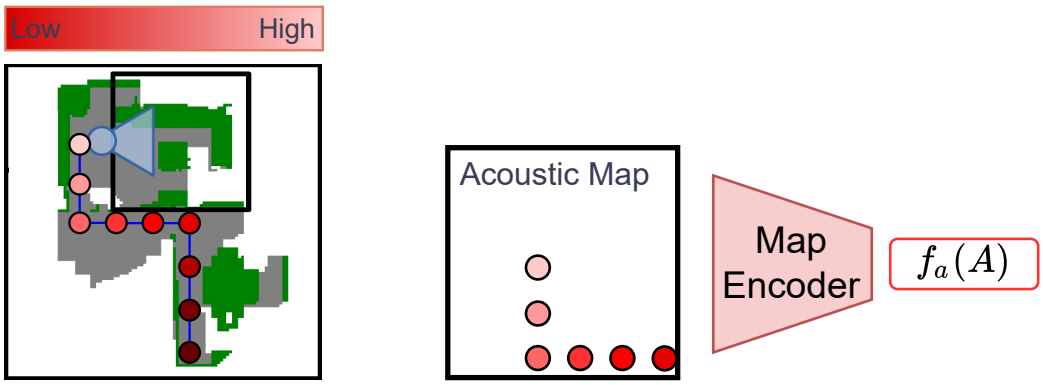
3.3 Acoustic Mapping

In addition to the geometrical map, we also implement an acoustic map following [13]. Given the input binaural audio observation, we compute the sound intensity of the audio signals by averaging the root mean square of the audio signals at the left and right channels. As a new intensity value can be obtained at each step, we can then maintain the acoustic map A_t with intensity values at different positions in an egocentric way similar to G_t. The acoustic map aggregates the audio intensity over positions and time by recording the moving averages of the intensity values. Unlike geometrical mapping where the global map is fed to the encoder as the visual perception, we clip a local acoustic map of size 20×20 as the input to the CNN-based acoustic map encoder.

As illustrated in Figure 3.3, the acoustic map made up of audio intensity values will reveal both directional and distance information to the audio goal. In detail, the agent can infer the relative direction of the audio goal by observing the gradient in the acoustic map. Meanwhile, the distance to the audio goal can also be estimated based on the changes in the audio intensity values, as the agent will receive audio signals with relatively higher intensity values near the audio goal. According to Chen et al. [13], the acoustic mapping gives a coarse sense of direction when the agent is far from the target and offers the directional clues with increasing precision as the agent keeps approaching the goal.

3.4 AFSO-based Audio Encoding

Modern AVN frameworks such as AV-NAV [12] and AV-WaN [13] use a CNN-based audio encoder to extract the goal-oriented features from the spectrograms and directly optimise the audio encoder with the corresponding RL algorithm. While the audio goal information is also extracted from the binaural spectrograms using a CNN-based encoder, our AV-GeN framework is equipped with the proposed Audio Feature Similarity Optimisation (AFSO) method to improve the audio encoder. As a novel optimisation method, AFSO can be conveniently deployed to train the audio encoder with an auxiliary loss, in addition to the standard RL losses applied in previous methods. With the AFSO module, our AV-GeN framework significantly improves the generalisation ability of the agents in terms of understanding



(A) (B)

FIGURE 3.3. A hypothetical graphical illustration of the acoustic map. (A) A navigation trajectory with recorded intensity values at each step. It can be observed that the closer the agent is to the audio goal, the larger the intensity value will be. (B) An egocentric latency map is cropped from the global latency map and fed into the acoustic map encoder for feature extraction.

the audio observation indicative of the relative goal positions. In this section, we detailly explain the proposed AFSO method.

3.4.1 Intuition

In all existing methods, the models show substantially better performance at navigating to familiar audio goals that have been learned during training, than navigating to audio goals where the emitted sound is never heard before. When changing the evaluation audio goals from heard sounds to unheard sounds, existing AVN frameworks implemented on the SoundSpaces simulator usually suffer from a decrease in performance of approximately 50%. For example, according to Chen et al. [13], if the evaluation sound split is changed from heard sounds to unheard sounds on the Matterport3D dataset, the performance of an AV-NAV agent decreased from 55.1% SPL to 25.9%, whereas the performance of an AV-WaN agent decreased from 72.3% to 36.2%. Such a relative decrease shows that the trained agent has severely overfit the training sound, and it has learned many audio-specific features, instead of audio-goal descriptors that can generalise to various unheard sounds. On the other hand, the sound overfitting phenomenon also indicates the promising potential of methods that are designed to improve the generalisation ability of the audio encoder.

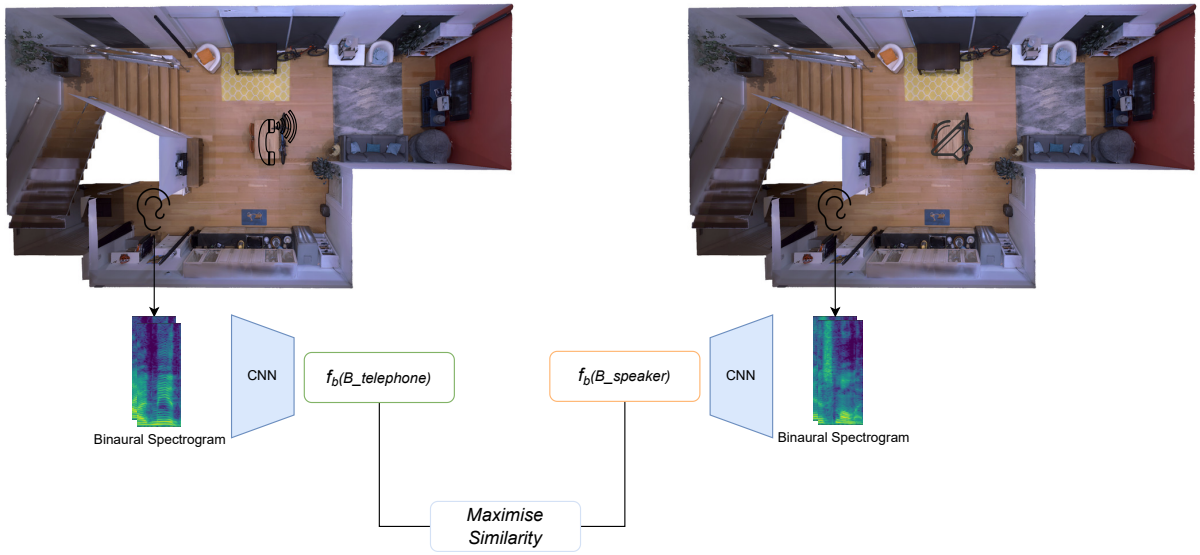
In this section, we introduce AFSO, a method designed to alleviate the overfitting problem by encouraging the audio encoder to learn sound-agnostic features that are not subject to different types of sounds.

Intuitively, the audio encoder does not need to learn features that distinguish what sound the agent heard. The only valuable information for the AVN task that the model needs to learn is the source-receiver spatial relationships implied by the audio signals. The AFSO method maximises the learned feature similarity between pairs of audio observations that are sourced from different sounds but imply highly similar source-receiver spatial relationships, i.e., information on the relative audio goal position. Meanwhile, the AFSO method minimises the similarity between audio observation pairs that imply distinct audio goal positions.

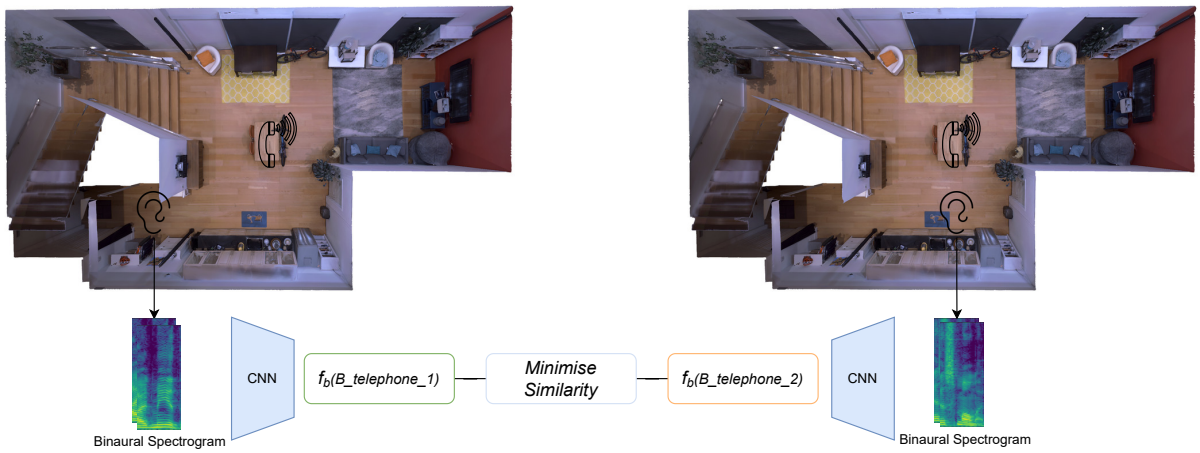
The high-level idea of the method is illustrated in Figure 3.4; On the left of Figure 3.4a, the ears at the bottom-left are receiving sounds emitted from a telephone, whereas in the image at the right, the position of the source and receiver is unchanged, but the agent is receiving sounds emitted from a speaker. In this case, we maximise the similarity between the extracted features from the telephone sound and the speaker sound, as they imply the identical source-receiver spatial relationships, even though the spectrograms received from the telephone and the speaker are very different. On the other hand, as shown in Figure 3.4b. The agent receives audio signals from the same sound source (telephone) at different receiver positions. In this case, although the agent acquires similar spectrograms from the same sound source, we minimise the feature similarity between such observation pairs, as they indicate distinct source-receiver spatial relationships. As a result, we force the audio encoder to learn only the necessary information for the navigation task and abandon other characteristics related to the sound itself.

3.4.2 Define Similar Audio Pairs

However, the concept of "similar source-receiver spatial relationships" is hard to be defined and estimated explicitly in 3D environments. If the similarity between pairs of audio observations is estimated based on their relative displacement vector to the audio-goal, i.e., the pairs of audios with similar relative source-receiver displacement will imply higher similarities in the source-receiver spatial relationship, then the disturbance in audio signals resulting from reflections and reverberations caused by obstacles between the audio source and the agent receiver will not be considered, as shown in Figure 3.5a. It can be observed that the pairs of audio observations from the two side of the figure shares the same source-receiver displacements but imply distinct source-receiver spatial relationships due to the change in the nearby environment (For example, the propagation of the audio wave on the left figure is blocked by the wall).



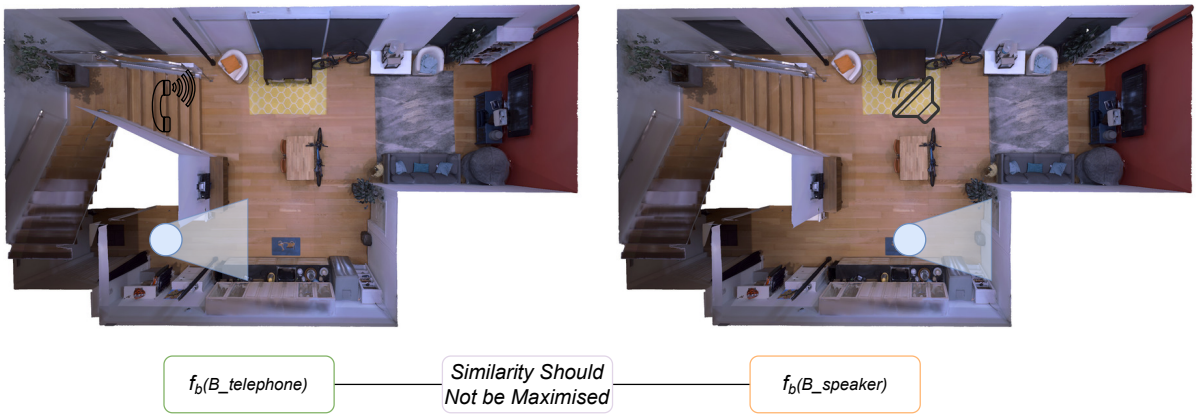
(A)



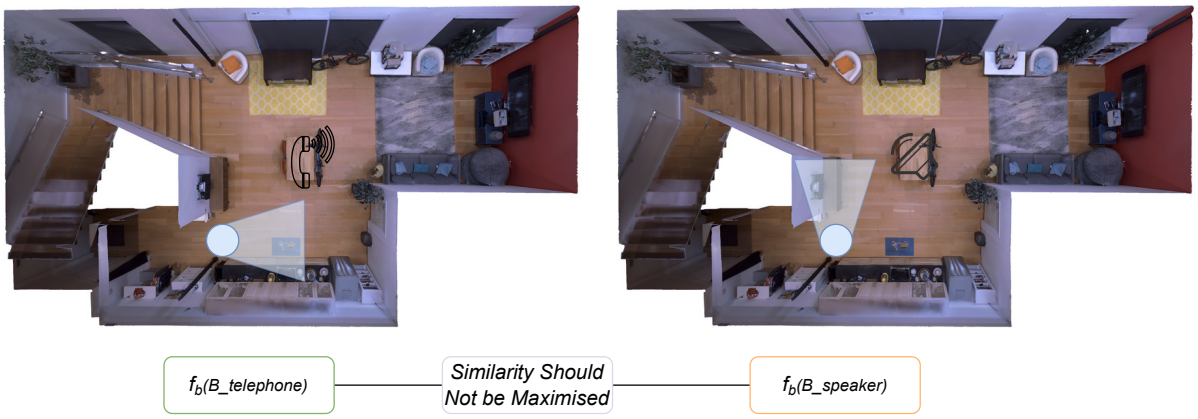
(B)

FIGURE 3.4. The high-level intuition of the Audio Feature Similarity Optimisation method. Take a telephone and a speaker as an example of the source sound. (A). AFSO maximise the similarity between audio features of different sounds if they imply the same position information of the audio goal. (B). AFSO minimise the similarity between audio features that imply different position information of the audio goal, even if they are emitted from the same sound source.

On the other hand, if the similarity between pairs of audio observations is estimated based on their relative orders of being heard in the navigation trajectory, i.e., consecutive audio observations during navigation steps will imply higher relative source-receiver position similarities, then the feature similarity between pairs of audio observations that indicates much different information on the audio goal positions will be maximised in cases the agent takes the actions of in-place rotation, as shown in Figure



(A)



(B)

FIGURE 3.5. The cases where the audio features similarity between the left and right scenarios should not be maximised. (A). The agent and the sound source in the environment have identical source-receiver displacement. However, the spatial hints in the heard audio will be different due to the change in the nearby environment. (B) The consecutive audio observations might indicate divergent spatial clues to the audio goal. Therefore, the audio similarity should not be defined according to the relative orders of being heard in the navigation trajectory.

3.5b. The figure presents the consecutive pairs of audio observations obtained by taking an in-place rotation action, where the binaural audio pair evidently imply distinct source-receiver spatial relationships in the egocentric view of the agent.

The failures on the above similarity estimation strategies have indicated the difficulties in explicitly modelling the desired audio pair feature similarities. Therefore, instead of computing pseudo-labels on the pair-wise similarity values explicitly and supervising the audio encoder to generate features that match the similarity measurements, we define the pairs of audio observations as ‘similar’ only if the pair of audio is sourced from the exact same scene, audio source position, and receiver position. Otherwise,

the pairs of audio signals will be considered dissimilar pairs, and the feature similarity between the pairs will be minimised. Under the SoundSpaces simulator, the above similarity criteria can be further summarised as the audio observations simulated with the same RIR file with be reckoned as positive pairs.

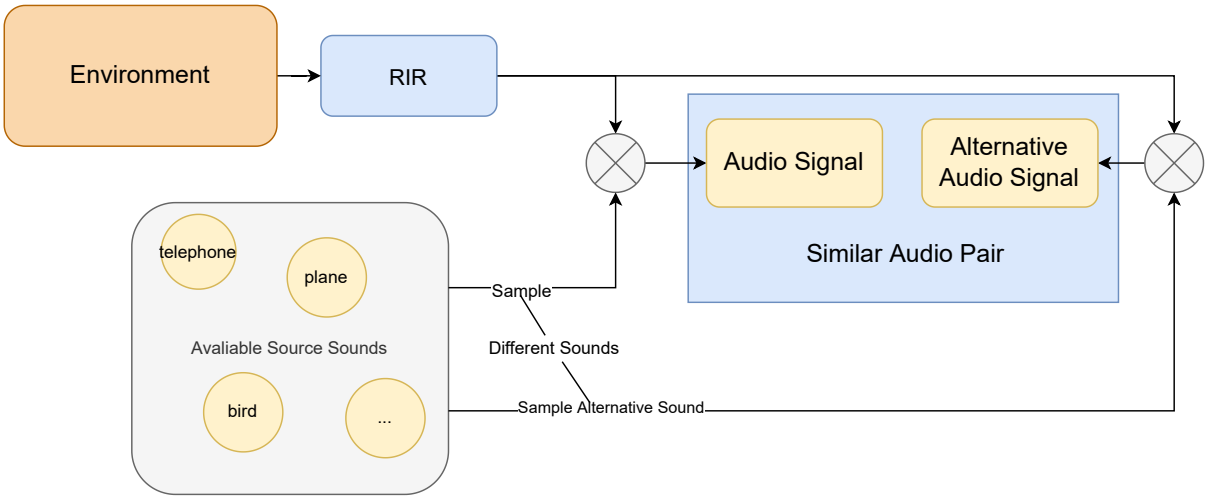


FIGURE 3.6. The graphical illustration of how we define and generate the pairs of audios of which the feature similarity is to be maximised.

To obtain the defined similar pairs of audio observations while avoiding disturbing the existing RL learning framework, at each navigation step, we manually simulate the audio signal $\widetilde{b}$ that should be considered 'similar' to the current audio observations b but with different sounds (we refer to such simulated sounds as the 'alternative' audio), as shown in Figure 3.6. We record the simulated alternative audios at each step and use them to pair with the audios heard during navigation.

In this way, we obtain similar pairs of audio signals with the same scene, audio source position, and receiver position. For the collected pairs simulated with the same RIR, we maximise their feature similarity in the latent space. On the other hand, we pair the audio observation at specific steps with alternative audios obtained at other navigation steps and minimise the similarity of their feature in the latent space, as demonstrated in Figure 3.7. Instead of searching for all similar pairs in the batch of audio observations, the described pairing method can be implemented easier and runs more efficiently, while not affecting the overall RL framework. Note that the AFSO method can also be adapted to AVN frameworks other than AV-GeN flexibly, a graphical illustration of how the AFSO method can be plugged into generic AVN frameworks is depicted in Figure 3.8.

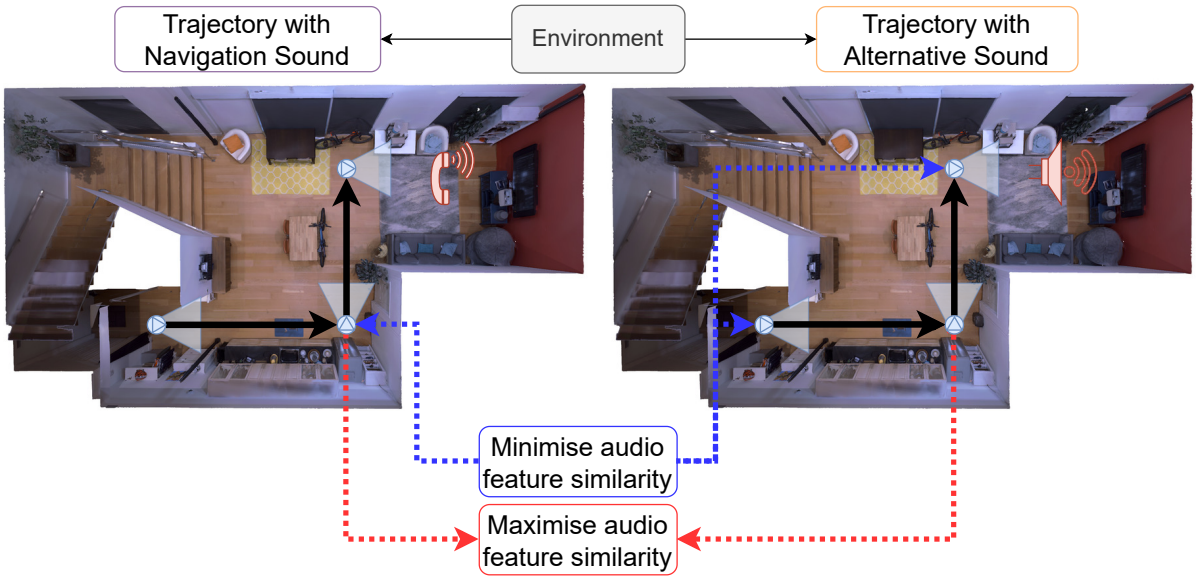


FIGURE 3.7. The graphical illustration of how we define the audio pairs of which the feature similarity is to be optimised. In the example shown, the audio signal of the second step-wise observation will be paired with all three audio observations in the trajectory with the alternative sounds. The similarity between the pair with identical navigation status will be maximised while others will be minimised.

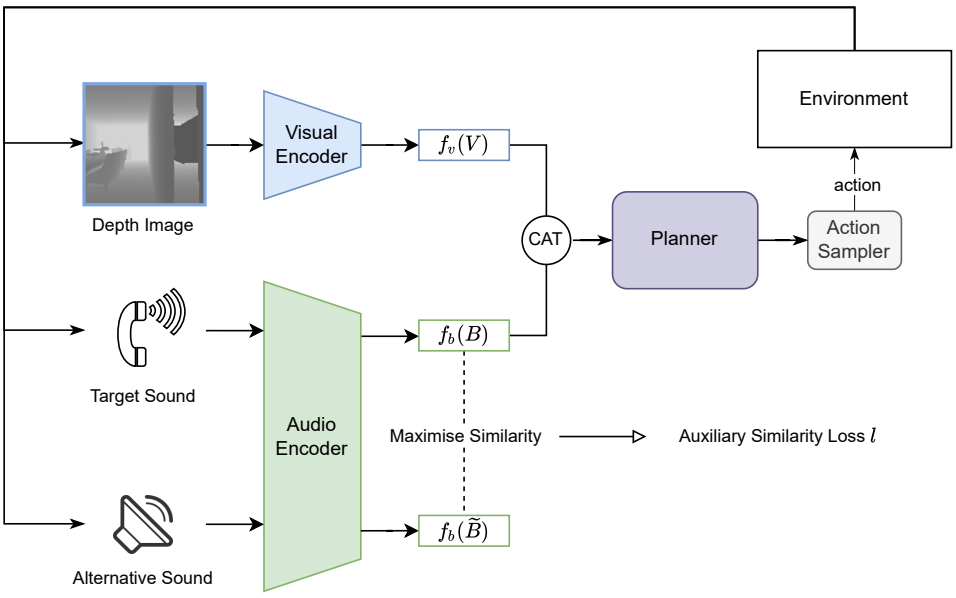


FIGURE 3.8. Schematic illustration of how our AFSO method is plugged into a generic AVN framework.

3.4.3 Optimisation Method

We implement the similarity optimisation process using a contrastive learning method [70, 18]. With the PPO algorithm, the RL framework collects trajectories as batches to stabilise the back-propagation. With the alternative sound generation method, we can obtain another batch of alternative sound observations at the loss computation stage. We forward the two batches of audio signals through the audio encoder f and a projection head g with trainable parameters, then we pair the features in the two batches to compute the pair-wise cosine similarity in the latent space. Note that the projection head is only used to estimate the similarities and has no impact on path planning. Finally, the audio encoder is updated using the InfoNCE loss [70], the loss function for a positive pair of audio signals (i, j) is defined as:

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)} , \tag{3.2}$$

where z denotes the binaural audio representation in the latent space, $\mathbb{1}_{k \neq i} \in \{0, 1\}$ is an indicator function evaluating to 1 iff $k \neq i$, sim denotes the cosine similarity function $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$, and τ denotes a temperature parameter. The InfoNCE loss can be viewed as a normalized temperature-scaled version of the standard cross-entropy loss, and it has proved to be successful in various studies [70, 76, 18, 19].

A high-level overview of the AFSO algorithm is shown in Algorithm 1. Following the alternative sound generation strategy introduced in the last section, we perform the similarity calculation between the batch elements to derive the contrastive loss. Finally, we apply a weight factor w to the similarity loss $\sum_B l_{i,j}$ and combine it with standard AVN losses for optimisation.

3.4.4 Batch Sampling

As introduced previously, the RL framework for the AVN task collects a batch of trajectories on different episodes and computes the accumulated gradients on the collected batch of observations to optimise the networks. By default, the AFSO method pairs and computes losses for each audio element in the batch, i.e., every audio observation in each trajectory will be paired with its alternative audio. However, such pairing implies that occasionally, negative pairs of sounds might be formed of audios and alternative audios that are collected at the same position, as the agent might travel to certain states multiple times during navigation. For example, consider the agent has taken the actions ‘turn left’ then ‘turn right,

Algorithm 1 Audio Feature Similarity Optimisation

Input: batch size N, constant τ, audio encoding function f, pairing sound simulator g, available training sounds $\mathcal{A}$.
for sampled audio batch $\{\boldsymbol{b}_k\}_{k=1}^{N}$ **do**
  for all $k \in \{1, \ldots, N\}$ **do**
    sample pairing sound $a \sim \mathcal{A}$.
    # simulate alternative audio
    $\tilde{b}_k = g(a, b_k)$
    # compute representations
    $h_{2k-1} = f(b_k)$
    $h_{2k} = f(\tilde{b}_k)$
    # projection
    $z_{2k-1} = g(h_{2k-1})$
    $z_{2k} = g(h_{2k})$
  end for
  for all $i \in \{1, \ldots, 2N\}$ and $j \in \{1, \ldots, 2N\}$ **do**
    $s_{i,j} = \boldsymbol{z}_i^{\top} \boldsymbol{z}_j / (\|\boldsymbol{z}_i\| \|\boldsymbol{z}_j\|)$ # pairwise similarity
  end for
  define $\ell(i, j)$ **as** $\ell(i, j) = -\log \frac{\exp(s_{i,j}/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(s_{i,k}/\tau)}$
  $\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$
  update audio encoder network f to minimize $\mathcal{L}$
end for
return the updated audio encoder network $f(\cdot)$

then the first and the last observations in the trajectory will be identical, as the agent moves back to its original state. If we apply the AFSO algorithm on such a collected batch, the feature similarity between the first and the third audio signals will be minimised, while they do imply the identical source-receiver spatial relationship, which conflicts with our intuition and is undesired.

A simple and effective way to eliminate such false-negative pairs of audio observations is to record the coordinate of the agent at each navigation step and removes the observations in the same trajectory with identical coordinates. While this method ensures the perfect removal of all mislabelled pairs, accessing the low-level information such as scenes and trajectories can be challenging to implement with the SoundSpaces simulator, meanwhile, it is also time-consuming and will severely affect the efficiency of the simulation.

To ensure the RL framework can run efficiently, we design a batch sampling strategy. As shown in Figure 3.9, at the agent update stage, we randomly sample m pairs of audio observations from the collected batch with n pairs of audio observations and input the sampled pairs to the AFSO algorithm 1.

The similarity losses are only computed among those sampled elements, which reduces the chances that false-positive audio pairs being included in the batch.

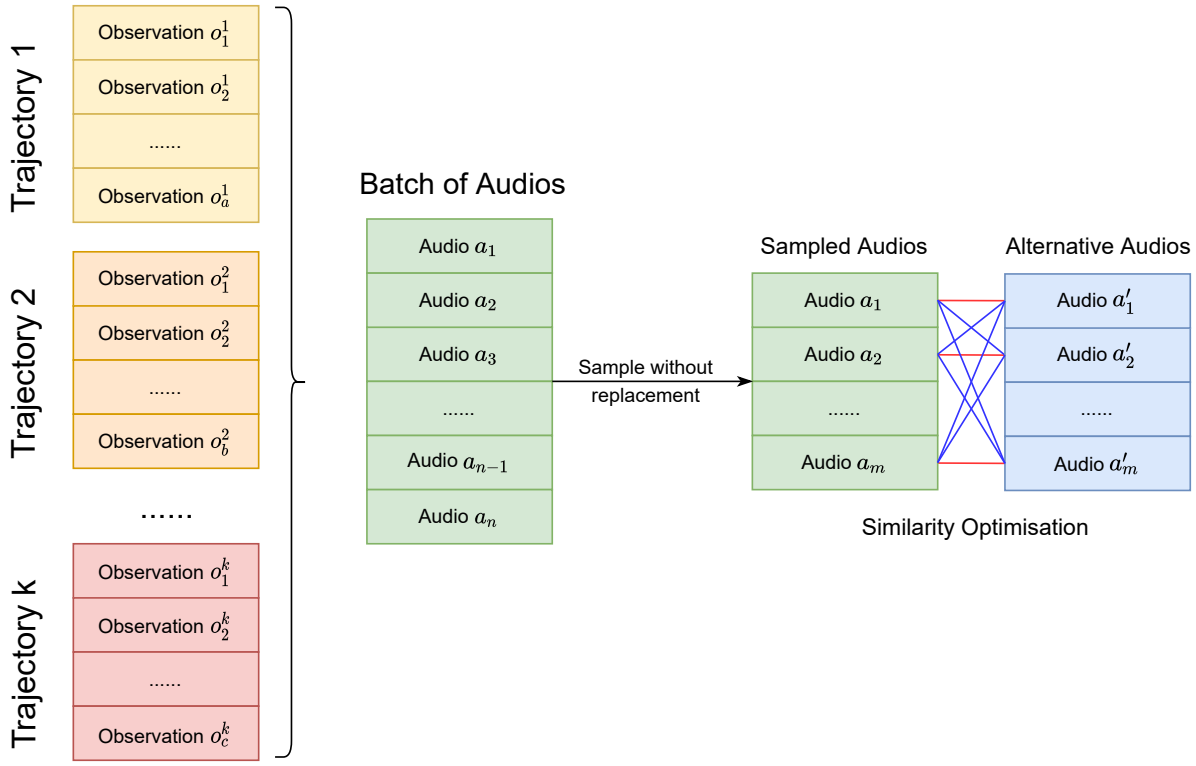


FIGURE 3.9. The graphical illustration of the batch sampling strategy. The left part shows that the RL framework works by collecting trajectories for batch gradient accumulation, o_c^k denotes the observation at step c from the k-th trajectory. The total number of audio observations from those trajectories is n, and the right part shows how we sample m audio to form pairs for AFSO.

Besides removing large proportions of the false-negative pairs, the batch sampling strategy further facilitates training by removing negative audio pairs made up of audios that imply not identical but still similar source-receiver spatial relationships. For example, if the audio source is far from the agent receiver, the source-receiver spatial relationships implied by the audio will still be partly similar between the consecutive audio signals in the trajectory when the agent takes an action 'move forward'. By sampling a small proportion of the audio observations, we implicitly enlarge the distance between the consecutive movements in the trajectories, hence avoiding minimising the feature agreement between pairs of audio observations that indicate similar audio goal information. In the meantime, the batch sampling strategy also substantially reduces the computational cost of forwarding audio signals and backpropagating errors through the encoder.

3.5 Source Sound Augmentation

One of the major reasons why existing AVN frameworks generalise poorly on unheard sounds is that the number of training sounds is extremely limited, only less than 100 different sounds are used to train the intelligent agent. As a result, there are substantial gaps in the distribution between the training, validation, and test splits, and it has crucially reduced the performance of the agent.

However, it is worth noting that the type of source sound the agent heard is irrelevant to the AVN task, as the agent should always navigate to the audio-goal position with louder volumes. This characteristic implies that arbitrary audio signals can be used as the simulation source sounds, as all the necessary information for AVN is provided in the RIR instead of the source sounds. Nonetheless, the specific source sounds to be rendered by RIR do need to share a similar volume compared to other sounds in this dataset. Therefore, random audio waves cannot be used as the source sound as they do not fit in the theoretical global distribution of the simulator sounds.

To alleviate the performance degradation caused by the highly-biased distribution of the training sounds, we propose to augment the source sounds in the training set to enrich the distribution. In detail, we propose two source sound augmentation strategies, namely audio mix-up and audio reverse. These augmentations can generate novel sounds that do not exist in the training set while still lying within the sound distribution of the SoundSpaces simulator.

3.5.1 Audio Mix-up

We generate novel sound sources by selecting and mixing two (potentially reversed) sounds randomly selected from the available training sounds. In detail, given two source audio s_1, s_2 as arrays of length equal to the sampling rate, the mixed new source audio is calculated as:

$$S_m = \lambda S_1 + (1 - \lambda) S_2, \tag{3.3}$$

where λ is a scalar sampled from a symmetric Beta distribution:

$$\lambda \sim Beta(\alpha, \alpha). \tag{3.4}$$

The above beta distribution is a symmetric distribution defined on the interval $[0, 1]$. For $\alpha = 1$, the beta distribution is equal to the uniform distribution in the close unit interval. For $0 < \alpha < 1$, the distribution will be skewed towards two ends, i.e., the probabilities of sampling 0 or 1 will be higher than 0.5, whereas the distribution will be more centralised in cases $1 < \alpha$.

By mixing up the training sounds, we expand the number of possible training sounds from less than one hundred to infinitely many sounds. Meanwhile, we do not try to increase the number of sounds we mixed (i.e., mixing more than two sounds), as they will not further increase the scale of the number of sounds. We believe that mix-up between pairs of sounds already resolves the overfitting issue caused by memorising limited sounds. While the models might still overfit specific patterns in sounds that cannot be removed through the mix-up approach, mixing more sounds will not help in handling this problem either.

3.5.2 Audio Reverse

To fill the gap that mix-up will not introduce any novel patterns as extra supervision to train the audio encoder, we introduce audio reverse, which can double the number of training sounds and potentially infuse new patterns to the training audios that can crucially improve the generalisation of the model. Unlike mix up, the audio reverse approach can be applied without sampling a second sound, and we directly reverse the audio array of the original sound with probability p to generate the reversed sound.

While the reversed audio signals might lack substantial semantic meanings, it does not change the general volume of the sounds. Therefore, the reversed sound satisfies the criteria discussed in Section 3.5 and can be used to generate novel sources that emit sound in the environment. With the novel patterns in the spectrogram, we alleviate the issue of overfitting specific patterns in training sound.

3.5.3 Sound Augmentation Module

We develop the source sound augmentation module by integrating audio mix-up and audio reverse. Figure 3.10 demonstrates an overview of the source sound augmentation pipeline.

At each episode, we generate the source sound of the navigation target for this episode with the source sound augmentations module. Meanwhile, we also augment the alternative sound for AFSO. While the source sound needs to be consistent throughout different steps in the same episode, we simulate a different alternative sound with the sound augmentation pipeline at each step during the navigation

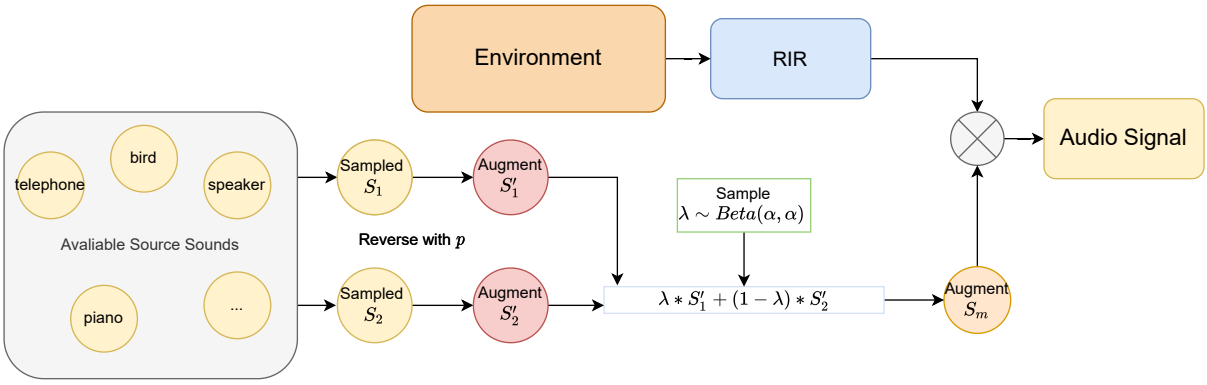


FIGURE 3.10. The sound augmentation pipeline combining audio mix-up and audio reverse approaches.

episode. The AFSO module is then plugged into the baseline AVN frameworks as depicted in Figure 3.8.

3.6 Waypoint Prediction

The features from the visual mapping branch, acoustic mapping branch, and the AFSO-based audio encoding branch are concatenated and fed into a GRU [20] together with the previous state descriptor h_{t-1}. The GRU will combine the observations at the current step and the state descriptor from the previous step, to output the descriptor on the navigation state at the current step h_t. The actor-critic RL architecture is then applied on S_t to optimise the downstream encoders, as well as predict the final output for navigation.

Following [13], we predict a probability distribution W_t of possible optimal waypoints in a local grid of (9×9) as intermediate sub-goals toward the audio goal. After masking out the occupied (hence un-navigable) nodes in W_t based on the geometrical map G_t, a waypoint $w_t = (\Delta x, \Delta y)$ is sampled from the grid based on the predicted 9×9 probability distribution W_t. The waypoint is then passed to the path planning module to generate the sequence of low-level navigation actions to reach the waypoint.

According to Chen et al. [13], the waypoint mechanism allows the agent to dynamically adjust the intermediate navigation goals based on current observations, as opposed to other methods where the agents suffer from myopic step-planning or fixed final goal prediction.

3.7 Path Planning

We implement the path planning module based on the AV-WaN framework [13]. Given the predicted waypoint w_t, the path planning module attempt to generate a sequence of low-level navigation actions to guide the agent moving to the intermediate goal specified by the waypoint. Based on the global occupancy map G_t, the planner module estimates a path from the agent to w_t using Dijkstra's algorithm. As the global map G_t contains non-integer values that are introduced by averaging operations between G_{t-1} and G_local, we consider cells with values above 0.5 as occupied or explored to avoid ambiguity. After the path has been estimated and decomposed into low-level commands, the agent will follow the actuation commands to reach to waypoint. With the deterministic waypoint navigation, the agent can effectively avoid colliding with obstacles if the predicted waypoint is reachable.

As the actor-network might predict navigation waypoints that are not reachable from the current position based on current G_t, the agent will execute a random action in cases where no valid paths are found. Meanwhile, the agent will only execute the first ten planned actions before re-planning the waypoints, to mitigate waypoints that are hard to reach or unreasonably predicted. Exceptionally, the agent will execute stop to end the current episode if the predicted waypoint $w_t = (0, 0)$.

CHAPTER 4

Experiments and Results

In this chapter, we evaluate our AV-GeN framework against two state-of-the-art AVN models, AV-NAV [12] and AV-WaN [13]. Following the SoundSpaces public AVN challenge hosted on the Matterport3D dataset, we evaluate our method and conduct the ablation studies on the Matterport3D dataset due to its abundant diversity and enormous scale of environments.

First, we introduce the experiment settings we used to evaluate our AV-GeN framework. Experiment details including task settings, evaluation metrics, 3D environment dataset used, and hyper-parameters used in our implementation will be described in the following sections. Then we present the quantitative and qualitative comparisons of the navigation results of different models to demonstrate the superiority of our AV-GeN framework. Afterwards, we show the results of several ablation studies to analyse the proposed AFSO and sound augmentation methods. In addition, we adapt the AV-GeN framework on a small-scale AVN dataset, Replica, and report the comparative navigation results to show that our framework can consistently outperform others in different environments.

4.1 Task Definition

While our AV-GeN framework can work under different AVN task settings, we focus on the most popular Audio-Goal Navigation task setting following the SoundSpaces challenge [12].

Task definition. At each step, the agent receives the binaural audio signal from the sounding object, indicating the target position of navigation. As various obstacles such as furniture and walls exist in the environment, the agent needs to get around the obstacles and approach the target position.

Agent and goal embodiment. Following the standard cylinder embodiment used in Habitat [64], the target has a height of 1.5m and a diameter of 0.2m. It should be noticed that the audio goal has no visual presence in the environment due to the limitation of the simulator, i.e., the agent might hear the

telephone ringing right at the front without seeing a telephone. The vision senses are only essential for the detection and avoidance of obstacles under the current task setting.

Action space. As mentioned in the last section, the SoundSpaces simulator maintains a navigability graph of the indoor scenes, where each node represents the physical placement of the agent. The agent is only allowed to move to the node in front of it if they are connected [12]. To summarise, the action space A of the agent is defined by a set of four actions: {*MoveForward*, *TurnLeft*, *TurnRight*, and *Stop* }. The physical position of the agent will not change if the agent executes *MoveForward* behind the obstacles.

Episode specification. An episode of AudioGoal is defined by an arbitrary 1) 3D scene, 2) agent start position, 3) agent start rotation, 4) audio goal position, and 5) source audio waveform. An episode is successful if the agent executes the *Stop* action at the position of the audio goal. Conversely, The agent fails if it executes the *Stop* action at incorrect positions, or if an upper bound of 500 number of actions is reached.

4.2 Dataset

Following the previous works and baselines, we implement our method using the SoundSpaces audio-visual platform. It augments the Habitat simulator [64], which is an open-source 3D simulator that offers fast rendering for RGB, depth, and semantic observations. The visual renderings for our experiment are adapted from the Matterport3D [9] dataset shown in Figure 4.1. It contains 85 real-world in-floor environments with 3D meshes and image scans, with an average floor space of $517m^2$. With the Habitat Simulator and Matterport3D environment, real-time rendering can be simulated flexibly by specifying camera poses, and the agent can move around in the environment while receiving egocentric visual observations [13].

The SoundSpaces simulator loads source sounds with a sampling rate of 16000Hz and provides acoustic renderings by pre-computing binaural audio RIRs at a discrete set of visitable navigation nodes. Specifically, let $\mathcal{S} = \{(x_i^s, y_i^s)\}_{i=1}^N$ denote the set of N possible sound-emitting positions, and let $\mathcal{L} = \{(x_l^s, y_l^s)\}_{i=1}^N$ denote the set of N possible positions of the listener (agent). The N locations are densely sampled as a grid with a spatial resolution of 1m, resulting in $N \in [20, 2103]$ for different scenes in the Matterport3D dataset. The sounding points are placed at a vertical height of 1.5m following the fixed height of the robot agent.

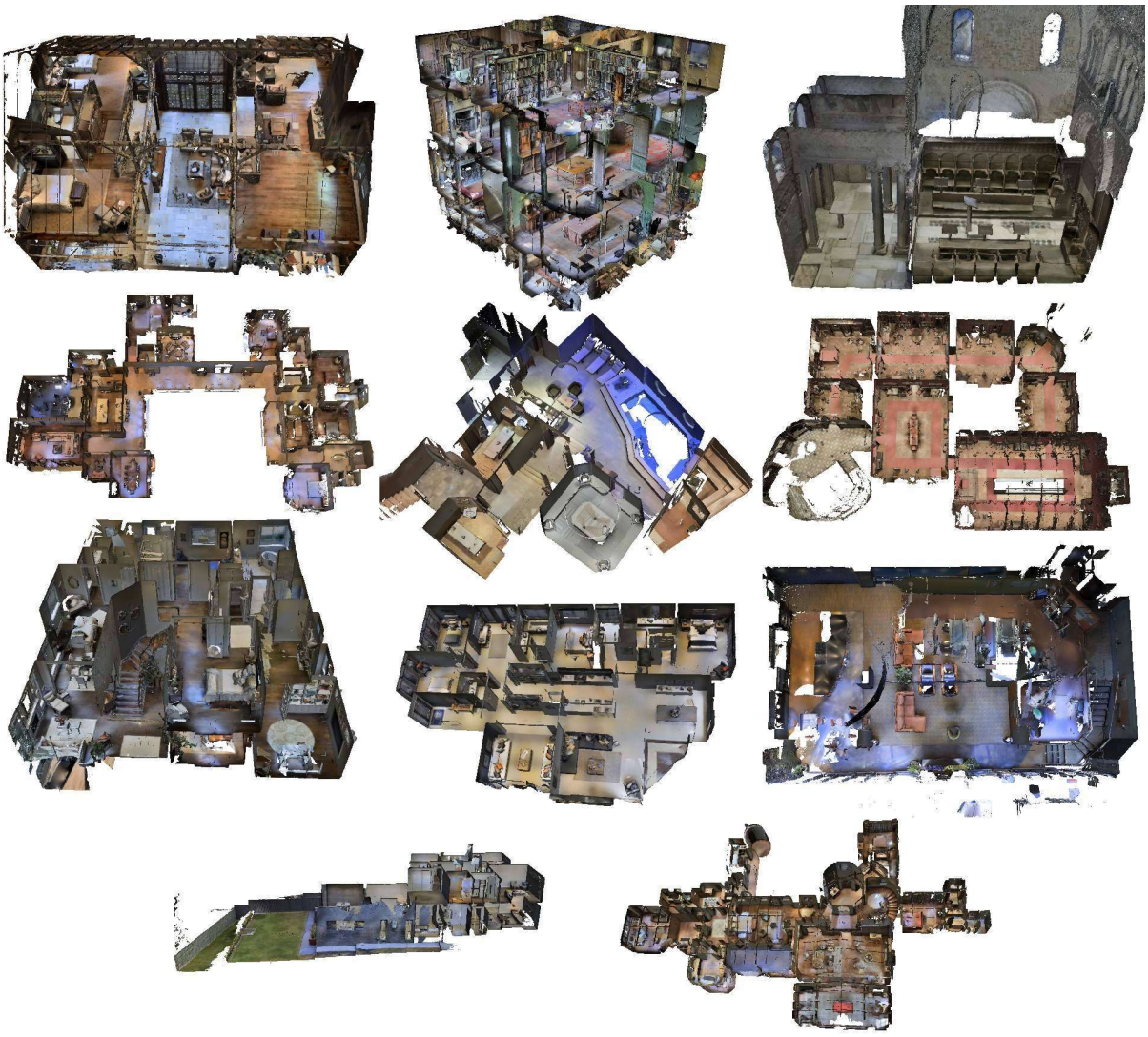


FIGURE 4.1. Exemplary visualisations of the scenes in the Matterport3D dataset [9].

Following the protocol of the SoundSpaces AudioGoal benchmark, we test the methods with train/-val/test splits of 59/10/11 disjoint scenes on the Matterport3D dataset. Meanwhile, we follow the official split of 73/11/18 non-overlapping sound for the sound sources.

While the previous works have additionally evaluated their methods on unseen scenes with one identical training and testing sound to study the generalisation ability of their path planning strategies, we do not experiment in such a setting as our methods focus on learning audio representations and do not involve path planning.

4.3 Metrics

The agent needs to approach and stop at the audio-goal position accurately and efficiently. Following AV-WaN [13], we evaluate all methods with the following metrics that are commonly used in the navigation literature [2]:

- Success Rate (SR): The overall percentage of episodes that the agent successfully stops at the audio-goal position.
- Success Weighted by Number of Actions (SNA): The success status times the minimum number of actions over the agent number of actions. The success status is represented as a boolean value of either 0 or 1 and NA represents the number of actions used:

$$\text{SNA} = success * \frac{\text{NA}_{min}}{\max(\text{NA}_{min}, \text{NA}_{agent})}. \tag{4.1}$$

- Success Weighted by Path Length (SPL): The success status times the shortest path length over the path length (PL) the agent achieves. Compared to SNA, SPL does not penalise the agent for rotation in place actions:

$$\text{SPL} = success * \frac{\text{PL}_{min}}{\max(\text{PL}_{min}, \text{PL}_{agent})}. \tag{4.2}$$

Following AV-NAV [12] and AV-WaN [13], we consider SPL the primary measurement, whereas SR and SNA will be used as secondary measurements. For all metrics, the higher values imply better navigation performance. Meanwhile, the metrics values will be presented in percentage in the following sections for better readability.

4.4 Implementation

We reproduce the AV-NAV [12] and AV-WaN [13] framework with default hyperparameters specified in the papers. Meanwhile, for the proposed AV-GeN framework, we adapt RL parameters similar to AV-WaN for fair comparisons. In detail, we use an initial learning rate of $2.5e-4$ with Adam optimiser [43] and decay it linearly through updates.

The network is trained for 7.5 million policy prediction steps with Proximal Policy Optimization (PPO) [66], with an entropy loss on the policy distribution with a coefficient of 0.02. The agent is rewarded based on the following set of rules:

- +10 for executing Stop at the goal location.
- +0.25 for reducing the geodesic distance to the goal.
- -0.25 for increasing the geodesic distance to the goal.
- -0.01 per time step.

For our novel AFSO module and sound augmentation strategies, we set the loss weight factor w as 0.1, the temperature τ as 0.07, the audio batch N as 256, the reverse probability p as 0.5, and the mix-up factor α as 1. We use a 2 layer MLP with 256 hidden units and the ReLU activation as the projection head g for similarity optimisation.

4.5 Quantitative Comparisons of Navigation Results

	SPL% ↑	SR% ↑	SNA% ↑
AV-NAV [12]	26.3	43.6	11.8
AV-WaN [13]	36.2	57.4	27.4
AV-GeN (Ours)	**48.4**	**73.9**	**37**

TABLE 4.1. Quantitative comparisons with SOTA methods on audio-visual navigation in the Matterport3D environments.

We compare the proposed AV-GeN framework with two state-of-the-art AVN models. As presented in Table 4.1, the AV-GeN framework significantly outperforms existing frameworks by a large boundary. It can be observed that AV-GeN produces improvements of 22.1%, 30.3%, and 25.2% in SPL, SR, and SNA respectively compared to the AV-NAV baseline, where we approximately double the rate of success in navigation while using only half of the actions. Meanwhile, compared to its counterpart AV-WaN which is not equipped with the AFSO and sound augmentation module, the AV-GeN achieves 12.2%, 16.5%, and 9.6% higher performance in terms of SPL, SR, and SNA. The results prove that the AV-GeN framework can navigate toward the audio goal positions more accurately and efficiently. We have submitted the produced AV-GeN model to the SoundSpaces challenge and currently, we achieved the 1-st rank in the challenge.

While we did not observe superior performances of AV-GeN over others during the training procedure, we reckon that the increases in the test performance attribute chiefly to the successful reduction in the generalisation errors on the unfamiliar sounds. Overall, we conclude that our methods can significantly benefit the existing AVN frameworks consistently on different datasets.

4.6 Navigation Results Visualisations

In this section, we visualise the top-down view of navigation trajectories on the test splits produced by the baselines and our methods. The shortest geodesic path is shown in green, and the agent path fades from dark blue to light blue during navigation. We randomly sample three navigation episodes, and the corresponding navigation visualisations are demonstrated in Figures 4.2, Figure 4.3, and Figure 4.4.

In Figure 4.2, it can be observed that the baseline AV-NAV agent kept oscillating around the starting position, and it failed to approach the audio. Meanwhile, although the AV-WaN agent successfully reached the audio goal, it took a long detour around the environment before it found its way to the target. On the other hand, while our AV-GeN agent was also distracted and explored the environments at the left, it soon realised the correct path and navigated towards the audio source using actions much less than the AV-WaN baseline.

Figure 4.3 shows a navigation episode in a different environment. In this case, all three agents reached the goal position successfully. However, it can be seen that the AV-NAV and AV-WaN agents could not accurately interpret the position of the goals, where these agents moved away from the audio goal during the navigation. Conversely, most of the actions the AV-GeN agent took reduced its distance to the audio goal, and these actions formed a path more similar to the optimal path. As a result, the proposed AV-GeN model produced a navigation trajectory with a shorter path and fewer actions, indicating its superior ability to interpret the received audio signals.

Figure 4.4 demonstrates trajectories on a challenging test episode. Both the AV-NAV and AV-WaN agents failed to reach the audio goal. Specifically, the AV-NAV agent managed to find its way to the audio source, but could not decide where is the correct position to stop, while the AV-WaN agent was trapped at the narrow corner and could not analyse the directions of the audio source. On the other hand, the AV-GeN agent demonstrated a more powerful ability to understand the relative position of the goal from the audio signal, and the agent planned its way to the target position efficiently with a relatively short path. Overall, the trajectories provide a clearer view of the superiority of the agents trained with our methods. And we conclude that our methods substantially improve the agent at navigation toward unfamiliar audio goals.

(A) AV-NAV

(B) AV-WaN

(C) AV-GeN

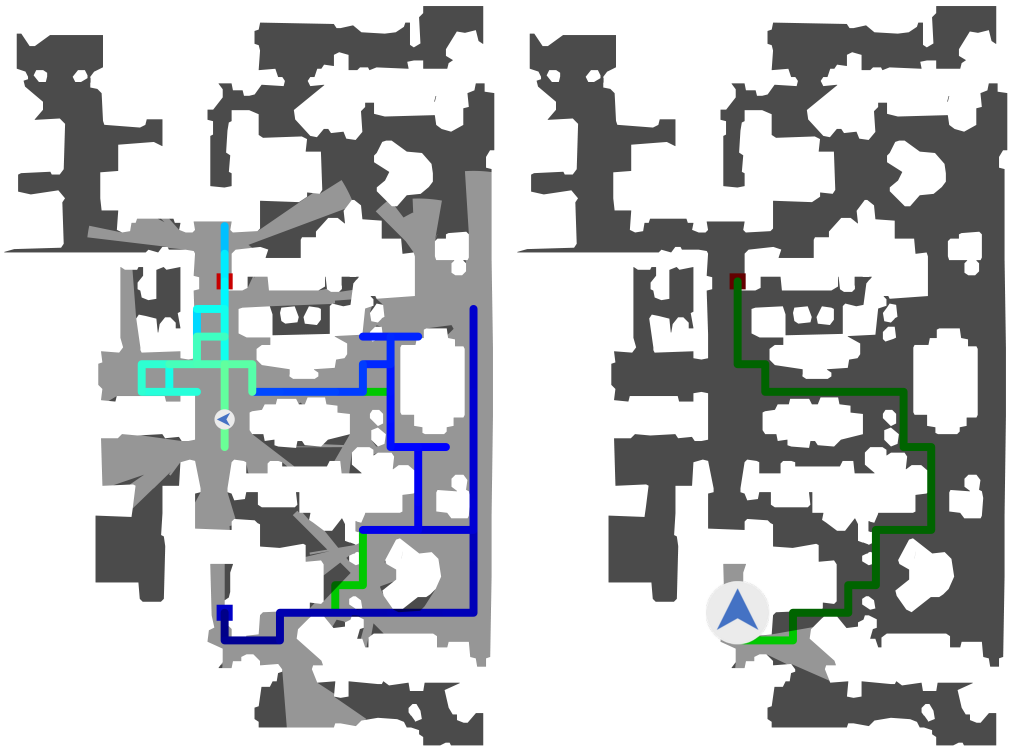
FIGURE 4.2. Navigation trajectories in top-down views of all methods in environment A.

(A) AV-NAV

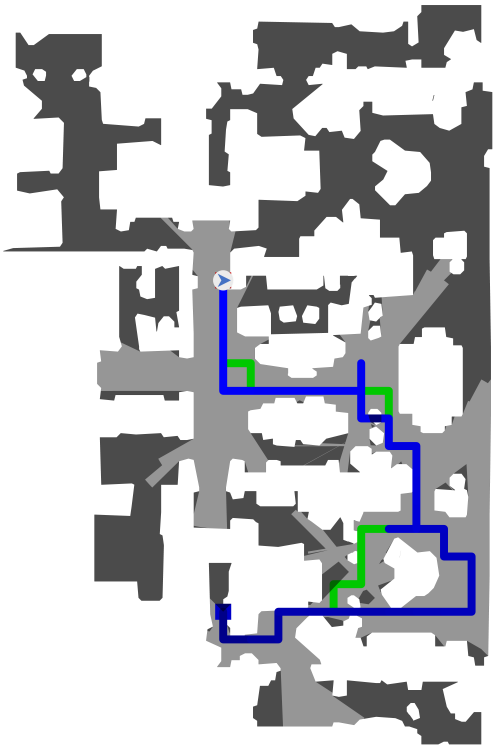
(B) AV-WaN

(C) AV-GeN

FIGURE 4.3. Navigation trajectories in top-down views of all methods in environment B.



(A) AV-NAV (B) AV-WaN



(C) AV-GeN

FIGURE 4.4. Navigation trajectories in top-down views of all methods in environment C.

4.7 Ablation Studies and Results

In this section, we conduct ablation studies to prove the effectiveness of our AFSO and sound augmentation methods. Moreover, we also experiment with the designed components of each method, to demonstrate the advantage of our designs.

In detail, we first try to implement the AV-GeN framework without the AFSO and the sound augmentation module and compare the resulting performance. Meanwhile, we study AFSO and sound augmentation separately. For AFSO, we conduct experiments to study the importance of batch sampling and projection head by conducting experiments on optimising the feature similarity on all collected audio observations without sampling (for batch sampling) and directly compute the InfoNCE loss on audio features without projecting them into another latent space (for projection head). For sound augmentation, we study the effect of applying mix-up and reverse strategies separately.

Method	AFSO		Aug		Preformance		
	BS	PH	Mix-up	Reverse	SPL% ↑	SR% ↑	SNA% ↑
AV-GeN	✓	✓	✓	✓	**48.4**	**73.9**	**37.0**
AV-GeN w/o Aug	✓	✓	-	-	43.3	66.4	33.9
AV-GeN w/o AFSO	-	-	✓	✓	39.9	68	31.0
AV-GeN w/o both	-	-	-	-	36.7	56.4	28.1
Ablations on AFSO	✓	✗	-	-	41.2	67.8	32.4
	✗	✓	-	-	41.5	65.6	31.9
	✗	✗	-	-	40.8	62.4	32.7
Ablations on Aug	-	-	✗	✓	37.0	66.2	28.3
	-	-	✓	✗	37.7	62.6	29.7

TABLE 4.2. Ablation results for AV-GeN. Aug represents the sound augmentation method, BS represents the batch sampling strategy and PH represents the projection head.

Table 4.2 presents the results of the ablation studies. We affirm the effectiveness of the proposed AFSO and sound augmentation approaches by observing an increase of 6.6% and 3.2% in SPLs by deploying these methods. Moreover, we witness a solid advantage in applying two methods collaboratively over individually. We obtain an 11.7% increase in SPL by applying both methods while the benefit of applying them individually is only $6.6 + 3.2 = 9.8\%$. While part of this could be due to the randomness and fluctuations among experiments, the results still signify that we can obtain promising profits by integrating AFSO and sound augmentation. Noticeably, the AV-GeN framework without both will have an architecture identical to the AV-WaN framework and should lead to similar results, but the evaluation

results of AV-GeN without both are slightly different from the AV-WaN results due to the randomness among runs.

The results in the middle rows of Table 4.2 also exhibit the benefit of implementing the batch sampling strategy to reduce the false-negative optimisation pairs, where models with the batch sampling mechanism consistently outperform those without it. Meanwhile, although AFSO without a projection head leads to a slightly better success rate, it takes more steps and actions for the agent to reach the target position, indicating that a projection head is overall beneficial to the AFSO method.

It could be observed from the last two rows that both mix-up and reverse are valuable strategies to improve the generalisation, and mix-up proves to be a more powerful augmentation technique than the reverse approach. We believe that the mix-up augmentation provides more benefits than the reverse because it enlarges the training distribution of possible source sounds from a limited number to infinity, while the reverse augmentation only doubles the number of source sounds.

4.8 Extensive Comparisons on Replica

In addition to Matterport3D, we also deploy the proposed AV-GeN framework to the Replica dataset, which contains indoor 3D environments of relatively small scales. It includes 18 hotel, apartment, office, and room scenes with 3D meshes and the visitable locations are sampled at a spatial resolution of $0.5m$, resulting in $N \in [38, 566]$ navigation nodes in different scenes. In addition, the SoundSpaces simulator renders the acoustic events in Replica with a sampling rate of 44100Hz. A visualisation of the environments in Replica is demonstrated in Figure 4.5, it can be seen that the environments in Replica are much simpler than Matterport3D, hence Replica is not considered as the standard evaluation dataset for the AVN task following the SoundSpaces challenge. Following [13, 12], we test the methods with train/val/test splits of 9/4/5 disjoint scenes on Replica.

The quantitative results are presented in Table 4.3. The AV-NAV with a relatively simple architecture performs slightly better than the AV-WaN on Replica, probably because the waypoints mechanisms are designed for long-term path planning and obstacle avoidance, whereas the environments in Replica are small with few obstacles. While our AV-GeN framework is developed based on AV-WaN, it still significantly outperforms both the baseline methods, with more than 10% increases in SPL, 20% increases in SR, and 10% increases in SNA for both baselines.

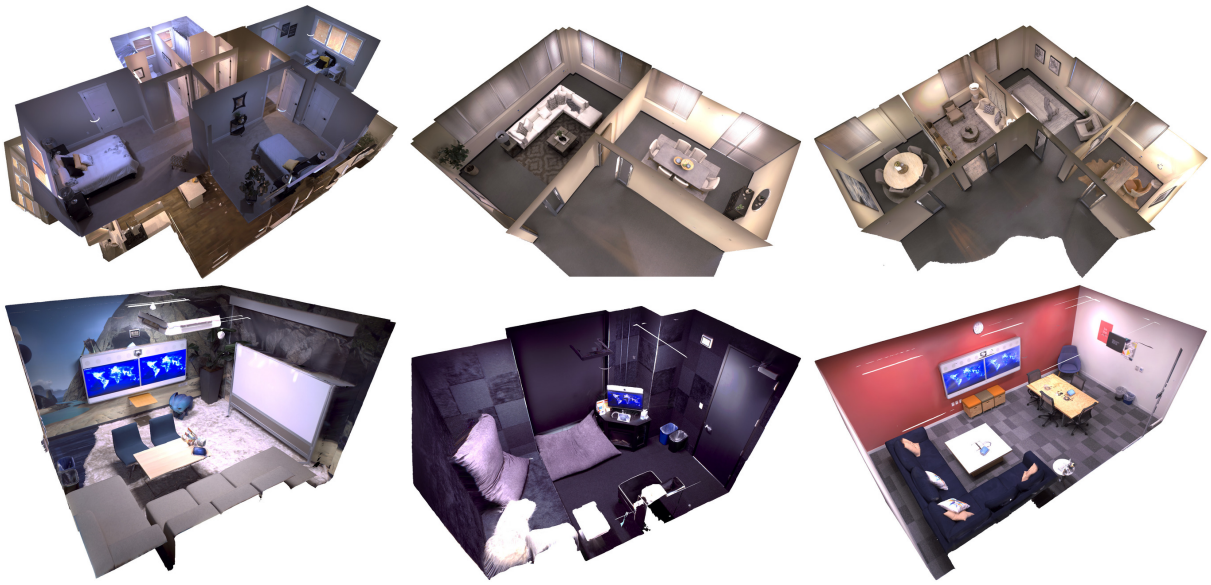


FIGURE 4.5. Visualisations of the environments in the Replica dataset. Image adapted from [68].

	SPL% ↑	SR% ↑	SNA% ↑
AV-NAV [12]	38.2	45.2	21.5
AV-WaN [13]	35.7	48.4	28.5
AV-GeN (Ours)	**49.1**	**69.8**	**38.6**

TABLE 4.3. Quantitative comparisons with SOTA methods on audio-visual navigation in the Replica 3D environments.

While we observed that the AV-GeN framework brings more advancement in the performance to its AV-WaN counterpart in Replica than in Matterport3D, we believe that the higher audio sampling rate used in the Replica dataset is the main reason for the higher increase in performance. According to [12], the higher audio sampling rate used in Replica enables the audio signals received at each step to carry more information. As a result, the input spectrograms in Replica are also more complex, and larger encoders were used in baselines to encode the audio observations. With the complex spectrograms that potentially carry more meaningful features for the navigation, the agent for Replica also suffers more from overfitting with a more heavily parameterised audio encoder. With our AFSO method and augmentation techniques, models on Replica benefit more from the alleviation of the overfitting problem, which reduces the generalisation error.

CHAPTER 5

Discussion

5.1 Designs of the AV-GeN Framework

The proposed AV-GeN framework substantially improves the generalisation of AVN agents by combining the proposed AFSO and sound augmentation methods with the AV-WaN framework. Compared to existing AVN frameworks, AV-GeN demonstrates dominant performance on navigation towards unfamiliar audio goals. Meanwhile, compared to methods that improve the AVN agent by reformulating the task definitions such as [79], AV-GeN does not require altering the task definition or any substantial modifications to the environment.

Moreover, it is worth noting that the proposed AFSO and sound augmentation method is not coupled with the AV-GeN framework. The designed AFSO-based audio encoding module and sound augmentation module can be decoupled from the AV-GeN framework and deployed flexibly to arbitrary AVN frameworks with a learning-based audio encoding scheme. For example, given the fact that AV-NAV outperforms AV-WaN on the Replica dataset, we can implement a variant of the AV-GeN framework by adapting the path planning modules in AV-NAV, which will lead to even better performance on Replica compared to the current AV-GeN design.

Nonetheless, the current AV-GeN framework might be further enhanced in terms of generalisable goal-driven representations. For example, the acoustic map in the current framework is built with intensity scalars estimated with the root mean square of the input audio signal. While such an estimation method can derive approximations of the intensity descriptor, the resulting intensity map cannot always address the goal orientation accurately due to noise in the intensity approximation. Learning-based methods should be explored to construct acoustic maps with more meaningful descriptors. Furthermore, the audio representation learning problem in AVN can be formulated as a few-shot learning (FSL) problem [26], where the encoder model needs to generalise from a few training sounds to predict the positions of

the audio goal implicitly. In this case, methods from the FSL literature can be studied and transferred to the AVN task. These directions could be explored in the future towards building AVN frameworks with better generalisation on the ability to navigate to unfamiliar audio goals.

5.2 AFSO Method

5.2.1 Implications Behind the Contrastive Optimisation

It can be observed from the optimisation method detailed in Algorithm 1 that from the algorithmic view, the AFSO method is similar to the well-known SimCLR contrastive learning framework [18]. While the SimCLR samples a positive pair of data by applying different sets of augmentations on the same input image, our AFSO method generates positive pairs of audio signals by simulating audio observations with different sounds but the same binaural RIRs. After obtaining the two batches of elements where the element at the same index corresponds to a positive pair, our AFSO is implemented in the same way as the SimCLR does. Overall, the audio observation pairs rendered with the same RIR but different sounds are equivalent to the differently augmented image pairs in SimCLR. While SimCLR is designed for self-supervised representation learning, which works by obtaining supervision in a contrastive style, the audio encoder is trained with an RL algorithm and does not lack supervision. Therefore, it might be un-intuitive why AFSO can help to improve navigation performance. Here we discuss the theoretical groundings of why the AFSO method can improve the generalisation.

The audio observations are simulated by convolving RIRs with source sounds. Among the two elements that determine the audio observations, the RIRs were computed by the simulator based on room geometry to summarise the propagations and reflections of sounds, which includes all necessary information for navigation. On the other hand, the source sounds are randomly chosen from the training set, and what class of sound the agent hears is irrelevant to the navigation planning. From this point of view, the semantic classes subject to different target sounds can be regarded as noises, while the RIRs are the only essential clues that determine the necessary navigation actions.

With the above theory, the generalisation of the agent on various audios can be enhanced by highlighting the RIR-related information while suppressing the audio-specific information in the binaural audio observation. Our AFSO method can successfully improve the generalisation of the audio encoder using a contrastive learning method because it manages to suppress the audio-specific features. And this explains why the generalisation error can be reduced by leveraging a contrastive learning approach.

The theory behind highlighting RIR-related signals also brings us to an interesting idea: Can we explicitly emphasise the RIR information by reconstructing the RIR signal and using it as the input to the audio encoder? Theoretically, if the RIR signals can be recovered, the agent should be able to achieve perfect generalisation. However, Reconstructing the input signals of the convolution operation based on the outputs is an ill-posed question with non-stationary solutions. Machine learning models cannot estimate the RIR signals reliably without knowing the source sound, just like humans cannot tell the relative positions of the sound without knowing the source sound. Nonetheless, humans can estimate the source sound by moving around and observing the changes in the sound signals. Similarly, CNN-based models might be capable of estimating the source sound and RIRs based on the experience of the heard audio signals at different positions. Putting it all together, we believe that estimating RIRs based on navigation experience can be a promising study to be carried out in the future.

5.2.2 Designs of Batch Sampling

Through ablation studies, we have found that the batch sampling strategy is crucial for the success of the AFSO-based audio encoding module. It reduces the amount of false-positive pairs in the AFSO method and increases the discrimination among the batch elements. Moreover, we find that the batch size for batch sampling should be selected adaptively based on the scale of the environment. Specifically, we find that it is more beneficial to use smaller audio batch sizes for the AFSO method on Replica than Matterport3D. We attribute this behaviour to the larger proportions of false-negative pairs in the navigation trajectories collected in the Replica environment.

According to the SoundSpaces simulator, the distance between adjacent navigation nodes is 0.5m on Replica and 1m on Matterport3D, indicating that audio observations between consecutive steps are less distinguishable on Replica than on Matterport3D. Meanwhile, the RIRs derived based on the room layout on Replica are less diverse than Matterport3D due to the limited floor-space space and scene homogenisation (For example, 5 out of the 18 scenes in Replica describe apartments of the same structures and layouts but decorated with different furniture). While these scenes provide distinct visual observations, the propagation and reflections of audio signals are essentially the same among these apartments. Constraint by the small floor-space and geometry homogenisation, Replica has a more compact distribution of RIRs where pairs of audio observations with analogous source-receiver spatial relationships are more likely to be sampled from different scenes, hence forming false-negative pairs. As the batch sampling mechanism aims to implicitly increase the heterogeneity between sampled audio signals to remove the

false-negative pairs in AFSO, a smaller audio batch size could be more suitable for the Replica dataset where larger proportions of false-negative audio pairs exist. As a result, we conclude that when applying the AV-GeN framework to different environments, the batch size used for batch sampling should be selected with values proportional to the scale of the deployed environment.

Besides the experimental findings, we acknowledge that several aspects remain to be studied regarding the designs of batch sampling. Firstly, the removal of audio pairs in the batch sampling strategy depends on random selections, and it cannot guarantee that the false-negative pairs in the audio batch will be removed precisely. While the current design of batch sampling can be adapted efficiently and conveniently, it is worth thinking about whether there exist more rigorous ways of removing false-negative pairs based on navigation status. Furthermore, can we avoid the deficiencies caused by simulating the alternative audio observations, and can we implement the AFSO by selecting the positive and negative pairs incrementally instead of removing undesirable pairs decreasingly? The solution to this problem might provide us with a clearer view of the effectiveness of the AFSO method.

5.2.3 Designs of Projection Head

We implement the InfoNCE loss following [18], where a projection head is used to project the audio features to a more compact latent space for similarity optimisation. According to [18], the projection head is an indispensable component as it prevents dimension collapse where learned representations trivially converge to constant solutions. However, Table 4.2 has illustrated that the removal of the projection head will not affect the performance of the AFSO module dreadfully. We reckon the reason could be that the audio encoder is also supervised by the actor-network and are potent against dimension collapse. The decreased importance of projection allows us to explore other solutions to facilitate the contrastive learning-based similarity optimisation. For example, we may try to optimise the similarity on sub-vectors of the feature following [42]. In this way, we might implement the AFSO module with comparable performance without training an additional projection head with extra parameters.

5.3 Source Sound Augmentation

The experiment results have shown that source sound augmentation is a simple but effective strategy for improving generalisation. Different from existing augmentation methods such as [8] in the audio processing literature, our approaches do not need to augment the corresponding audio labels. Meanwhile,

compared to [79] which also applied audio augmentations to the AVN task, our methods directly augment the source sound in the environment for more diverse patterns in the emitted sound signals, instead of augmenting the observed audio spectrogram at each step with crops or masks [79], which potentially leads to losses of critical information.

An absorbing empirical finding is that the AV-NAV and AV-WaN agents start to overfit the training sounds halfway through the training process, resulting in a mountain-shaped validation curve. Conversely, the AV-GeN agent equipped with a sound augmentation module will be able to improve its performance on the validation set over the whole training process (On the other hand, such climbing learning curves will not be observed on AV-GeN agents trained without augmentation). The observation implies that the source sound augmentation method reduces overfitting by introducing novel source sounds every episode. As a result, the agent can keep learning and improving without memorising specific sound patterns. Therefore, agents tend to perform better when training with more episodes, which is a critical reason why the AV-GeN agent can beat the other models.

If we extend the idea of the source sound augmentation method, it will inevitably come to our mind that we could directly synthesise source sound to introduce infinitely many novel acoustic patterns to train the AVN agents. While we admit that synthesising fake sounds could be a more powerful strategy to augment the training sounds, we need to design, implement, and train extra deep models. Meanwhile, using synthesised sounds might contradict the intuition of the AVN task, where the agent should learn from **limited** sounds and generalise to others. On the contrary, the proposed source sound augmentation method can be implemented with cheap computational cost, with all training sounds still originating from the training set.

5.4 Limitations in the SoundSpaces AVN Simulator

During research on the AVN task, we have identified a few limitations in the SoundSpaces acoustic simulator. Firstly, the data distribution varies significantly between validation and test splits, due to the small amounts of scenes and sounds in these splits. Through practical experience, we found that a model achieving an SPL of s on the evaluation split could potentially be evaluated to any SPL values lying in the interval $(s-3.5,\ s+3.5)$ on the test split. As a result, evaluating the model checkpoint with superior validation performance will not often lead to optimal performance value on the test splits. Consequently, conducting the same experiment several times could result in numerically unstable results, hence many

of our experiment results might not accurately reflect the quantitative strength and weaknesses among different methods. Admittedly, we could have reduced the variance among the runs by repeating the experiments several times and reporting the average performance. However, each experiment run takes around three days on average, making it less feasible to derive more reliable results due to limitations on available devices and time.

Another limitation we identify is that the simulator does not simulate the event of receiving acoustic signals realistically. In real life, after an acoustic event occurs, it takes time for the signal to propagate through the environment and reaches the receiver. Such sound travelling periods are usually unobservable in our life as we will not be aware of the occurrence time of the sound emission event before we hear it. Nevertheless, the SoundSpaces simulator permits the agent to be aware of the time when acoustic events occur since the agent starts to acquire the surrounding audio signals at the moment when the sound is emitted from the source. Consequently, the agent might 'cheat' on judging the relative distance to the goal, by calculating the sound propagation time based on the silence period before the emitted sound arrives in the recorded acoustic signal. However, such sound propagation time will not be detected by sensors in real-life applications and agents relying on these hints will undoubtedly fail to generalise in this case.

Finally, a crucial and intricate drawback in current AVN simulation is the lack of support for simulating acoustic events at a continuous scale. While visual perceptions can be rendered at arbitrary positions and angles, the acoustic events simulations are only available at the fixed set of positions in the environment. It forbids the agent from learning audio representations based on the gradients between the received audio signals at nearby locations. While the authors do not release the method to compute RIRs at arbitrary positions but only provide pre-computed RIRs at certain places, workarounds to approximate RIRs at undefined positions could be studied, e.g., linear interpolations might be used to estimate intermediate RIRs between two nodes.

CHAPTER 6

Conclusion

Embodied AI has been considered the future for developing general-purpose AI systems. Among the Embodied AI tasks, Audio-Visual Navigation (AVN) is of particular importance, where an intelligent agent needs to navigate to a constantly sound-making object in complex 3D environments based on its audio and visual perceptions. It challenges the embodied agent in multiple aspects, including look, listening, movement, exploration, reasoning, and multi-modality signals processing. However, Modern AVN frameworks failed to generalise to unfamiliar audio sources, and their performance critically degrades when tested on various unheard sounds. In this thesis, we study how to improve the generalisation of AVN agents to unfamiliar audio goals.

We proposed the Audio-Visual Generalisable Navigation (AV-GeN) framework based on two novel methods, namely the Audio Feature Similarity Optimisation (AFSO) and Source Sound Augmentation, to improve the generalisation of AVN agents on navigation with unfamiliar audio goals. More specifically, the AFSO method regularises the audio encoder with a contrastive approach, where the sound-agnostic goal-driven latent representations can be learnt from various audio signals of different classes. In addition, the source sound augmentation method enriches the set of training sounds with audio mix-up and audio reverse. Our evaluation results on the Matterport3D dataset show that the proposed AV-GeN framework performs better than the state-of-the-art AVN frameworks. Moreover, the AVN agent trained with the AV-GeN framework has achieved top-rank performance in the SoundSpaces AVN challenge. Part of this work has been summarised into a paper and submitted to the 2022 CVPR Embodied AI Workshop.

Bibliography

[1] Abhishek Kadian*, Joanne Truong*, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. 2020. Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance? *IEEE Robotics and Automation Letters (RA-L)*, 5(4):6670–6677.

[2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir R. Zamir. 2018. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference Computer Vision Patt. Recogn. (CVPR)*, volume 2.

[4] Relja Arandjelović and Andrew Zisserman. 2017. Look, listen and learn. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 609–617.

[5] Relja Arandjelović and Andrew Zisserman. 2018. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[6] Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiau-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. 2021. Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

[7] Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. 2016. Deepmind lab. *arXiv preprint arXiv:1612.03801*.

[8] Léo Cances, Etienne Labbé, and Thomas Pellegrini. 2021. Improving deep-learning-based semi-supervised audio tagging with mixup. *arXiv preprint arXiv:2102.08183*.

[9] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *Proceedings of the International Conference on 3D Vision (3DV)*, pages 667–676.

[10] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. 2020. Object goal navigation using goal-oriented semantic exploration. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

[11] Changan Chen, Ziad Al-Halah, and Kristen Grauman. 2021. Semantic audio-visual navigation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15511–15520.

[12] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. 2020. Soundspaces: Audio-visual navigation in 3D environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 17–36.

[13] Changan Chen, Sagnik Majumder, Al-Halah Ziad, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. 2021. Learning to set waypoints for audio-visual navigation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[14] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. 2021. Learning audio-visual dereverberation. *arXiv preprint arXiv:2106.07732*.

[15] Kevin Chen, Junshen Chen, Jo Chuang, Marynel V'azquez, and Silvio Savarese. 2021. Topological planning with transformers for vision-and-language navigation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11271–11281.

[16] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

[17] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

[19] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. Big self-supervised models are strong semi-supervised learners. *arXiv preprint arXiv:2006.10029*.

[20] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*.

[21] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. 2020. See, hear, explore: Curiosity via audio-visual association. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14961–14972. Curran Associates, Inc.

[23] Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. 2020. Evolving graphical planner: Contextual global planning for vision-and-language navigation. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, NIPS'20. Curran Associates Inc., Red Hook, NY, USA.

[24] Jiafei Duan, Samson Yu, Tangyao Li, Huaiyu Zhu, and Cheston Tan. 2022. A survey of embodied ai: From simulators to research tasks. *Proceedings of the IEEE Transactions on Emerging Topics in Computational Intelligence (TETCI)*, 6:230–244.

[25] Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. 2019. Scene memory transformer for embodied agents in long-horizon tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 538–547.

[26] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2006. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 28:594–611.

[27] Chuang Gan, Xiaoyu Chen, Phillip Isola, Antonio Torralba, and Joshua B. Tenenbaum. 2020. Noisy agents: Self-supervised exploration by predicting auditory events. *arXiv preprint arXiv:2007.13729*.

[28] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Michael Lingelbach, Aidan Curtis, Kevin Feigelis, Daniel M. Bear, Dan Gutfreund, David Cox, Antonio Torralba, James J. DiCarlo, Joshua B. Tenenbaum, Josh H. McDermott, and Daniel L. K. Yamins. 2020. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*.

[29] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B. Tenenbaum. 2020. Look, listen, and act: Towards audio-visual embodied navigation. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707.

[30] Ruohan Gao, Changan Chen, Ziad Al-Halab, Carl Schissler, and Kristen Grauman. 2020. Visualechoes: Spatial image representation learning through echolocation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[31] Xiaofeng Gao, R. Gong, Tianmin Shu, Xu Xie, Shu Wang, and Song-Chun Zhu. 2019. Vrkitchen: an interactive 3d virtual environment for task-oriented learning. *arXiv preprint arXiv:1903.05757*.

[32] Rishabh Garg, Ruohan Gao, and Kristen Grauman. 2021. Geometry-aware multi-task learning for binaural audio generation from video. In *Proceedings of the British Machine Vision Conference (BMVC)*.

[33] Saurabh Gupta, David F. Fouhey, Sergey Levine, and Jitendra Malik. 2017. Unifying map and landmark based representations for visual navigation. *arXiv preprint arXiv:1712.08125*, abs/1712.08125.

[34] Saurabh Gupta, Varun Tolani, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. 2019. Cognitive mapping and planning for visual navigation. *Proceedings of the International Journal of Computer Vision (IJCV)*, 128:1311–1330.

[35] Simon Haykin and Zhe Chen. 2005. The cocktail party problem. *Neural Comput.*, 17(9):1875–1902.

[36] Joao F. Henriques and Andrea Vedaldi. 2018. Mapnet: An allocentric spatial memory for mapping environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8476–8484.

[37] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

[38] Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9240–9249.

[39] Jisu Hwang and Incheol Kim. 2021. Joint multimodal embedding and backtracking search in vision-and-language navigation. *Sensors*, 21:1012.

[40] Unnat Jain, Luca Weihs, Eric Kolve, Ali Farhadi, Svetlana Lazebnik, Aniruddha Kembhavi, and Alexander G. Schwing. 2020. A cordial sync: Going beyond marginal policies for multi-agent embodied tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. First two authors contributed equally.

[41] Unnat Jain, Luca Weihs, Eric Kolve, Mohammad Rastegari, Svetlana Lazebnik, Ali Farhadi, Alexander G. Schwing, and Aniruddha Kembhavi. 2019. Proceedings of the two body problem: Collaborative visual task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. First two authors contributed equally.

[42] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. 2022. Understanding dimensional collapse in contrastive self-supervised learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[43] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[44] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[45] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv preprint arXiv:1712.05474*.

[46] Vijay Konda and John Tsitsiklis. 2000. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller, editors, *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 12. MIT Press.

[47] Heinrich Kuttruff. 2019. *Room acoustics*. CRC Press.

[48] Kenneth Leidal, David F. Harwath, and James R. Glass. 2017. Learning modality-invariant representations for speech and images. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 424–429.

[49] Chengshu Li, Fei Xia, Roberto Martín-Martín, Michael Lingelbach, Sanjana Srivastava, Bokui Shen, Kent Vainio, Cem Gokmen, Gokul Dharan, Tanish Jain, Andrey Kurenkov, Karen Liu, Hyowon Gweon, Jiajun Wu, Li Fei-Fei, and Silvio Savarese. 2021. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. *arXiv preprint arXiv:2108.03272*.

[50] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

[51] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. 2021. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 275–285.

[52] Lina Mezghani, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Alahari Karteek. 2021. Memory-augmented reinforcement learning for image-goal navigation. *arXiv preprint arXiv:2101.05181*.

[53] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

[54] Rui Nian, Jinfeng Liu, and Biao Huang. 2020. A review on reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering*, 139:106886.

[55] Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc Q. K. Duong, Patrick Pérez, and Gaël Richard. 2020. Weakly supervised representation learning for audio-visual scene analysis. *Proceedings of the IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 28:416–428.

[56] Kranti K. Parida, Siddharth Srivastava, and Gaurav Sharma. 2022. Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2151–2160.

[57] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, pages 16–17.

[58] Rolf Pfeifer and Josh C. Bongard. 2006. How the body shapes the way we think - a new view on intelligence.

[59] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. *arXiv preprint arXiv:1806.07011*.

[60] S. Purushwalkam, S. Amengual Gari, V. Krishna Ithapu, C. Schissler, P. Robinson, A. Gupta, and K. Grauman. 2021. Audio-visual floorplan reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1163–1172. IEEE Computer Society, Los Alamitos, CA, USA.

[61] Claudia Pérez-D'Arpino, Can Liu, Patrick Goebel, Roberto Martín-Martín, and Silvio Savarese. 2021. Robot navigation in constrained pedestrian environments using reinforcement learning. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1140–1146.

[62] Omkar Ranadive, Grant Gasser, David Terpay, and Prem Seetharaman. 2020. Otoworld: Towards learning to separate by learning to move. *arXiv preprint arXiv:2007.06123*.

[63] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. 2018. Semi-parametric topological memory for navigation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[64] 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. 2019. Habitat: A platform for embodied ai research. *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9338–9346.

[65] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1889–1897. PMLR, Lille, France.

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

[67] David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. 2014. Deterministic policy gradient algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, page I–387–I–395. JMLR.org.

[68] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. 2019. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

[69] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *Proceedings of the International Conference Robot. Learning (CoRL)*, page 394–406.

[70] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[71] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

[72] Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. 2021. Structured scene memory for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[73] Saim Wani, Shivansh Patel, Unnat Jain, Angel X. Chang, and Manolis Savva. 2020. Multion: Benchmarking semantic map memory using multi-object navigation. *arXiv preprint arXiv:2012.03912*.

[74] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. 2021. Visual room rearrangement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[75] Yi Wu, Yuxin Wu, Aviv Tamar, Stuart J. Russell, Georgia Gkioxari, and Yuandong Tian. 2019. Bayesian relational memory for semantic visual navigation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2769–2779.

[76] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3733–3742.

[77] Fei Xia, William B. Shen, Chengshu Li, Priya Kasimbeg, Micael Edmond Tchapmi, Alexander Toshev, Roberto Martín-Martín, and Silvio Savarese. 2020. Interactive gibson benchmark: A benchmark for interactive navigation in cluttered environments. *IEEE Robotics and Automation Letters (RA-L)*, 5(2):713–720.

[78] Claudia Yan, Dipendra Misra, Andrew Bennnett, Aaron Walsman, Yonatan Bisk, and Yoav Artzi. 2019. Chalet: Cornell house agent learning environment. *arXiv preprint arXiv:1801.07357*.

[79] Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. 2021. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *arXiv preprint arXiv:2111.14843*.

[80] Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, and Xiaohong Liu. 2022. Sound adversarial audio-visual navigation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[81] Fengda Zhu, Xiwen Liang, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2021. Soon: Scenario oriented object navigation with graph-based exploration. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12684–12694.

[82] Yuke Zhu, Daniel Gordon, Eric Kolve, Dieter Fox, Li Fei-Fei, Abhinav Kumar Gupta, Roozbeh Mottaghi, and Ali Farhadi. 2017. Visual semantic planning using deep successor representations. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 483–492.

[83] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. 2016. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*.