Single-Target Audio-Visual Learning and Navigation in Search and Rescue Scenarios: Transfer Application to Physical Robot

Nathanael Oliver, Md Amanullah Kabir Tonmoy

I. INTRODUCTION

Autonomous Navigation is a pivotal advancement in the field of robotics. Its significance derives from the ability of robots to independently operate in complex and dynamic environments. This capability allows robots to transcend their traditional limitations by harnessing technologies like computer vision, machine learning, and sensor fusion. Autonomous robots can execute tasks with precision and efficiency by adapting to obstacles and planning optimal pathing. This is especially important in fields such as search and rescue (SAR) because it reduces the need for human intervention in hazardous settings.

While visual and sensor-based data have long been primary sources for navigation, integrating audio cues as input for autonomous navigation introduces a profound dimension of significance in robotics. Sound can offer crucial information about the surroundings, including the direction and proximity of objects, potential hazards, and even human presence. In the context of SAR, audio cues are extremely important in locating human destinations. Audio-based navigation has the potential to be transformative, enabling machines to navigate with a level of awareness that was previously unattainable. In this project, our goal is to create a model based on single-source audio which will be transferred to a physical robot to examine the utility of the model.

II. PROBLEM STATEMENT

The problem at hand is to train an agent to navigate to a single audio source based on audio and visual cues with no reference map and evaluate its performance in real world. Specifically, we aim to adapt and fine-tune this model using transfer learning techniques to work on a Jetson Nano, which will be controlling a simple robot. The datasets to be used for training and evaluation will primarily consist of acoustic data collected from complex indoor environments, with a focus on the Replica3D datasets.

III. RELATED WORKS

In autonomous navigation, where a robot must operate without any refence map, Active Simultaneous Localization and Mapping (ASLAM) is the popular approach to construct a map and create a path through environment based on sensor data – such as camera, lidar etc. [1] However, in complex environment this method become unreliable due to inaccuracy of the sensor data. On the other hand, computational cost for this approach is also very high, which makes this approach impractical for time-sensitive tasks [2,3]. Consequently, many alternative approaches have emerged, advocating for the

replacement of ASLAM with end-to-end policy learning [4,5,3]. End-to-end policy learning can be applied to extract semantic features such as audio, smell, vision etc. to facilitate a directed search towards a target object [6,7]. Chen et al. [8,9] explored the advantages of audio in indoor environments, proposing a multi-modal deep reinforcement learning approach to train navigation policies to locate a single sound source in an unknown environment. Their method used audio and visual cues to create a path towards the target.

Vision based navigation combined with other tasks is implemented to attain intelligent behavior such as visual recognition [10] and following instruction [11]. Audio based navigation is already employed in audio-based equipment that helps blind people to navigate and avoid any obstacle [12]. Most state-of-the-art methods in audio-visual (AV) learning primarily focused on leveraging video content rather than embodied perception. These methods heavily rely on human-captured video and give different directions for separation of sound source [13], spatializing sound [14], sound synthesis for video [15] and audio-visual tracking [16]. Chen et al. [8,9] addressed embodied navigation which relies on real-time sensor data and employed multi-modal transfer learning based on audio and visual cues. This method gives efficient path planning as both audio and visual cue help to locate the target and navigate faster by avoiding any obstacles in the path.

IV. TECHNICAL APPROACH

a) Dataset:

Replica3D [17] dataset was used to train and evaluate the model. Replica3D dataset contains a total of 18 scenarios among them 12 scenes are of different constructions (apartment, hotel, office and room) (Figure 1) and 6 are of different configurations of an apartment (Figure 2).

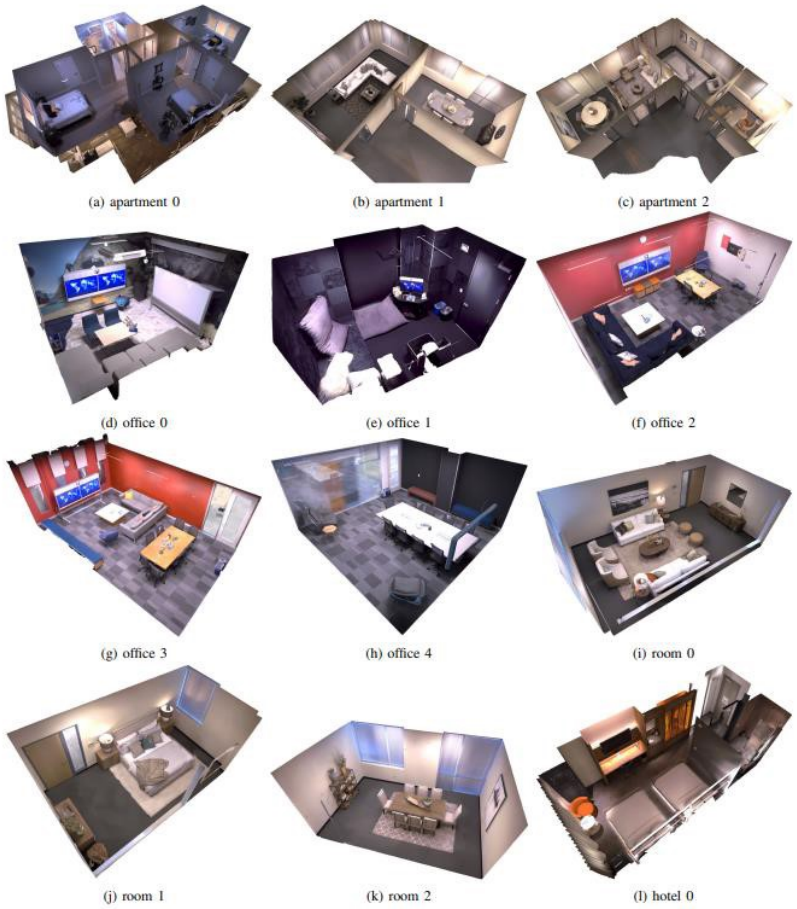

Figure 1: 12 scenes of different constructions in Replica3D dataset.

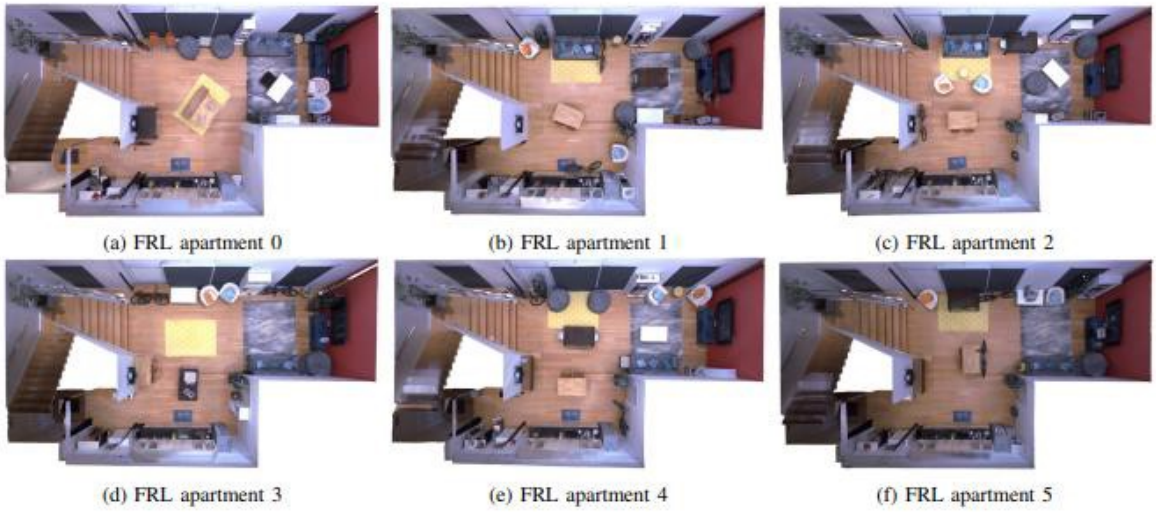

Figure 2: 6 scenes of different configurations in an apartment.

b) Simulation Setup:

Soundspaces [8,9] and Habitat Simulator [18] are used to train and evaluate the model for single source navigation. Habitat offers fast and user-friendly API to simulate the cases using RGB and depth sensors. For each scene, acoustics of the environment are simulated using room impulse responses (RIR). The RIR is the transfer function between a sound source and microphone, which varies as a function of the room geometry, materials, and the sound source location [19]. The task for the navigation agent is to reach an audio source based on audio and visual cues. To get those cues audio, RGB and depth sensors are used. The action space is: MoveForward, TurnLeft, TurnRight and Stop. If MoveForward is valid, the agent takes a step of 0.5m.

c) Training:

Figure 3 summarizes the model used by Chen et al [8,9]. RGB and Depth sensors are used to get the visual cues which go through a CNN that extracts important information from the visual cues. Similarly, audio cues are obtained from spectrogram which also passes through CNN to extract useful information from audio cues. GPS data can also be used to train the model. In this case, it was unused. The extracted information then passes through a GRU layer to accumulate history over time and then value of the state and policy distribution are estimated using critic and actor heads.

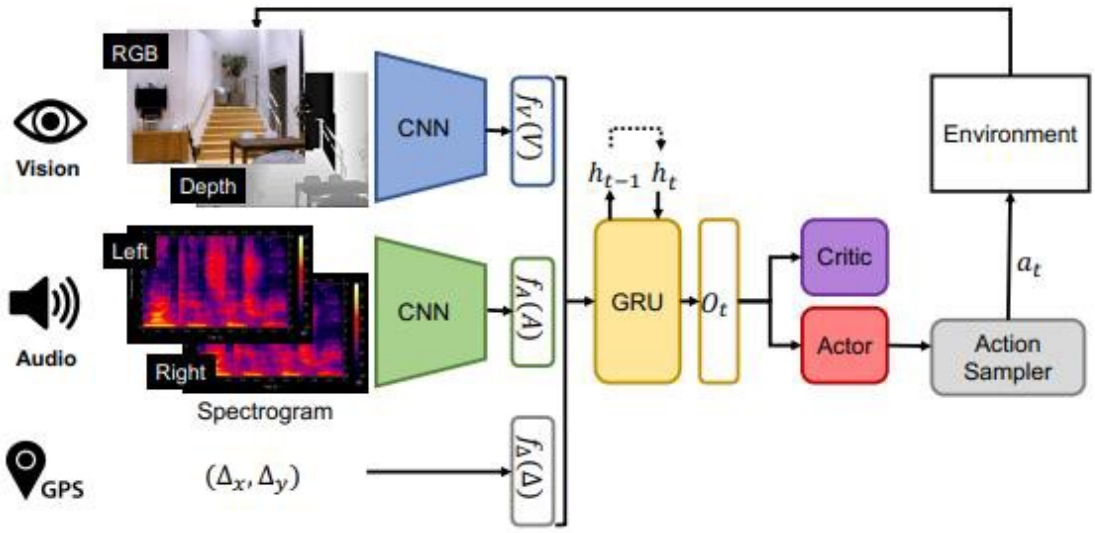

Figure 3: Model for Training.

The model is trained with Proximal Policy Optimization (PPO). The agent is rewarded for reaching the goal quickly. Specifically, it receives a reward of +10 for executing Stop at the goal location, a negative reward of −0.01 per time step, +1 for reducing the geodesic distance to the goal, and the equivalent penalty for increasing it. An entropy maximization term was added to the cumulative reward optimization, for better action space exploration as per Chen et al [8,9]. The model was trained using ADAM optimizer and the learning rate was 2.5e-4.

d) Evaluation Matrices:

The model was evaluated using Average Success and SPL (Success weighted by Normalized Path Length).

$$Average\ Success = \frac{1}{N}\sum_{i=1}^{N} S_i$$

$$Average\ SPL\ = \frac{1}{N}\sum_{i=1}^{N} \frac{S_i l_i}{\max\,(p_i, l_i)}$$

S_i is the flag whether the i-th episode is successful or not ,l_i is the shortest path distance to succeed in i-th episode and p_i is the path length traversed by agent in i-th episode.

e) Experiments:

1. Evaluate a pre-trained model from Soundspaces which was trained on a part of Replica3D dataset (office-1, room-1, apartment –2, hotel -0 and frl-apartment-1).
2. Train the model from scratch using Replica3D dataset and evaluate the performance.

Once the multi-sound source model is established, the next challenge is making it runnable on resource-constrained systems like the Jetson Nano. This step often involves simplifying the model architecture and optimizing it for efficient execution. This might include reducing the number of parameters, employing quantization techniques, and optimizing inference processes. The goal is to strike a balance between maintaining model accuracy and minimizing computational resource requirements, allowing the model to operate smoothly on hardware with limited processing power and memory. This final step is crucial for practical deployment, enabling autonomous robots or devices with constrained resources to benefit from the enhanced audio perception capabilities achieved through transfer learning.

V. RESULTS

a) Pretrained Model:

The results from a pretrained model are summarized in Table 1. The higher success rate (0.946) indicates the agent can successfully navigate to the audio source most of the time whereas the moderate SPL (0.79) indicates it does not follow the optimal path while reaching the target. The explanation can be verified from Figure 4.

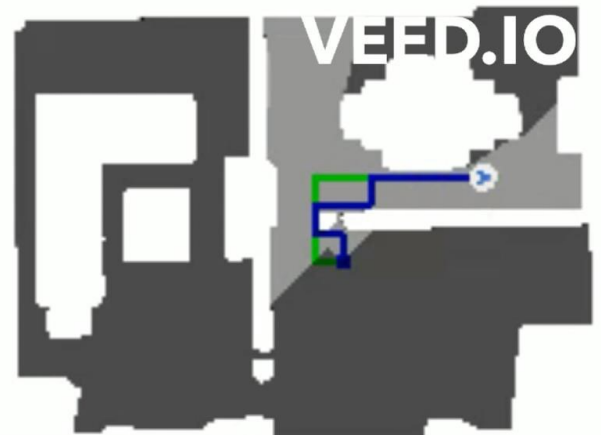

Figure 4. Evaluation of the Pre-trained model on Apartment-1. Green Line indicates optimal path and Blue Line indicates the agent's path. The Gray area is seen by the robot and the black area is unseen. The white area is occupied by obstacles.

Table 1: Summary Results

Model	Success	SPL
Pre-trained	0.946	0.79
Training from scratch	0.644	0.35

b) Training from scratch:

Results from the training from scratch are summarized in Table 1. Both success rate (0.644) and SPL (0.35) are lower than pre-trained model. The lower success means the agent cannot navigate to the target audio source properly and the lower SPL means it does not follow the optimal path most of the time (Figure 5).

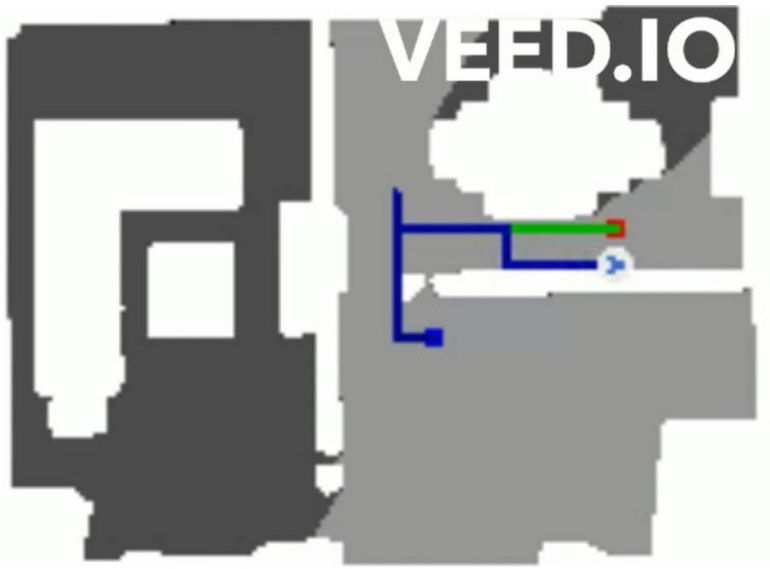

Figure 5. Evaluation of the model training from scratch on Apartment-1. Green Line indicates optimal path and Blue Line indicates the agent's path. The Gray area is seen by the robot and the black area is unseen. The white area is occupied by obstacles.

The reason behind the poor performance of the training from scratch model is the non-convergence. Figure 6 shows that the model did not reach convergence fully before stopping the training. Due to the high computational cost of training and validation of the model, only 500 updates were used to train the model. If it was run for more updates, results similar as pretrained model can be expected.

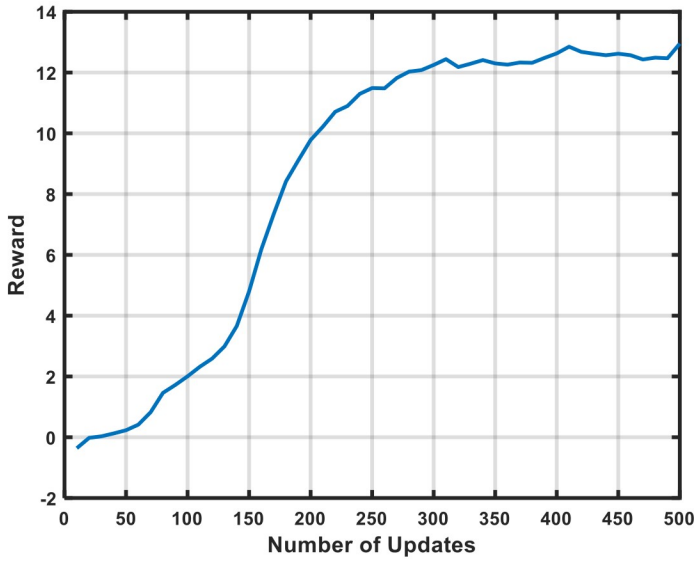

Figure 6: Reward Curve

Table-2 also shows the improvement of results after 500 updates by comparing the performance of model after 300 updates and 500 updates which further verifies our claim.

Table 2: Summary Results of Model Trained from Scratch

Model	Success	SPL
After 300 epochs	0.51	0.32
After 500 epochs	0.644	0.35

VII. CONCLUSION

The main objective of this project is to replicate the work of Chen et al. [8,9] on Replica3D dataset. Due to the computation cost, the model is run for fewer updates as a result the model did not perform as well as the pretrained model. The version control system of Soundspaces and Habitat Simulator is not user-friendly. It takes a lot of time to find and install the optimal packages to run the simulation. In future, more audio sensors can be added to train and evaluate the model. By doing this project, hands on experience on end-to-end policy learning and multi-modal training of a navigating agent to a single-source audio is obtained.

REFERENCES

[1] Iker Lluvia, Elena Lazkano, and Ander Ansuategi. “Active Mapping and Robot Exploration: A Survey”. In: Sensors 21.7 (2021). ISSN: 1424-8220.

[2] Hamid Taheri and Zhao Chun Xia. “SLAM; definition and evolution”. In: Eng. Appl. Artif. Intell. 97 (2021), p. 104032.

[3] Kai Zhu and Tao Zhang. “Deep reinforcement learning based mobile robot navigation: A review”. In: Tsinghua Science and Technology 26.5 (2021), pp. 674– 691.

[4] Lei Tai, Shaohua Li, and Ming Liu. “A deep-network solution towards model-less obstacle avoidance”. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2016, pp. 2759– 2764.

[5] Hartmut Surmann et al. “Deep Reinforcement learning for real autonomous mobile robot navigation in indoor environments”. In: ArXiv abs/2005.13857 (2020).

[6] Wei Yang et al. “Visual Semantic Navigation using Scene Priors”. In: 7th International Conference on
Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.
[7] Arsalan Mousavian et al. “Visual Representations for Semantic Target Driven Navigation”. In: 2019 International Conference on Robotics and Automation (ICRA). 2019, pp. 8846–8852.
[8] Changan Chen et al. “SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning”. In: NeurIPS 2022 Datasets and Benchmarks Track. 2022.
[9] Changan Chen et al. “SoundSpaces: Audio-Visual Navigaton in 3D Environments”. In: ECCV. 2020.
[10] Jayaraman et al. “End-to-end policy learning for active visual categorization”. In: TPAMI .2018.
[11] Anderson et al. “Vision-and-language navigation: Interpreting visually grounded navigation instructions in real environments”. In: CVPR .2018)
[12] Merabet et al. “Audio-based navigation using virtual environments: combining technology and neuroscience”. AER Journal: Research and Practice in Visual Impairment and Blindness. 2009
[13] Zhao et al. “The sound of pixels”. In: ECCV. 2018
[14] Morgado et al. “Self-supervised generation of spatial audio for 360 video”. In: NeurIPS. 2018
[15] Owens et al. “Visually indicated sounds”. In: CVPR. 2016
[16] Gebru et al. “Tracking the active speaker based on a joint audio-visual observation mode”. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 15–21. 2015
[17] Straub et al. “The replica dataset: A digital replica of indoor spaces”. arXiv preprint arXiv:1906.05797. 2019
[18] Manolis et al. “Habitat: A Platform for Embodied AI Research.” In: ICCV. 2019
[19] Kuttruff et al. “Room acoustics”. CRC Press. 2016