Semantic Audio-Visual Navigation through Distractor Silencing

Sagnik Majumder
EID: sm72878
Email: sagnik@cs.utexas.edu

Shailesh Mani Pandey
EID: smp4228
Email: shailesh.pandey@utexas.edu

Abstract*—In recent times, embodied AI has witnessed significant progress in single-source audio-visual navigation, which tasks an agent to reach the only audio source in an unknown environment by relying on audio and visual cues. In this work, we further generalize this setting and introduce the task of *semantic audio-visual navigation*, in which the agent is instead tasked to navigate to the audio source of a certain semantic class while there are other sources of different classes playing in the environment at the same time. We propose two unique approaches to solve this task, both of which rely on the principle of extracting the target audio from the mixed audio and thus effectively silencing the distractor sources for accurate and efficient navigation. One of the approaches does implicit extraction by learning disentangled latent features for navigation that are conditioned on the audio classes in the mixture. The other approach extracts the target audio more explicitly by using an additional class-conditional extraction module. We demonstrate our approaches on Replica, a challenging dataset of real-world 3D scans. Our approaches improve over the current end-to-end reinforcement learning based state-of-the-art audio-visual navigation agent that is customized to account for audio semantics, in multiple challenging evaluation settings, thus demonstrating the effectiveness of target audio extraction based agents for successful navigation in multi-audio settings. Project slides with navigation videos: https://bit.ly/semanticAudioNavigation*

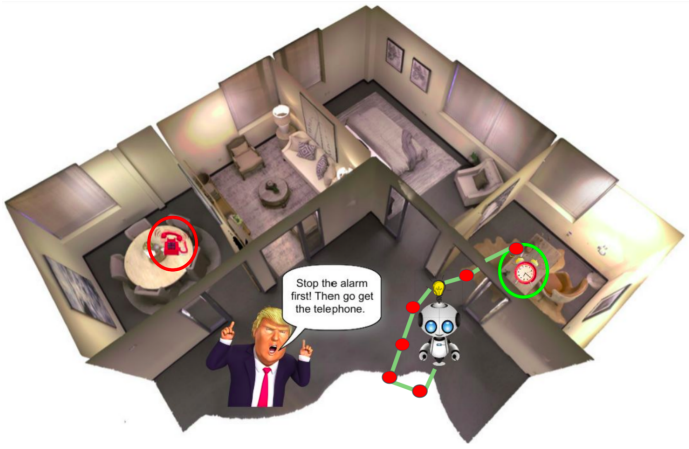


Fig. 1: **Semantic AudioGoal Navigation**. An agent is tasked with navigating to an audio source of a certain semantic class while there are other sources, each of a different class other than the target class, playing in the 3D environment. For successful navigation, the agent must strictly follow the target audio cue while shutting out the distractor audio sources.

I. Introduction

Accurate and efficient navigation to a target is often a first important step in an embodied agent's process of solving a certain downstream task. For example, a bomb detonating robot will have to first reach the location of the bomb before proceeding to detonate it. End-to-end training of such navigation agents with egocentric image inputs and without the use of an explicit geometric map for motion planning in a 3D environment [18, 19, 41, 32] has shown a lot of promise recently. The development of high-quality simulators with realistic rendering abilities has further helped research in this direction by trying to narrow the Sim2Real gap [18, 34, 8, 46].

Although photorealistic RGB and/or depth images can often provide rich navigation cues, they are highly local in nature, and can also be unreliable in low-light and high-occlusion situations. Besides, humans often use multi-modal sensory data (vision, sound, or smell) to compensate for the low or insufficient fidelity of one or more of the sensory streams. Further, human navigation also doesn't depend on an external explicit directional signal like GPS which is very widely used currently for robot navigation but can be very weak or even absent in real-world indoor environments. To that end, [9, 15] recently studied the problem of audio-visual navigation (AudioGoal) where an agent is tasked with finding an audio source by just relying on egocentric visual and binaural audio inputs. [10] improves upon [9, 15] by hierarchically predicting intermediate navigation waypoints of auto-adaptive granularity in an end-to-end fashion, and using a geometric-mapper and an analytical planner to reach the waypoints. Further, [10] shows the benefit of maintaining an explicit acoustic memory to improve navigation performance.

Despite being realistic in terms of audio and visual rendering, [15, 9, 10] essentially suffer from the over-simplistic assumption that there is a single audio source in the environment that's at worst corrupted by microphone noise. A more likely scenario in a live indoor setting is one where an autonomous agent often has to deal with a mixture of sounds from multiple sources and has to robustly navigate to a specific source while being able to ignore the distractors. This aspect

*Please view the slides by signing in using a Google account to play the videos.

of realism adds to the complexity of the vanilla AudioGoal task by requiring the navigation agent to implicitly or explicitly extract the target audio from the mixture on the basis of some given acoustic attribute(s) and use it as the cue for navigation.

In this report, we aim to study one such realistic setting, where the attribute of choice is the semantic class of the audio target, and introduce the task of *semantic audio-visual navigation (Semantic AudioGoal)*. In simpler terms, we consider the scenario where there are multiple audio sources, each of a different semantic class, playing in the environment and the agent is supposed to navigate to the source of a certain class where the class label could be an input from a human user (see Figure 1). A simple approach to this problem is to augment the input space of one of the single-source (in this context and all our discussions henceforth, *single-source* means that there is one dominant audio source in the environment and has nothing to do with the number of sources the agent is supposed to navigate to) navigation methods proposed in [15, 9, 10] with the class label and expect it to identify and extract the target audio on its own for navigation. However, our experiments show that such a minor modification results in low navigation performance, thus highlighting an increase in task complexity in comparison to the vanilla AudioGoal task.

We propose two multi-modal deep reinforcement learning (RL) approaches to this task. In our first approach, we design an end-to-end RL agent that takes egocentric visual and mixed audio inputs and learns a disentangled latent feature representation of them that is conditioned on the audio classes present in the mixture. Such latent disentanglement allows the agent to implicitly identify and extract only those features that correspond to the target audio class, and use them for navigation. Our second approach involves the use of a separate class-conditional audio extraction module that explicitly extracts the audio corresponding to the target class from the mixed audio before feeding it to the RL-based navigation policy. Such explicit extraction, if perfectly done, practically converts the mixed-audio navigation task into clean single-source navigation, thus freeing up the navigation policy network for learning only navigation-specific features and providing the flexibility of using any off-the-shelf single-source audio navigation policy.

For our experiments, we use certain commonly-occurring indoor audio and their class labels from the the Environmental Sound Classification (ESC-50) [37] dataset. We extend the SoundSpaces [9] audio simulation platform for scanned real-world 3D environments in the Replica [48] dataset with the ESC-50 audio data for rendering mixtures of audio from two or more of these classes. Our results show that both our implicit disentanglement-based and explicit target-audio extraction approaches outperform naive heuristical baselines that don't have access to the ground-truth target location but have the ability to stop perfectly, and a state-of-the-art end-to-end RL based AudioGoal navigation method that is modified to take the target audio class as input, on both heard and unheard sounds even when the number of total sound sources is increased from 2 to 3. The performance gains of our proposed methods are larger for heard sounds – 4 to 23 points on heard against 2 to 8 points on unheard. Furthermore, we qualitatively compare trajectories of our proposed navigation methods with those of the modified state-of-the-art AudioGoal navigation method, and also discuss some common failure cases for our methods. Finally, we visualize the learned latent features in our disentanglement-based approach that interestingly shows some correlation between the clusters and classes of audio present in the mixture.

II. Related Work

A. Learning to navigate in 3D environments

Most of the initial works on robot navigation have heavily relied on simultaneous localization and mapping (SLAM) approaches that involve the continuous building of a geometric map of the agent's 3D environment, the estimation of the agent's pose with respect to that map, and eventually planning the path to a certain goal location in the map [49, 14]. Recent advances in deep learning, however, have made it feasible to learn implicit state representations and end-to-end navigation policies directly from the egocentric RGB(D) images [18, 19, 32, 41]. However, a large number of these works tackle the PointGoal task where the main assumption is that the agent has access to a 2D displacement vector to the target location in the form of a GPS signal [18, 32, 8, 42, 43].

On the other hand, in ObjectGoal navigation the agent is tasked with navigating to a target of a certain semantic class instead of a specific location in its 3D environment. In particular, navigation is considered successful if the agent is able to reach the nearest instance of the given object class label [58, 3, 55, 33, 54, 7, 6].

The recently proposed AudioGoal task also attempts to loosen the assumption that the agent has continuous access to PointGoal inputs, and introduces the notion of using audio as the navigation cue instead. Specifically, the agent instead hears an audio in addition to input from its other image sensors, and has to navigate to the location of the audio source [9, 15, 10].

Our proposed Semantic AudioGoal task extends the vanilla AudioGoal navigation task by adding audio semantics to it. The agent is now tasked with navigating to a specific audio class while hearing a mixture of audio of different classes placed around the environment. Besides, in contrast to ObjectGoal navigation, in our setup there is at most one source from every audio class in the environment and the agent's navigation is considered successful if and only if it's able to reach the only source for the target class that's present in the environment.

B. Audio source localization

Sound source localization in robotics is mostly achieved using microphone arrays [35, 39], and active control is often used to improve the localization [36, 52]. Prior work has shown that audio signals can be used to infer partial geometric and spatial information about the environment [13, 16]. Audio-visual signals have also been used previously for several tasks, such as surveillance [53, 38], speech recognition [56], and most recently, robot navigation [9]. Moreover, sound sources in the presence of distractor sounds have been accurately localized

in 2D video frames by leveraging audio-visual association cues [20, 2]. In our proposed task of Semantic AudioGoal Navigation, the agents have no explicit supervision to localize the target sound source. Instead, they have to build an implicit understanding of the location of the target audio type from a mixture of different types for successful navigation.

C. Audio-only source separation

Audio source separation is an extensively studied problem in classical signal processing. One approach to solve it is to use multiple microphones to capture directional cues of different sources that are important for separation. Another way to tackle it is to do "blind" separation of monaural audio [23, 50, 25, 51], most recently with deep learning [23, 21, 47]. Mix-and-separate style training [57, 24, 21] is also commonly used nowadays to augment training data for improving separation performance. Instead of doing separation, one of our proposed approaches tries to do class-conditional extraction of the target audio. The alternate disentanglement-based approach separates mixed audio in the sense that it disentangles audio features but it does so inside the latent embeddings in an implicit fashion.

D. Disentangled representation learning

Disentangled feature representations in deep learning models are often known to be more interpretable or semantically meaningful [12, 28] and more generalizable [45]. Most of these approaches are unsupervised in nature and based on variational auto-encoders (VAEs) [27] or generative adversarial networks (GANs) [17].

VAEs tend to dominate the disentanglement landscape because of better training stability. While some of the VAE-based approaches tackle the problem from the persepctive of limiting the bottleneck capacity [22, 5], others penalize the total correlation [26, 11] or match factorized priors [28]. However, all of these works attribute disentanglement to factorizing the distribution of representations [26, 11, 30, 29].

Among GANs, InfoGAN [12] penalizes the mutual information of representations, and qualitatively shows that different factors in representations correspond to different visual concepts. The authors in [4] propose to penalize the Jensen-Shannon divergence between the distribution of representations and its factorized distribution with a discriminator, based on Independent Component Analysis.

Unlike these unsupervised disentanglement methods, our disentanglement-based implicit audio extraction approach is fully supervised in nature. It draws motivation from bottleneck-constriction in VAEs and conditions its factorization of latent features on the audio classes present in the mixed audio input to the agent by using a class-conditional supervised regularization loss during training.

III. Audio-visual Simulation

We build on top of the publicly available AI-Habitat [31] based SoundSpaces audio-visual simulation platform for our Semantic AudioGoal task. Specifically, we use SoundSpaces for the Replica dataset of real-world 3D scans. Replica consists of 18 environment meshes constructed from scans of apartments, hotels, rooms and offices. The simulation generates realistic and real-time audio-visual observations as a navigation agent traverses these 3D environments. While the realism and high fidelity of the egocentric visual images arise from the Replica data being dense scans of real-world scenes, the state-of-the-art spatial audio renderings leverage room impulse responses (RIR) that capture how sound propagates and interacts with the surrounding geometry and surface materials, modeling all of the major acoustic phenomena: direct sound, early specular/diffuse reflections, reverberation, binaural spatialization, and frequency dependent effects from materials and air absorption (see [9] for more details).

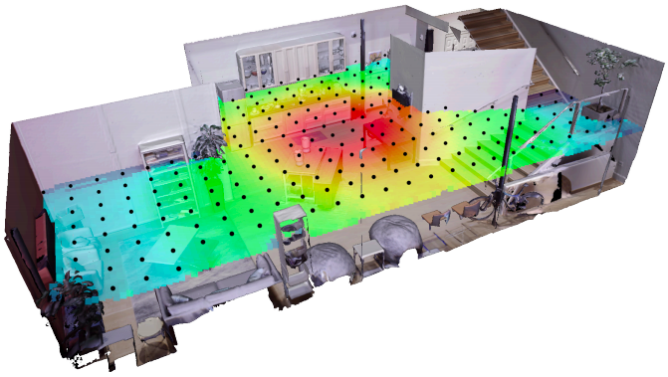


Fig. 2: **Audio simulation in SoundSpaces [9]**. This shows the audio pressure fields, shown using a heatmap with the pressure being higher at redder locations than at bluer locations, at a densely sampled grid inside 'FRL apartment 0' in Replica when the sound source is at the center of the room. At each grid point, the agent hears a binaural audio that captures the source's local intensity, direction of arrival and frequency texture.

For adding audio semantics to this simulation setup, we take monaural audio clips from the public ESC-50 dataset. In particular, we look at 10 indoor audio classes: 1) *vacuum cleaner*, 2) *door wood knock*, 3) *can opening*, 4) *mouse click*, 5) *clock alarm*, 6) *keyboard typing*, 7) *glass breaking*, 8) *washing machine*, 9) *clock tick* and 10) *door wood creaks*. The original ESC-50 dataset has 40 audio clips per class and each clip is 5 seconds long. We sample a 1 second chunk from each 5 second clip by using a sliding-window approach that ensures that the energy of the chosen window is not below a certain threshold. The energy threshold is enforced so that it has enough discernible features for a human listener to listen to it and agree on the class label that's been originally assigned to it. The chosen threshold for every clip is the average energy of the whole clip. However, we discard a few 1 second audio chunks after manually listening to them because of lack of class-specific information in them. This step reduces the total number of audio samples in our dataset from 400 (40 clips per class and 10 classes chosen in total) to 387. Finally, we resample all monoaural chunks, which are originally at 44.1

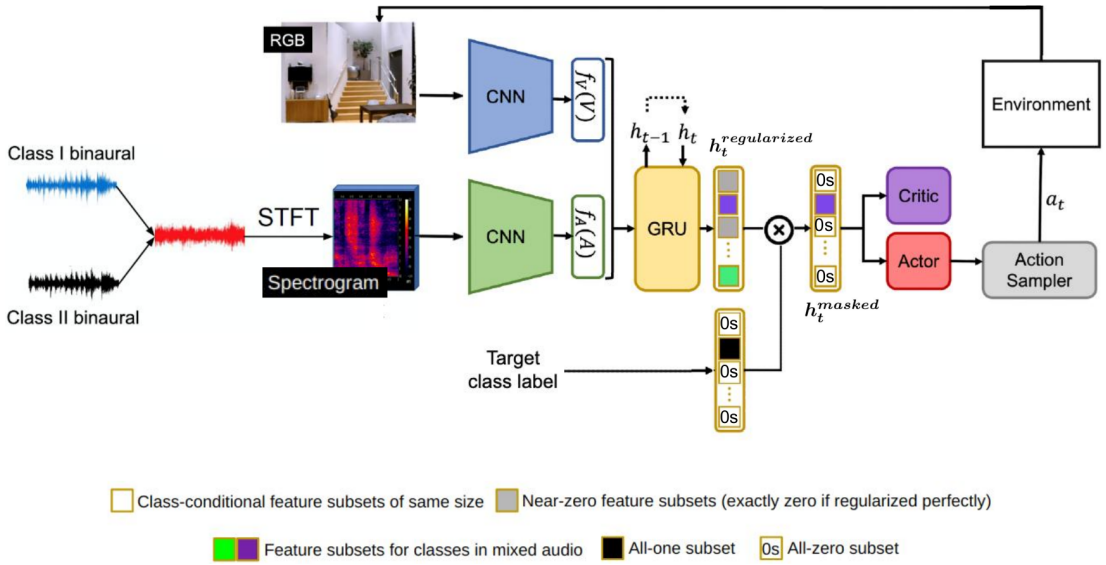


Fig. 3: **Class-conditional latent disentanglement:** our implicit audio extraction based navigation approach that disentangles the latent features in the policy into subsets where each subset corresponds to a different audio class and only the subsets corresponding the classes present in the audio mixture are active during navigation. Such disentanglement allows for using only the features for the target class to to navigate accurately and efficiently.

kHz, at 16 kHz and normalize them so that they have the same energy value. The chosen value is the average energy of the resampled audio chunks before normalization.

Next, we take this preprocessed monaural audio and generate navigation episodes for training and evaluation. We consider two scenarios: one where all the sounds in evaluation have been heard during training, henceforth referred to as *heard sounds*, and the other where there is no overlap of sounds among the training, validation and testing episodes but all the audio classes are shared, henceforth referred to as *unheard sounds*. For unheard sounds, we split the audio data in each category in the training/validation/testing ratio of 8:1:1. We modify the navigation episode definition in [9] by replacing the single audio source in every episode with two or more sources where each source is from a different audio class. We also enforce a constraint on the spatial positioning of the audio sources: the Euclidean distance for every source-source pair and agent-source pair at the start of navigation is at least 1.5 meters. This is to make sure that the agent doesn't fail to hear the target sound during its initial steps or when it gets very close to the target source. Furthermore, the modified episode also contains the class label of the target audio source. For every episode, the simulation times out after 500 navigation steps, which is a configurable parameter, and the agent is reset to start a new episode.

For rendering mixed audio at every navigation step, we generate binaural waveforms by convolving the monaural waveform of each source with the RIR for that source at the agent's current pose, take a mean over the individual binaural waveforms and convert the mean to a spectrogram using the short time Fourier transform (STFT). For more details on spectrogram computation, refer to section I in Supp.

The simulator maintains a navigability graph of the environment (unknown to the agent). The agent can only move from one node to another if there is an edge connecting them and the agent is facing that direction. The action space $\mathcal{A}$ has four actions: $MoveForward$, $TurnLeft$, $TurnRight$ and $Stop$, where a successful $MoveForward$ takes the agent forward by 0.5 meter. The step size is the result of the spatially-discrete audio simulation in SoundSpaces (see Figure 2 for an example).

IV. Approach

We propose two unique deep RL based approaches specific to the Semantic AudioGoal task. Both our methods are motivated by the need to extract the target audio from the mixture either implicitly or explicitly for the agent to navigate accurately and efficiently. The first approach involves implicit extraction of target audio in the form of class-specific disentangled latent feature representation. The other approach does more explicit extraction of it by using an audio extraction component that tries to predict the audio spectrogram for the target class while taking the mixed audio as input.

Despite these differences, the navigation policies used by the approaches are inspired by the actor-critic architecture proposed in [9] and have many elements in common. For visual and acoustic perception, they take the agent's first-person RGB view and a binaural audio spectrogram as inputs and encode them using separate CNNs into two feature vectors, $f_V(V)$ and $f_A(A)$ respectively. $f_V(V)$ and $f_A(A)$ are then concatenated and fed into a GRU that acts as a memory for navigation and aggregates the audio-visual features over time. Depending on

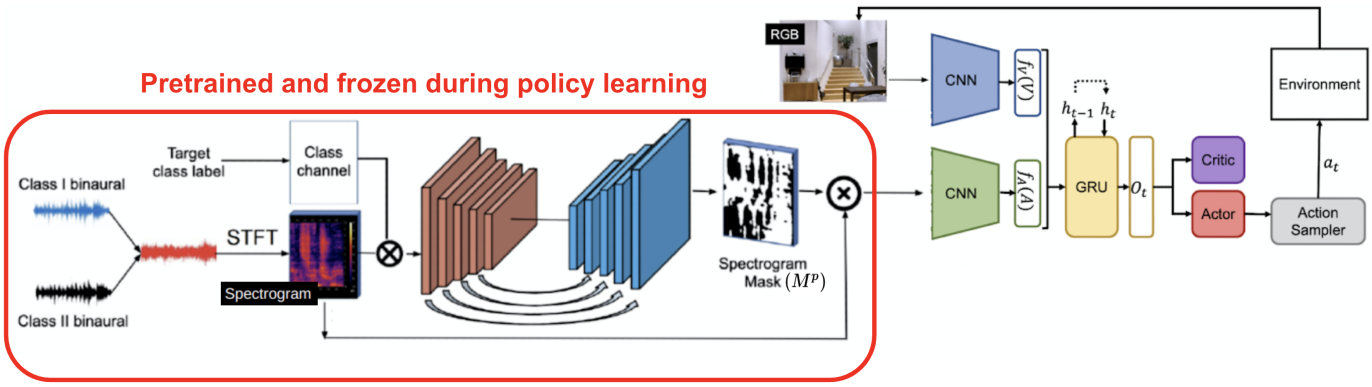


Fig. 4: **Class-conditional target audio extraction:** our other proposed approach that explicitly extracts the target audio from the mixture by using an encoder-decoder architecture. The extraction module is pre-trained and frozen during navigation policy learning.

the approach, the output of the GRU (h_t) is either directly or indirectly fed into the actor and the critic networks that respectively predict the next action and the value of the current agent state. See section IIA in Supp. for architectural details of our navigation policy.

Following typical navigation reward defintions [42, 9, 10], we reward the agents with +10 if it succeeds in reaching the goal and executing the Stop action there, plus an additional reward of +0.25 for reducing the geodesic distance to the goal and an equivalent penalty for increasing it. Finally, we issue a time penalty of 0.01 per executed action to encourage efficiency.

While the policy in the explicit extraction based approach is trained using the proximal policy optimization (PPO) [44] algorithm, the training loss for the other approach includes a loss for enforcing latent disentanglement in addition to the action, value and entropy losses in the vanilla PPO algorithm. The PPO updates are done after every 150 navigation steps for both the policies. See section IIB in Supp. for more details on policy training.

Next, we discuss in detail the main differences between the two approaches and their training.

A. *Class-conditional latent disentanglement*

The core idea behind this method is to leverage class-conditional sparsity in learned latent features for enforcing their disentanglement into class-specific subsets so that only the features corresponding to the target class can be used for accurate navigation. In other words, we modify the end-to-end deep RL based navigation policy, discussed above, so that the temporally aggregated audio-visual feature output from the GRU of the policy can be split into subsets where each subset is of the same size and corresponds to one audio class in the dataset, and the subsets corresponding to all classes absent in the current audio mixture take all-zero values when trained perfectly (see Figure 3). To achieve this sparsity, we regularize the GRU output (h_t) by masking it with a binary vector b in which all the indices corresponding to the classes present in the mixture are 0 and the rest are 1, and adding the weighted l_2-norm of the masked vector to the PPO training loss. The modified training loss can be written as follows:

$$\mathcal{L}_{training} = \mathcal{L}_{PPO} + \alpha * ||h_t \odot b||_2^2 \quad (1)$$

where α is the weight for the regularization loss. Next, these sparse features are masked using the class label of the target audio so that only the feature subset corresponding to the target class can be implicitly extracted and fed into the actor and the critic of the policy.

B. *Class-conditional target audio extraction*

In this method, we explicitly extract the audio corresponding to the target class from the audio mixture and feed it to the audio encoder of our navigation policy. The extractor module uses a U-net [40] backbone that takes the spectrogram of the mixed binaural audio concatenated with a single channel of the target audio's integer class-label as input and predicts a ratio mask of the spectrogram of the target binaural (see Figure 4). The final output of the extractor is an element-wise product of the predicted ratio mask and the mixed audio spectrogram, which represents the module's estimate of the target audio spectrogram. The extractor is pre-trained with the mean-squared error (MSE) loss between the target audio spectrogram ground-truth and the estimate, and kept frozen during policy training. For this pre-training, we build a dataset of audio source and receiver locations that are randomly sampled from Replica training environments while obeying the inter-source minimum Eucliean distance criterion of 1.5 meters. For more details on the extractor architecture and the pre-training dataset, refer to section III in Supp.

V. Experiments

A. *Environment*

We train and evaluate our proposed approaches in the 3D environments of Replica using our extension of the SoundSpaces

Model	2 sources						3 sources					
	Heard			*Unheard*			*Heard*			*Unheard*		
	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑
Random w/ perfect stopping	5.3	20.8	1.8	5.3	20.8	1.8	5.3	20.8	1.8	5.3	**20.8**	1.8
Move forward w/ perfect stopping	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Modified Chen and Jain et al.	7.8	13.0	3.1	4.3	7.2	1.7	2.2	4.6	0.8	2.2	4.0	0.9
Class-conditional latent disentangler	12.8	22.5	4.6	8.8	15.6	3.3	6.1	10.6	2.5	4.3	8.4	1.7
Class-conditional audio extractor	**30.5**	**53.5**	**12.4**	**12.7**	**23.1**	**4.9**	**26.4**	**44.8**	**10.3**	**6.5**	12.8	**2.2**

TABLE I: **Semantic AudioGoal navigation results.** Our proposed navigation approaches (bottom 2 rows) outperform naive heuristical baselines that can stop perfectly, and a state-of-the-art end-to-end RL based AudioGoal navigation agent that's been customized for the Semantic AudioGoal task, on important navigation metrics like SPL and SNA in multiple challenging evaluation settings. All metrics are reported as percentages and for all of them, higher is better. All values are statistically significant with the maximum standard deviation being less than 0.4% across 3 different random seeds.

audio rendering platform and AI-Habitat simulator. We follow the SoundSpaces AudioGoal protocol with train/val/test splits of 9/4/5 scenes on Replica. Note that these splits are completely disjoint in nature. We do two sets of experiments: one where every navigation episode has 2 sound sources and the other where every episode has 3 sound sources placed in the environment. For both sets, we evaluate with both heard and unheard sounds. Finally, the number of episodes in training, validation and testing are 109676, 500 and 1000 respectively in all our experiments.

B. Evaluation metrics

In all our experiments, we use the following metrics for evaluating navigation performance: 1) success rate (SR) – the fraction of episodes in which the agent has stopped exactly at the target source, 2) success weighted by path length (SPL) – the standard metric [1] that weights successes in individual episodes by the ratio of the length of the shortest geodesic path from the agent's starting location to the target and that of the path actually traversed by the agent, and 3) success weighted by number of actions (SNA) – a recently introduced navigation metric [10] that weights successes with the ratio of the optimal number of actions and the actual number of actions taken by the agent in its trajectory, thus also penalizing in-place-rotations in addition to redundant $MoveForward$ actions.

C. Existing methods and baselines

We compare our proposed approaches with the following methods:

– Random w/ perfect stopping: this is a naive heuristical baseline that executes a random action from the action-space $\mathcal{A}' : \{MoveForward, TurnLeft, TurnRight\}$ until it reaches the target location in which case it always takes the $Stop$ action perfectly.

– Move forward w/ perfect stopping: this is another heuristical baseline that always takes the $MoveForward$ action until it reaches the target location in which case it always takes the $Stop$ action perfectly.

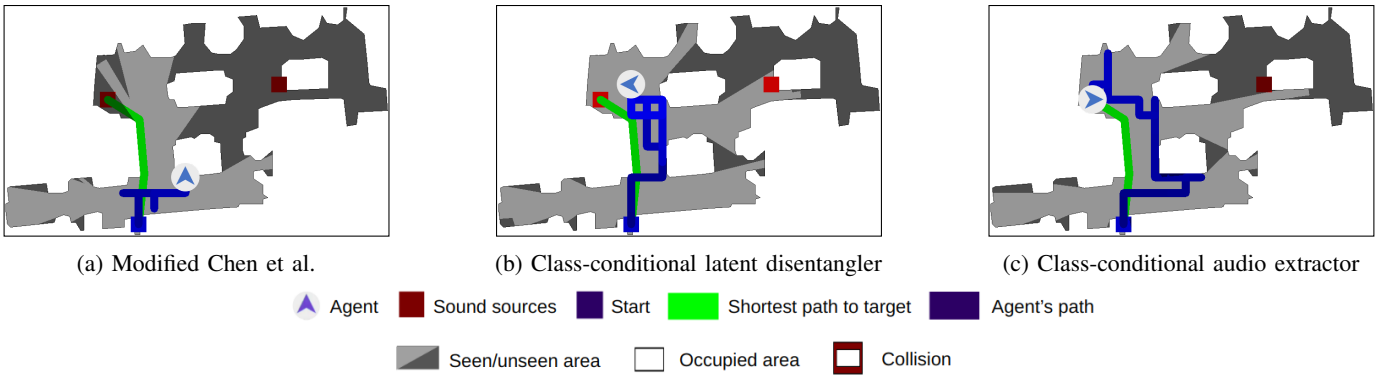
– Modified Chen and Jain et al. (2020): this is a state-of-the-art end-to-end RL based audio-visual navigation agent introduced in [10] that we modify by augmenting the spectrogram input to the model's audio encoder by adding an extra channel that stores the integer class label of the target audio source. This method is also trained with PPO like our proposed methods. For architectural and training details of this approach, we point the reader to section II in Supp.

For a fair comparison, we make certain architectural choices so that both our approaches and the modified Chen and Jain et al. model have approximately similar representation capacity in terms of the number of learnable features (see Supp. section II). Also note that both our naive heuristical baselines have access to the privileged information of ground-truth goal location, which neither modified Chen and Jain et al. nor our approaches have. Perfect stopping is needed for the naive baselines or else they always give 0 performance consistently across all evaluation metrics for multiple random seeds.

D. Navigation results

Table I lists our results with 2 and 3 sound sources with both heard and unheard sounds, where the bottom two rows show the results for our proposed approaches. We reiterate that for unheard sounds, there is no overlap among the sound samples in training, validation and testing splits while all the sound categories are shared. Further, for all our experiments, there is no overlap of Replica environments among the training and evaluation splits.

Among the heuristical baselines, random with perfect stopping does decently well mostly because of its knowledge of the exact target location and partially due to the small size of Replica scenes, especially the rooms and hotels. The other privileged MoveForward baseline gives 0 scores on all metrics even though it can also stop perfectly. That's because of the highly non-linear and complex nature of the ground-truth trajectories that either make the MoveForward agent collide very frequently with surrounding obstacles, go completely in the wrong direction or both. The modified Chen and Jain et al. agent suffers a very large drop in performance from the vanilla AudioGoal setting (refer to [10] for the exact numbers) even though it has access to the target audio class label and in principle, it should be able to learn an association between the given class label and the correct set of actions through



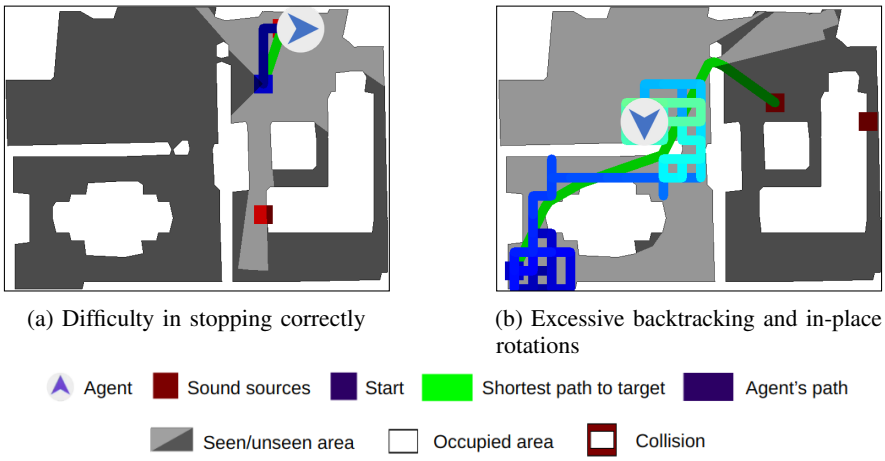
(a) Modified Chen et al. (b) Class-conditional latent disentangler (c) Class-conditional audio extractor

Fig. 5: **Navigation behaviors of the modified Chen et al. model and our approaches.** While the modified Chen et al. agent is unable to follow the target audio because of its inability to learn the target-class to source association and ends up moving towards a wrong source. Our latent disentanglement based agent initially moves in the right direction but suffers from bad implicit extraction of the target audio when the distractor sound gets louder. On the other hand, explicit target audio extraction helps our other agent to successfully reach the target.

RL training. Figure 5a shows an example 2-source scenario in which this agent is unable to use the target audio as the navigation cue and ends up moving towards the source of the other class. These point to the very high complexity of the Semantical AudioGoal task and justifies the need to design task-specific agents for solving it.

Our class-conditional latent disentanglement based approach outperforms the modified Chen and Jain et al. model on all metrics in all 4 evaluation settings. The performance gains are larger for 2 sources and for heard sounds – for SPL, there is a 5 point gain on heard versus a 4.5 point gain on unheard with 2 sources, and a 3.9 point gain on heard versus a 2.1 point gain on unheard with 3 sources. Besides, the disentangler gets higher SPL scores than the privileged random on all settings except the one with 3 sources and unheard sounds. This implies that although the random baseline has a very high success rate due to its ability to perfectly stop, especially in smaller scenes, the disentangler is more efficient in terms of navigation. In the same example episode as the one shown in 5a, the disentangler is able to initially navigate very close to the target but fails to stop at the right location, possibly due to its inability to implicitly extract the target audio when the distractor audio becomes stronger (see Figure 5b).

Our other proposed approach that does class-conditional target audio extraction beats all other models on the SPL and SNA metrics in all 4 evaluation settings. While the smallest SPL gain over the next best model is 1 point (unheard with 3 sources where the next best is the random agent), the highest gain is as large as 20.3 SPL points (for heard with 3 sources where the disentangler is the next best performer). On the SR metric, the extractor is not able to beat the random agent on the only 1 setting – unheard sounds with 3 sources. Although this could be attributed to perfect stopping in the random agent, the pre-trained extractor module doesn't generalize very well to unheard sounds and might be partially responsible for the worse navigation performance of this approach on unheard sounds in comparison to heard sounds. However, the overall highly promising performance of our extraction-based



(a) Difficulty in stopping correctly (b) Excessive backtracking and in-place rotations

Fig. 6: **Common failure cases in our proposed methods.** The two most common failure cases in our proposed methods arise either from the agent's inability to stop correctly owing to its over-reliance on misleading audio cues near the target for stopping, or excessively reactive action-taking i.e. very frequent backtracking and in-place rotations.

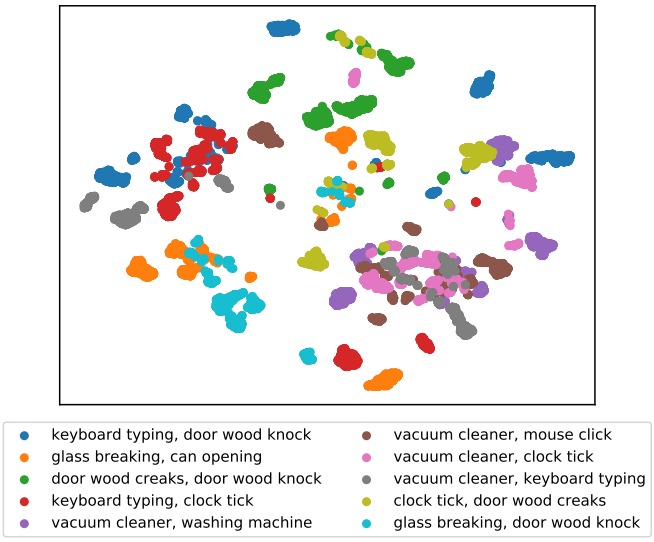


Fig. 7: **Visualization of disentangled features**. Latent features of our disentanglement based approach when visualized using 2D tSNE projections tend to show some clustering on the basis of the audio classes present in the audio mixture, thus hinting at partially successful latent disentanglement.

approach in comparison to the other methods demonstrates how explicit extraction of the target audio can make navigation more accurate and efficient by practically converting mixed-audio navigation into noisy single-source navigation at worst. One such example is shown in Figure 5c where the extraction-based agent successfully reaches the target while silencing out the distractor sound even though the distractor gets louder as the agent nears the target.

E. Cases of unsuccessful navigation

We manually look at navigation trajectories of our proposed agents and identify two very common fault types. In one, the agent is able to navigate very close (within one grid step) to the target but is not able to stop at the right location. This happens due to the near-uniform audio field at locations very close to the target and the lack of embodiment of the target in the simulation setup, which denies the agent the ability to use a visual cue for stopping when the audio cue becomes misleading. We show one such example episode in Figure 6a.

In the other failure type, the agent goes in wrong directions, backtracks or does in-place rotations a lot until the episode times out. Although, in a few such cases, the agent's general motion direction is towards the target, the highly reactive decision making prevents it from successful navigation. Such a behavior arises mostly from the agent's inability to plan long-term due to the complex nature of the task and the navigation environments. See Figure 6b for an example.

F. Class-conditional disentangled embeddings

To understand the extent of disentanglement in the learned latent features of our disentanglement-based approach, we evaluate the model on our test split with 2 sound sources and both heard and unheard sounds, and collect the GRU feature outputs of the model (h_t in Figure 3) at every navigation step. These features are originally 640-dimensional and for effective visualization, we embed them in 2 dimensions using t-SNE . While a maximum of $\binom{10}{2} = 90$ pairs are possible with 10 audio classes in the dataset in total, 2 sources in each episode and a different class at each source, this particular evaluation setup has 88 audio-class pairs in total. For the sake of clarity, we randomly sample 10 class-pairs from the total of 88 and show their 2D plot in Figure 7. Although the embedding doesn't show perfect disentanglement of the features, it reflects some correlation between the spatial location of the features and the corresponding class-pair in the mixed audio. Particularly, the class-pairs – {(*door wood creaks*, *door wood knocks*); (*keyboard typing*, *clock tick*); (*vacuum cleaner*, *vacuum machine*); (*vacuum cleaner*, *clock tick*); (*glass breaking*, *door wood knock*)} – tend to show spatially-localized clustering to a certain extent.

The imperfect nature of clustering helps us identify one flaw in our model design for the disentanglement-based approach. Our current model tries to disentangle the GRU output that, in addition to acoustic information, has visual and temporal information, which are not necessarily from distributions that can be factorized on the basis of audio class. A better alternative could be to disentangle the feature output of the audio encoder ($f_A(A)$ in Figure 3) that only has features of the mixed audio. This change might also bring about an improvement in the navigation performance of the model.

VI. Conclusion

We introduced a new audio-visual navigation task where an embodied agent is tasked with navigating to an audio source of a given semantic class when multiple audio sources, each of a different class other than the target class, are playing around the environment. To solve this task, we also propose two novel techniques that try to either explicitly or implicitly extract the target audio from the mixed audio input for navigating accurately and efficiently. Our methods improve over a current state-of-the-art end-to-end reinforcement learning based single-source audio-visual navigation agent that has been customized for our proposed task, and our analysis shows the direct impact of the new technical contributions. In future work, we plan to address the existing issues with our proposed approaches, namely overfitting to heard sounds during pre-training of the extractor module of our explicit extraction based approach and trying to do class-conditional disentanglement of temporally-aggregated audio-visual features instead of pure audio features. We also hope to extend our simulation setup and proposed methods to settings in which the audio sources are placed at semantically meaningful locations in the environment instead of random locations, are more realistic in that they are time-varying in nature, and play only sporadically and not for the whole duration of the agent's navigation.

REFERENCES

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

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[3] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020.

[4] Philemon Brakel and Yoshua Bengio. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*, 2017.

[5] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. arxiv 2018. *arXiv preprint arXiv:1804.03599*, 1804.

[6] Matthew Chang, Arjun Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos. In *arXiv*, 2020.

[7] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *arXiv preprint arXiv:2007.00643*.

[8] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *ICLR*, 2020.

[9] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigaton in 3d environments. In *ECCV*, 2020.

[10] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Audio-visual waypoints for navigation. *arXiv preprint arXiv:2008.09622*, 2020.

[11] Ricky T. Q. Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 2610–2620. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf.

[12] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29:2172–2180, 2016.

[13] Ivan Dokmanić, Reza Parhizkar, Andreas Walther, Yue M Lu, and Martin Vetterli. Acoustic echoes reveal room shape. *Proceedings of the National Academy of Sciences*, 110(30):12186–12191, 2013.

[14] Jorge Fuentes-Pacheco, José Ruiz-Ascencio, and Juan Manuel Rendón-Mancha. Visual simultaneous localization and mapping: a survey. *Artificial intelligence review*, 43(1):55–81, 2015.

[15] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. *arXiv preprint arXiv:1912.11684*, 2019.

[16] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. *arXiv preprint arXiv:2005.01616*, 2020.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

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

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

[20] John R Hershey and Javier R Movellan. Audio vision: Using audio-visual synchrony to locate sounds. In *Advances in neural information processing systems*, pages 813–819, 2000.

[21] John R Hershey, Zhuo Chen, Jonathan Le Roux, and Shinji Watanabe. Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 31–35. IEEE, 2016.

[22] I. Higgins, Loïc Matthey, A. Pal, C. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[23] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis. Deep learning for monaural speech separation. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1562–1566, 2014.

[24] Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, and Paris Smaragdis. Joint optimization of masks and deep recurrent neural networks for monaural source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2136–2147, 2015.

[25] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*,

13:411–430, 2000.

[26] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/kim18b.html.

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[28] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. VARIATIONAL INFERENCE OF DISENTANGLED LATENT CONCEPTS FROM UNLABELED OBSERVATIONS. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1kG7GZAW.

[29] Francesco Locatello, Damien Vincent, Ilya Tolstikhin, Gunnar Ratsch, Sylvain Gelly, and Bernhard Scholkopf. Clustering meets implicit generative models, 2018. URL https://openreview.net/forum?id=rk4QYDkwz.

[30] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/locatello19a.html.

[31] Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019.

[32] Dmytro Mishkin, Alexey Dosovitskiy, and Vladlen Koltun. Benchmarking classic and learned navigation in complex 3d environments. *arXiv preprint arXiv:1901.10915*, 2019.

[33] A. Mousavian, A. Toshev, M. Fišer, J. Košecká, A. Wahid, and J. Davidson. Visual representations for semantic target driven navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8846–8852, 2019.

[34] Matthias Müller, Alexey Dosovitskiy, Bernard Ghanem, and Vladlen Koltun. Driving policy transfer via modularity and abstraction. *CoRR*, abs/1804.09364, 2018. URL http://arxiv.org/abs/1804.09364.

[35] Kazuhiro Nakadai and Keisuke Nakamura. Sound source localization and separation. *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–18, 1999.

[36] Kazuhiro Nakadai, Hiroshi G Okuno, and Hiroaki Kitano. Epipolar geometry based sound localization and extraction for humanoid audition. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, volume 3, pages 1395–1401. IEEE, 2001.

[37] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press. ISBN 978-1-4503-3459-4. doi: 10.1145/2733373.2806390. URL http://dl.acm.org/citation.cfm?doid=2733373.2806390.

[38] Jianzhao Qin, Jun Cheng, Xinyu Wu, and Yangsheng Xu. A learning based approach to audio surveillance in household environment. *International Journal of Information Acquisition*, 3(03):213–219, 2006.

[39] Caleb Rascon and Ivan Meza. Localization of sound sources in robotics: A review. *Robotics and Autonomous Systems*, 96:184–210, 2017.

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[41] Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation. *arXiv preprint arXiv:1803.00653*, 2018.

[42] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9339–9347, 2019.

[43] Alexander Sax, Bradley Emi, Amir R Zamir, Leonidas Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. *arXiv preprint arXiv:1812.11971*, 2018.

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

[45] Xander Steenbrugge, Sam Leroux, Tim Verbelen, and Bart Dhoedt. Improving generalization for abstract reasoning tasks using disentangled feature representations. *arXiv preprint arXiv:1811.04784*, 2018.

[46] Gregory J. Stein, Christopher Bradley, and Nicholas Roy. Learning over subgoals for efficient navigation of structured, unknown environments. In Aude Billard, Anca Dragan, Jan Peters, and Jun Morimoto, editors, *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 213–222. PMLR, 29–31 Oct 2018.

[47] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Adversarial semi-supervised audio source separation applied to singing voice extraction. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2391–2395. IEEE, 2018.

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

[49] Sebastian Thrun. Probabilistic robotics. *Communications of the ACM*, 45(3):52–57, 2002.

[50] T. Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(3):1066–1074, 2007.

[51] Beiming Wang and Mark D. Plumbley. Investigating single-channel audio source separation methods based on non-negative matrix factorization.

[52] Yu Wang, Mubbasir Kapadia, Pengfei Huang, Ladislav Kavan, and Norman I Badler. Sound localization and multi-modal steering for autonomous virtual agents. In *Proceedings of the 18th meeting of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 23–30, 2014.

[53] Xinyu Wu, Haitao Gong, Pei Chen, Zhi Zhong, and Yangsheng Xu. Surveillance robot utilizing video and audio information. *Journal of Intelligent and Robotic Systems*, 55(4-5):403–421, 2009.

[54] Yi Wu, Yuxin Wu, Aviv Tamar, Stuart Russell, Georgia Gkioxari, and Yuandong Tian. Bayesian relational memory for semantic visual navigation. In *Proceedings of the 2019 IEEE International Conference on Computer Vision (ICCV)*, 2019.

[55] Wei Yang, X. Wang, Ali Farhadi, A. Gupta, and R. Mottaghi. Visual semantic navigation using scene priors. *ArXiv*, abs/1810.06543, 2019.

[56] Takami Yoshida, Kazuhiro Nakadai, and Hiroshi G Okuno. Automatic speech recognition improved by two-layered audio-visual integration for robot audition. In *2009 9th IEEE-RAS International Conference on Humanoid Robots*, pages 604–609. IEEE, 2009.

[57] Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2017.

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