Pay Self-Attention to Audio-Visual Navigation

Yinfeng Yu[1, 3, *]
yyf17@mails.tsinghua.edu.cn

Lele Cao[1, 2, *]
lele.cao@eqtpartners.com

Fuchun Sun[1, ✉]
fcsun@mail.tsinghua.edu.cn

Xiaohong Liu[1]
liu-xh17@mails.tsinghua.edu.cn

Liejun Wang[3]
wljxju@xju.edu.cn

[1] Department of Computer Science and Technology, State Key Lab on Intelligent Technology and Systems, Tsinghua University, Beijing, China

[2] Motherbrain, EQT, Stockholm, Sweden

[3] College of Information Science and Engineering, Xinjiang University, Urumqi, China

Abstract

Audio-visual embodied navigation, as a hot research topic, aims training a robot to reach an audio target using egocentric visual (from the sensors mounted on the robot) and audio (emitted from the target) input. The audio-visual information fusion strategy is naturally important to the navigation performance, but the state-of-the-art methods still simply concatenate the visual and audio features, potentially ignoring the direct impact of context. Moreover, the existing approaches requires either phase-wise training or additional aid (e.g. topology graph and sound semantics). Up till this date, the work that deals with the more challenging setup with moving target(s) is still rare. As a result, we propose an end-to-end framework FSAAVN (feature self-attention audio-visual navigation) to learn chasing after a moving audio target using a context-aware audio-visual fusion strategy implemented as a self-attention module. Our thorough experiments validate the superior performance (both quantitatively and qualitatively) of FSAAVN in comparison with the state-of-the-arts, and also provide unique insights about the choice of visual modalities, visual/audio encoder backbones and fusion patterns.

Figure 1: Audio-visual embodied navigation with a moving sound source as the target: a blue robot chases a moving target (red) that is a low-speed robot emitting sound.

1 Introduction

Embodied navigation [13, 26, 34, 48] involves robotic agents (with egocentric observation) [6, 14, 27, 33, 38] exploring the unknown environment [25, 29, 35] to reach target (sometimes moving) locations. Embodied navigation has been a hotspot research topic in the broader domain of embodied intelligence [3]. Up till this date, most embodied navigation work relies on sensors such as vision and lidar [5, 28], ignoring other vital senses like hearing heavily utilized by some animals [15, 43, 45]. Hearing is a unique and important sense because it is known to be both temporal and spatial informative [24, 36, 46], enabling the visually impaired subjects to navigate properly [19]. Inspired by the simultaneous use of visual and hearing by animals and humans [18, 49], audio-visual assistance is believed to be beneficial to the efficiency and robustness [44, 52] of many different robotic tasks, such as audio-visual association [16], moving vehicle tracking [20], visual sound separation [21, 30], object detection [47], audio-visual dereverberation [10], audio-visual matching [11, 12, 31], audio-visual floor plan reconstruction [2, 37], and finally the main focus of this work: audio-visual embodied navigation [7, 8, 9, 22, 50, 52].

    Among the state-of-the-art audio-visual navigation researches, some [22] require multiple sequential steps to reach the goal; some [9, 22] need to build topology graph; and some [8, 23] rely on sound meta information. Unfortunately, all of these methods can only deal with single sound target that stays at the same position throughout the entire navigation task. Most recently, the authors of [52] and [50] experiment with the multiple sound sources and moving sound source, respectively. However, they simply apply concatenation to fuse the visual and audio information, as shown in Fig.2(a). Under this status-quo, we propose an end-to-end framework – Feature Self-Attention Audio-Visual Navigation (FSAAVN), which supports chasing a moving sound target without the need of topology graph or sound meta information. More importantly, as shown in Fig.2(d), we propose a novel audio-visual fusion module FSA (feature self-attention) to learn a context-aware strategy to determine the relative contribution of each modal in real-time. The main contributions of this work are:

- we propose a end-to-end framework (FSAAVN) to address a currently under-researched problem: audio-visual navigation to chase a moving sound target;

- we design a novel audio-visual fusion module (FSA) to learn a context-aware strategy to determine the relative contribution of each modal in real-time;

- we experimentally benchmark our approach towards the state-of-the-arts in 3D environments, showing the superior performance of FSAAVN;

- the thorough comparison of different variants of the fusion module (Fig.2) and visual/audio encoder[1] provides useful insights for future practitioners in this field.
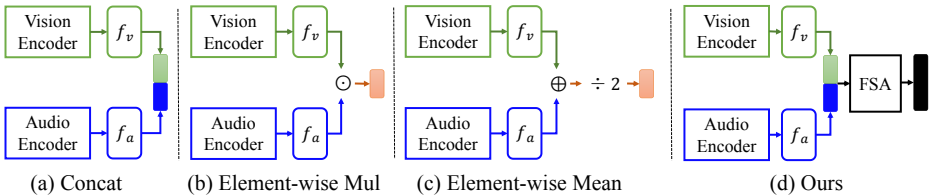

(a) Concat   (b) Element-wise Mul   (c) Element-wise Mean   (d) Ours

Figure 2: High-level illustration of different feature fusion methods in audio-visual navigation.

---

[1]The considered encoders: CNN (convolution neural network), ViT (vision transformer) [1], Capsule [69].

2   Related Work

Audio-visual embodied navigation is largely grouped into two categories in accordance with the behavior of sound source(s): static-sound and moving-sound sources. We visually illustrate the landscape of our literature survey in Fig.3.
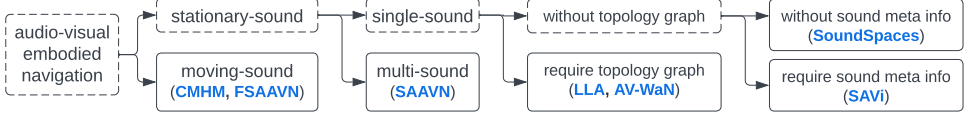
Figure 3: The landscape of the related work addressing audio-visual embodied navigation.

Stationary-sound refers to an environment where the location of the sound source is preset and remains the same. The number of sound sources could be one (*single-sound*) [1, 8, 9, 27] or many (*multi-sound*) [52]. To this date, the majority researches of audio-visual navigation have been carried out in *single-sound* environment, such as LLA (look, listen and act) [27], SoundSpaces [1], AV-WaN (audio-visual waypoint navigation) [9], SAVi (semantic audio-visual navigation) [8]. LLA [27] is a classical phase-wise navigation solution that needs to build a topological graph to aid shortest path planning. SoundSpaces [1] is the first end-to-end approach that does not rely on any topology graph or sound meta information (e.g. the category of the sound source: telephone, doorbell, alarm, etc). AV-WaN [9] predicts waypoints (represented as a topological graph) to improve long-distance navigation performances. The authors of SAVi [8] address the situation where the sound is not emitted continuously by incorporating the sound meta information. In the *multi-sound* case, SAAVN (sound adversarial audio-visual navigation) [52] propose and end-to-end framework to deal with acoustically complex environments where the target sound (usually only one source) is mixed up with other noisy sounds (usually more than one source).

Moving-sound, as the name implies, refers to a sound source that keeps changing its position, and the navigation target is that moving sound source. To the best of our knowledge, CMHM (catch me if you hear me) [50] is the first and only published work to tackle the moving-sound problem within unexplored environments. This work (FSAAVN) is closely related to CMHM with significant advancements mainly in audio-visual fusion strategy.

Audio-visual fusion: normally, the audio and visual inputs are encoded (using different encoders) into audio and visual feature vectors, respectively. They need to be fused before fed to the downstream neural networks. Unfortunately, all of the aforementioned approaches simply concatenate them to form a fused feature vector, as shown in Fig.2(a). In this work, we attempt to solve the moving-sound navigation problem using a more advanced audio-visual fusion mechanism: Feature Self-Attention (FSA) as illustrated in Fig.2(d).

3   Setting the Stage and Goal

In 3D environments demonstrated in Fig.1, a robot learns to chase after and catch up with a moving sound target. To concretize the research, we adopt the commonly used 3D environments collected using the SoundSpaces platform [1] and Habitat simulator [40]. They are publicly available as several datasets: Replica [42], Matterport3D [4] and SoundSpaces (audio) [1]. Replica contains 18 environments in the form of grids (with a resolution of 0.5 meter) constructed from accurate scans of apartments, offices and hotels. Matterport3D has 85 scanned grids (1 meter resolution) of indoor environments like personal homes. In
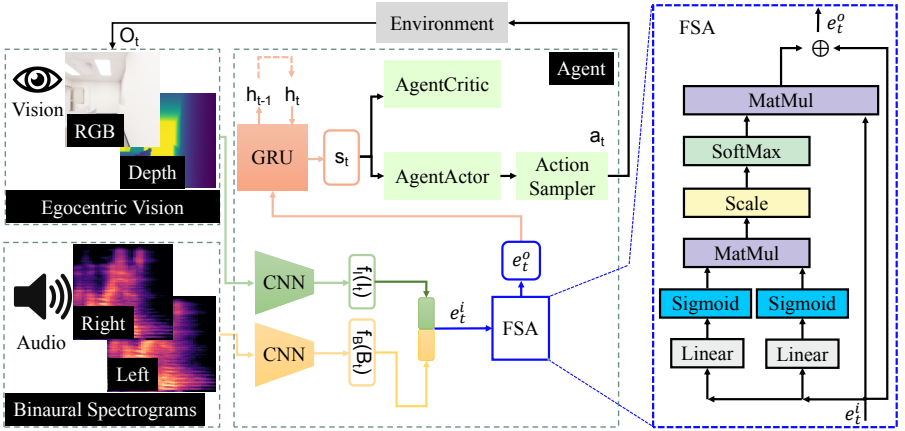
Figure 4: The overview of FSAAVN: Feature Self-Attention Audio-Visual embodied Navigation.

SoundSpaces, a sound source emits omnidirectional sound that is convolved with the corresponding binaural RIR (room impulse response); the convolved result is a binaural environmental response that is received by the navigating robot from its facing direction.

At each step t (cf. Fig.4), the robot receives the current observation $O_t = (I_t, B_t)$, where I_t denotes the current visual input that can be RGB ($128 \times 128 \times 3$ pixels) and/or depth (with a dimension of $128 \times 128 \times 1$) image[2], B_t represents the received binaural target sound spectrum[3] consisting of the audio signal from the left and right "ears". Although there exists a *navigability graph* (with nodes and edges) of the environment, this graph is hidden from the robot, hence it must learn from the accumulate observations O_t to understand the geometry of the scene. At each step, the agent at a certain node A can only move to another node B in the navigability graph if 1) an edge is connecting both nodes, and 2) the robot is facing node B. The viable robotic action space is defined as $\mathcal{A}$ ={MoveForward, TurnLeft, TurnRight, Stop}, where the Stop action should be executed when the robot and the moving audio target are on the same node in the navigability graph. The overall goal of the navigating robot is to **learn a policy to catch up with the moving audio target as fast as possible using both the visual and audio input**.

One might ask about the behavior of the moving audio target, we adopt the same approach used in CMHM [50]. In the beginning, the sound source randomly selects a destination position to move to; the selected destination position must have a traversable path to the current position of the sound source. Then, the sound source starts to follow the shortest path to reach (step by step) the designation position. However, at each step, the robot may only turn left/right while the sound source always match to the next node in the navigability graph. As a result, each move step of the sound source occurs with only 30% probability, so that the robot has a possibility to eventually catch up with the moving sound source. Once the sound source reaches the current destination position, a new destination position is randomly selected again so that the sound source will continue moving (in exactly the same manner described previously) to the new destination until the robot catches it.

---

[2]Both RGB and depth images capture the 90-degree field of view in front of the navigating robot.

[3]The audio spectrum is prepared in the same way as [□, □]: we compute the short-time Fourier transform with a window length of 512 samples and a hop length of 160 samples, corresponding to a duration of 12ms and 32ms in Replica [□] and Matterport3D [□]. Taking an aggregate of 1 sec gives matrices with dimensions of 257×257 (Replica) and 257×101 (Matterport3D). We further down-sample each matrix by a factor of 4, and stack the ones from left and right "ear" in a third channel, resulting in the final spectrum matrices in dimensions of $65 \times 69 \times 2$ (Replica) and $65 \times 26 \times 2$ (Matterport3D).

4   The Proposed Approach

We formalize the problem as a reinforcement learning task in which the navigating robot learns a policy to catch up with a moving audio target quickly in an unknown environment. Our solution to the problem is abbreviated FSAAVN (feature self-attention audio-visual navigation). FSAAVN is composed of four main parts (Fig.4) Specifically, given the egocentric vision and audio input, our model 1) encodes visual/audio inputs into visual/audio features using **CNNs** (convolution neural networks); 2) fuse the visual and audio features with **FSA** (feature self-attention) producing the fused audio-visual embedding, 3) transform a series of such embeddings into a temporal-ware state representation using a **GRU** (gated recurrent unit); and finally 4) use an **actor-critic** network to perform action prediction, evaluation, and optimization. The robot agent repeat this process until it catches the moving audio target. We will introduce each part consecutively in the upcoming paragraphs.

As introduced in Section 3, the robot receives an audio-visual observation $O_t = (I_t, B_t)$ at the t-th step. The visual (I_t) part is encoded into visual feature vector using a CNN encoder: $f_I(I_t)$. The audio feature vector is obtained in the same way using a different CNN encoder: $f_B(B_t)$. Visual and auditory CNN encoders are constructed in the same way (from the input to output layer): `Conv8x8`, `Conv4x4`, `Conv3x3` and a 512-dim linear layer; ReLU activations are added between any two neighboring layers. We denote the concatenation of visual and audio features as $e_t^i = [f_I(I_t), f_B(B_t)]$.

To determine the relative contribution of each modal in real-time according to the varying context, we design a trainable audio-visual fusion mechanism (i.e. FSA) to transform the encoded features e_t^i to a fused embedding vector e_t^o:

$$e_t^o = \text{softmax}\left(\frac{\text{sigmoid}(W_1 Q)\,\text{sigmoid}(W_2 K)^T}{\sqrt{d}}\right) V \oplus V, \quad \text{s.t. } Q = K = V = e_t^i, \quad (1)$$

where Q, K, V are the query, key, and value input to the FSA module; W_1 and W_2 are both weight matrices to be optimized; d is a scalar factor with a value of 256; and e_t^o is the resulting fused embedding vector. We collectively denote the encoder and FSA weights as **W** hereafter for simplicity. The right part of Fig.4 can be referred to for more details.

A bidirectional GRU (with one 512-dim hidden layer) is applied to further transform a series of fused embeddings (i.e. $e_1^o \ldots e_t^o$) into a temporal-aware state representation s_t. Concretely, at time t, the GRU cell takes in both the current embedding e_t^o and the previous cell state h_{t-1} to produce s_t and h_t. Essentially, $s_t = GRU(e_t^o, h_{t-1})$.

The state vectors (i.e. $s_1 \ldots s_t$) is then fed to an actor-critic network to 1) predict the conditioned action probability distribution $\pi_{\theta_1}(a_t|s_t)$, and 2) estimate the state value $V_{\theta_2}(s_t)$. The actor and critic are implemented with a single linear layer parameterized by θ_1 and θ_2, respectively. For the sake of conciseness, we use $\boldsymbol{\theta}$ to denote the compound of θ_1 and θ_2 hereafter. The action sampler in Fig.4 samples the actual action (i.e. a_t) to execute from $\pi_{\theta_1}(a_t|s_t)$. The training aims to maximise the expected discounted return $\Re$:

$$\Re = \mathbb{E}_\pi\left[\sum_{t=1}^T \gamma^t r(s_{t-1}, a_t)\right], \quad (2)$$

where γ is a discount factor; T is the maximum number of time steps; and π is the policy of the robot agent. $r(s_{t-1}, a_t)$ is the reward given by the environment at the time step t. The reward is calculated based on three simple rules: (1) $+10$ point when the robot successfully reaches the target and executes the `Stop` action, (2) $+0.25$ point when the Manhattan distance between the robot and target is reduced, and (3) a time penalty of -0.01 on each action

performed to encourage navigation efficiency. Proximal Policy Optimization (PPO) [41] is adopted in this work to optimize (2). The entire procedure is described in Algorithm 1.

Algorithm 1: FSAAVN: feature self-attention audio-visual navigation

> **Data:** Environment $\mathcal{E}$, stochastic policies π, initial actor-critic weights $\boldsymbol{\theta}_0$, initial encoder and FSA weights $\mathbf{W}_0$, # updates M, # episode N, max time steps T.
>
> **Result:** Trained weights: $\boldsymbol{\theta}_M$ and $\mathbf{W}_M$

1 **for** *i=1, 2, ... M* **do**
2   // Run policy $\pi_{\boldsymbol{\theta}_{i-1}}$ in environment for N episodes T time steps ;
3     $\{(o_t, h_{t-1}, a_t, r_t)_i\}_{t=1}^T \leftarrow \text{roll}(\mathcal{E}, \pi_{\boldsymbol{\theta}_{i-1}}, T)$ at i-th update ;
4   Compute advantage estimates ;
5   // Optimize w.r.t. $\boldsymbol{\theta}$ and $\mathbf{W}$;
6     $\boldsymbol{\theta}_i, \mathbf{W}_i \leftarrow$ new $\boldsymbol{\theta}$ and $\mathbf{W}$ from PPO algorithm w.r.t. maximizing Equation (2);
7 **end**

While the self-attention mechanism (i.e. FSA) over the latent dimension is effective (experimentally validated in Section 5), why don't we choose to adopt attention over the temporal dimension (i.e. replace the GRU with Attention)? The reason is two fold:

- The input data for audio-visual navigation is different from the typical input we see in NLP (natural language processing) and CV (computer vision) tasks. At any time point, the observation is always incomplete, meaning the future states are unknown. As a result, learning the temporal attention can be extremely unstable.

- The maximum number of navigation steps is 500 (or one can choose a much larger number), which leads to a spacial-temporal transformer model with high capacity and expressivity, hence requiring large amount of samples/episodes to converge.

5 Experiments

We carry out experiments on the environments and datasets described in section 3. FSAAVN is benchmarked towards several state-of-the-art baselines: **SoundSpaces** [7], **SoundSpaces-EMul**, **SoundSpaces-EM**, **CMHM** [50], and **AV-WaN** [9]. SoundSpaces-EMul/-EM is the extension of basic SoundSpaces (concatenation fusion) with different audio-visual feature fusion methods: "EMul" stands for element-wise multiplication and "EM" is element-wise mean, as shown in Fig.2(b,c). Based on the widely used SPL (success weighted by path length) metric [1], we calculate several evaluation metrics, such as **SPLT** (SPL for tracking), **SSPLT** (soft SPLT), and **SRT** (success rate for tracking).

$$\text{SPLT} = \frac{1}{N}\sum_{i=1}^N \frac{S_i \cdot l_i}{\max(p_i, l_i)}, \quad \text{SSPLT} = \frac{1}{N}\sum_{i=1}^N \frac{l_i \cdot \max(0, 1 - \frac{d_i^a}{d_i})}{\max(p_i, l_i)}, \quad \text{SRT} = \frac{1}{N}\sum_{i=1}^N S_i, \quad (3)$$

where N is the number of episodes; S_i is a binary indicator of success in the i-th episode; p_i stands for the length of the executed path; l_i is the length of the shortest path from the robot's start position to the target's final position. When computing SSPLT by the end of the i-th episode, d_i^a is the robot's distance to the target, d_i is the distance from the robot's start position to the target's final position. For the definition of other evaluation metrics used

Table 1: Overall performance comparison (STDEV≤0.01) using depth and sound input.

Model	Fusion	Replica									Matterport3D								
		Telephone			Multiple heard			Multiple unheard			Telephone			Multiple heard			Multiple unheard		
		SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT
		(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)
FSAAVN	FSA	**0.541**	**0.635**	**0.925**	**0.438**	**0.541**	**0.812**	**0.182**	**0.316**	**0.358**	**0.520**	**0.585**	**0.832**	**0.438**	**0.496**	**0.844**	**0.207**	**0.299**	**0.391**
SoundSpaces	Concat	0.531	0.604	0.892	0.354	0.462	0.764	0.152	0.255	0.317	0.454	0.511	0.797	0.431	0.475	0.818	0.180	0.254	0.350
SoundSpaces-	EMul	0.493	0.597	0.861	0.430	0.522	0.770	0.168	0.304	0.326	0.457	0.523	0.801	0.433	0.481	0.821	0.182	0.258	0.355
SoundSpaces-	EM	0.487	0.592	0.816	0.435	0.531	0.796	0.154	0.258	0.319	0.481	0.543	0.817	0.435	0.492	0.832	0.183	0.266	0.375
CMHM	Concat	0.335	0.338	0.791	0.259	0.302	0.692	0.121	0.202	0.314	0.114	0.125	0.606	0.086	0.099	0.528	0.052	0.085	0.267
AV-WaN	Concat	0.218	0.224	0.764	0.220	0.271	0.533	0.010	0.189	0.233	0.111	0.114	0.409	0.012	0.034	0.093	0.010	0.043	0.057

in this research, please refer to Appx.B.1 [51]. For all metrics, completion of an episode indicates that the robot either catches the target in less than 500 steps, or selects the stop action precisely at the location of the moving target. The reported metric values are averaged over 5 trials.

We train our model with Adam (to optimize an entropy loss on the policy distribution) with a learning rate of 2.5×10^{-4} with a limit of time horizon corresponding to 500 actions in a scene. We train the framework for $40M$ steps on Replica and $60M$ on Matterport3D, which amounts to 200 and 320 GPU hours, respectively. Since there are 102 sounds from SoundSpaces, we test three different sound source splittings. (1) **Telephone**: the target sound source (telephone) is the same in the training, validation and testing sets; (2) **Multiple heard**: all 102 sounds exist in three sets; (3) **Multiple unheard**: the 102 sounds are divided into non-overlapping 73/11/18 splits for train/validation/test.

5.1 Overall performance comparison

We generally discover that depth works the best with sound input (cf. Table 2 and Appx.B.3 [51]) that coincide the results in [7]. Therefore, in Table 1, we only illustrate the metrics obtained using depth and sound input. It shows that FSAAVN using FSA fusion constantly performs the best (in boldface) on both datasets in all splitting settings.

To gain a qualitative and vivid impression of robot behavior from different approaches, we visualize the recorded navigation trajectories (using depth and sound input) on a scene map from Replica and Matterport3D datasets (Fig.5). On both datasets, we largely see that FSAAVN often executes a shorter trajectory and a higher SPLT value, implying a better performance than other methods.

5.2 Comparing different visual modalities

There can be different types of egocentric visual input such as RGB, depth and RGBD, where RGBD is a combination of RGB and depth in different channels. To understand the relative effectiveness of these visual modalities when fused with audio input, we compare the navigation performance between FSAAVN and SoundSpaces using RGB, depth, RGBD and blind (no visual input) in addition to the auditory input from the moving target. From Table 2, we can concluded that: 1) among all tested modalities, depth alone achieves the best performance that is consistent with the conclusion in [7] when tackling with a stationary sound source; 2) with the same modality, FSA performs better than simple concatenation; 3) in the case of blind, FSA seems slightly worse than concatenation, probably caused by FSA's higher flexibility and complexity than concatenation.
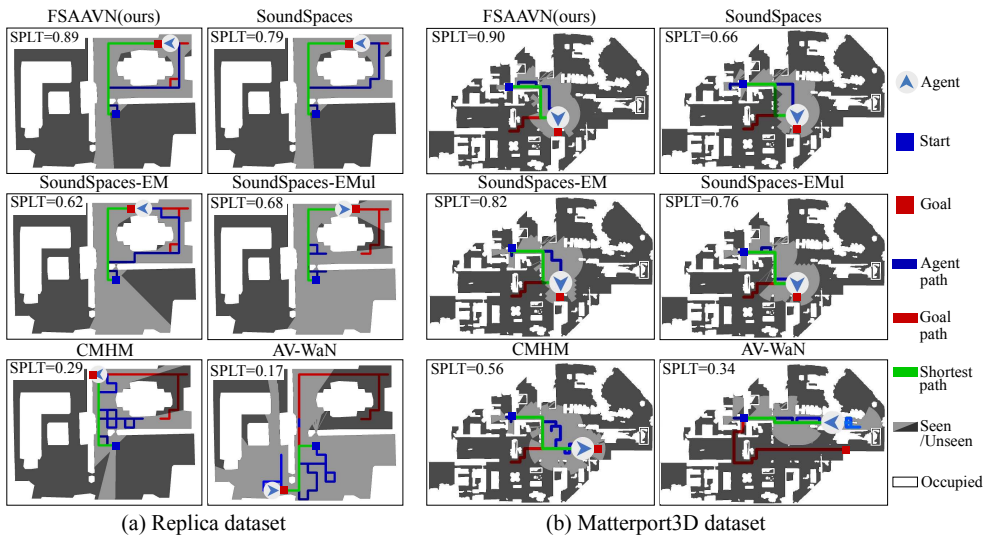
Figure 5: Navigation trajectories by the end of a particular episode from (a) Replica and (b) Matter-port3D dataset. Higher SPLT values and shorter blue paths indicate better performances.

Table 2: Performance Comparison (STDEV≤0.01) of different vision modalities.

Model	Fusion	Vision	Replica									Matterport3D								
			Telephone			Multiple heard			Multiple unheard			Telephone			Multiple heard			Multiple unheard		
			SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT
			(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)
FSAAVN	FSA	Depth	**0.541**	**0.635**	**0.925**	**0.438**	**0.541**	**0.812**	**0.182**	**0.316**	**0.358**	**0.520**	**0.585**	**0.832**	**0.438**	**0.496**	**0.844**	**0.207**	**0.299**	**0.391**
SoundSpaces	Concat	Depth	0.531	0.604	0.892	0.354	0.462	0.764	0.152	0.255	0.317	0.454	0.511	0.797	0.431	0.475	0.818	0.180	0.254	0.350
FSAAVN	FSA	RGBD	0.532	0.611	0.837	0.402	0.485	0.792	0.185	0.285	0.349	0.454	0.510	0.834	0.440	0.492	0.827	0.191	0.281	0.373
SoundSpaces	Concat	RGBD	0.527	0.605	0.835	0.393	0.475	0.756	0.182	0.276	0.339	0.435	0.502	0.798	0.412	0.469	0.809	0.186	0.277	0.369
FSAAVN	FSA	RGB	0.530	0.601	0.872	0.413	0.500	0.767	0.166	0.295	0.305	0.449	0.505	0.820	0.393	0.453	0.781	0.196	0.270	0.417
SoundSpaces	Concat	RGB	0.522	0.593	0.829	0.386	0.477	0.741	0.140	0.260	0.267	0.397	0.451	0.815	0.371	0.429	0.772	0.193	0.269	0.375
FSAAVN	FSA	Blind	0.470	0.544	0.833	0.328	0.425	0.703	0.141	0.229	0.294	0.369	0.424	0.787	0.339	0.387	0.766	0.162	0.241	0.356
SoundSpaces	Concat	Blind	0.472	0.545	0.839	0.334	0.425	0.725	0.142	0.229	0.331	0.385	0.443	0.790	0.319	0.372	0.724	0.163	0.257	0.370

5.3   Comparing different visual/audio encoders

Practically, nothing stops us from choosing a different visual/audio encoder other than CNN. Nowadays, the common options are ViT (vision transformer) [17] and Capsule [39]. Here, we will compare the audio-visual navigation performance using different visual/audio en-coder backbones (i.e. CNN, Capsule and ViT-based). For ViT, we test two variants: ViT-V and ViTScratch-V, where the former is initialised with pretrained weights while the latter is trained from random weights. The implementation details of these encoders can be found in Appx.A [51]. Table 3 shows the results of using different encoders while keeping the other parts the same. It can been seen that the overly complex encoders (Capsule and ViT-based ones) turns out to be inferior to the CNN encoder. Our assumption is that higher complexity increases the convergence difficulty and thus makes policy learning more challenging.

5.4   On dynamic modality importance

Since the environmental context and target location (relative to the robot) changes all the time during navigation, we assume that the relative influence of audio and visual input on

Table 3: Performance Comparison (STDEV≤0.01) of different visual/audio encoder backbones.

| Model | Fusion | Encoder | Replica | | | | | | | | | Matterport3D | | | | | | | | |
			Telephone			Multiple heard			Multiple unheard			Telephone			Multiple heard			Multiple unheard		
			SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT	SPLT	SSPLT	SRT
			(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)	(↑)
FSAAVN	FSA	CNN	**0.541**	**0.635**	**0.925**	**0.438**	**0.541**	**0.812**	**0.182**	**0.316**	**0.358**	**0.520**	**0.585**	**0.832**	**0.438**	**0.496**	**0.844**	**0.207**	**0.299**	**0.391**
SoundSpaces	Concat	CNN	0.531	0.604	0.892	0.354	0.462	0.764	0.152	0.255	0.317	0.454	0.511	0.797	0.431	0.475	0.818	0.180	0.254	0.350
ViT-V	Concat	ViT	0.521	0.584	0.871	0.329	0.415	0.713	0.138	0.233	0.304	0.412	0.465	0.797	0.012	0.188	0.027	0.013	0.170	0.020
Capsule	Concat	Capsule	0.426	0.503	0.810	0.262	0.372	0.580	0.154	0.278	0.330	0.317	0.382	0.742	0.246	0.302	0.623	0.178	0.255	0.445
ViTScratch-V	Concat	ViT	0.293	0.375	0.700	0.220	0.321	0.529	0.089	0.199	0.189	0.325	0.388	0.762	0.265	0.312	0.691	0.167	0.232	0.422
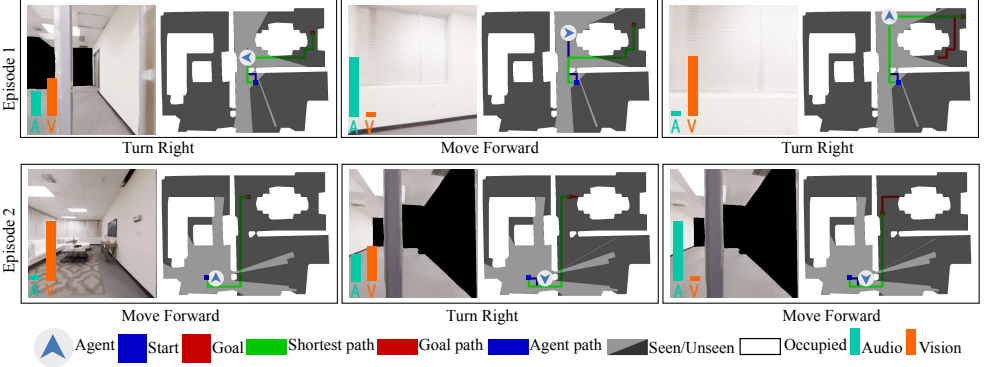
Figure 6: Dynamic visual and echo impact for two episodes. Columns corresponds to three sampled time steps. The green and orange bars represent the importance of audio and vision, respectively.

the robot's action can vary at different time points. To quantify visual (audio) impact, we replace the visual (audio) input with random noise; and then we compute the visual (audio) impact score as the absolute difference (normalized) of the logarithmic action probabilities from the semi-corrupted model and the intact one. Fig.6 shows the impact scores (for two episodes in two rows) on the egocentric robot view at different time steps. We can see that FSAAVN dynamically re-weight the modalities (according to the current surroundings) while chasing after the moving audio target.

5.5 Visualization of learned features and states

In FSAAVN framework (cf. Section 4), the vision encoder generates visual feature $f_I(I_t)$; the auditory encoder produces audio feature $f_B(B_t)$; and the GRU transforms historical feature vectors into audio-visual state representations s_t. The disengagement quality of these learned features and states is important to the downstream policy learning. In Fig.7(a), we examine the semantics of visual features by overlaying the output of the visual encoder (from different layers) over the RGB images. It is easy to see that the visual encoder has learned to pay more attention to the area (in red color) where the robot can walk. This effect becomes more evident as the encoder becomes deeper. It is more challenging to visualize the disengagement quality of audio features and the state representations, as a result, we choose to perform dimension reduction (to two dimensions) and clustering using UMAP [32]. The UMAP result is shown in Fig.7(b) with a color coding representing the action selected by the robot. Seen from Fig.7(b), the learned audio features and state representations are naturally correlated with the robot's action selection.
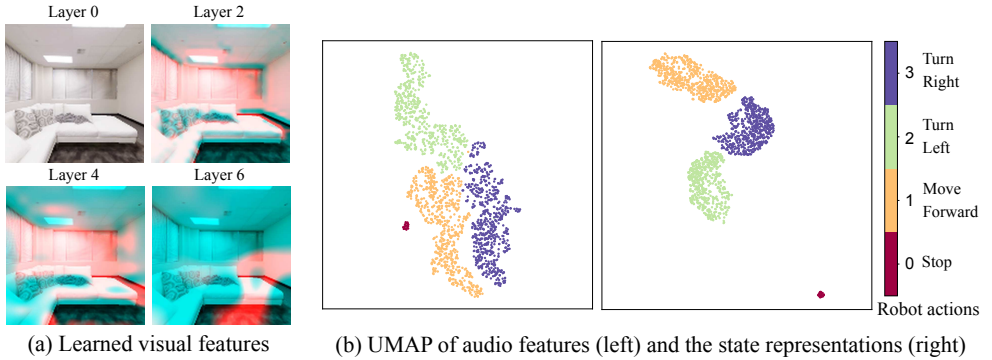
(a) Learned visual features   (b) UMAP of audio features (left) and the state representations (right)

Figure 7: Visualization of visual feature, audio feature, and state representations from Replica dataset.

6   Conclusion

To realize a more effective (than the existing methods) audio-visual feature fusion strategy during audio-visual embodied navigation, we design a trainable Feature Self-Attention (FSA) module that determines the relative contribution of visual/audio modal in real-time in accordance with the ever-changing context. We propose an end-to-end framework (FSAAVN: feature self-attention audio-visual navigation) incorporating FSA to train robots to catch up with a moving audio target. FSAAVN is easy to train since it requires no extra aid like topology graph and sound semantics. Our comprehensive experiments validate the superior performance (both quantitatively and qualitatively) of FSAAVN in comparison with the state-of-the-arts. We also carry out a set of thorough ablation studies on mainstream visual modalities, signal (visual/audio) encoders and audio-visual fusion strategies, providing useful insights for practitioners and researchers in this filed.

Acknowledgments

References

[1] Peter Anderson, Angel X. Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, and Amir Roshan Zamir. On evaluation of embodied navigation agents. *CoRR*, abs/1807.06757, 2018. URL http://arxiv.org/abs/1807.06757.

[2] Federico Boniardi, Abhinav Valada, Rohit Mohan, Tim Caselitz, and Wolfram Burgard. Robot localization in floor plans using a room layout edge extraction network. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2019, Macau, SAR, China, November 3-8, 2019*, pages 5291–5297. IEEE, 2019. doi: 10.1109/IROS40897.2019.8967847. URL https://doi.org/10.1109/IROS40897.2019.8967847.

[3] Angelo Cangelosi, Josh Bongard, Martin H Fischer, and Stefano Nolfi. Embodied intelligence. In *Springer Handbook of Computational Intelligence*, pages 697–714. Springer, 2015. URL https://link.springer.com/chapter/10.1007/978-3-662-43505-2_37.

[4] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision, 3DV 2017, Qingdao, China, October 10-12, 2017*, pages 667–676. IEEE Computer Society, 2017. doi: 10.1109/3DV.2017.00081. URL https://doi.org/10.1109/3DV.2017.00081.

[5] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural SLAM. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HklXn1BKDH.

[6] Devendra Singh Chaplot, Helen Jiang, Saurabh Gupta, and Abhinav Gupta. Semantic curiosity for active visual learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 309–326. Springer, 2020. doi: 10.1007/978-3-030-58539-6_19. URL https://doi.org/10.1007/978-3-030-58539-6_19.

[7] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3d environments. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VI*, volume 12351 of *Lecture Notes in Computer Science*, pages 17–36. Springer, 2020. doi: 10.1007/978-3-030-58539-6_2. URL https://doi.org/10.1007/978-3-030-58539-6_2.

[8] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*

2021, virtual, June 19-25, 2021, pages 15516–15525. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Semantic_Audio-Visual_Navigation_CVPR_2021_paper.html.

[9] Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=cR91FAodFMe.

[10] Changan Chen, Wei Sun, David Harwath, and Kristen Grauman. Learning audio-visual dereverberation. *CoRR*, abs/2106.07732, 2021. URL https://arxiv.org/abs/2106.07732.

[11] Changan Chen, Ruohan Gao, Paul Calamia, and Kristen Grauman. Visual acoustic matching. *CoRR*, abs/2202.06875, 2022. URL https://arxiv.org/abs/2202.06875.

[12] Changan Chen, Carl Schissler, Sanchit Garg, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W. Robinson, and Kristen Grauman. Soundspaces 2.0: A simulation platform for visual-acoustic learning. *CoRR*, abs/2206.08312, 2022. doi: 10.48550/arXiv.2206.08312. URL https://doi.org/10.48550/arXiv.2206.08312.

[13] Kevin Chen, Juan Pablo de Vicente, Gabriel Sepulveda, Fei Xia, Alvaro Soto, Marynel Vázquez, and Silvio Savarese. A behavioral approach to visual navigation with graph localization networks. In Antonio Bicchi, Hadas Kress-Gazit, and Seth Hutchinson, editors, *Robotics: Science and Systems XV, University of Freiburg, Freiburg im Breisgau, Germany, June 22-26, 2019*, 2019. doi: 10.15607/RSS.2019.XV.010. URL https://doi.org/10.15607/RSS.2019.XV.010.

[14] Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SyMWn05F7.

[15] Jesper Haahr Christensen, Sascha Hornauer, and Stella X. Yu. Batvision: Learning to see 3d spatial layout with two ears. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 1581–1587. IEEE, 2020. doi: 10.1109/ICRA40945.2020.9196934. URL https://doi.org/10.1109/ICRA40945.2020.9196934.

[16] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ab6b331e94c28169d15cca0cb3bbc73e-Abstract.html.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[18] Ross Flom and Lorraine E Bahrick. The development of infant discrimination of affect in multimodal and unimodal stimulation: The role of intersensory redundancy. *Developmental psychology*, 43(1):238, 2007.

[19] Madeleine Fortin, Patrice Voss, Catherine Lord, Maryse Lassonde, Jens Pruessner, Dave Saint-Amour, Constant Rainville, and Franco Lepore. Wayfinding in the blind: larger hippocampal volume and supranormal spatial navigation. *Brain*, 131(11):2995–3005, 2008.

[20] Chuang Gan, Hang Zhao, Peihao Chen, David D. Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7052–7061. IEEE, 2019. doi: 10.1109/ICCV.2019.00715. URL https://doi.org/10.1109/ICCV.2019.00715.

[21] Chuang Gan, Deng Huang, Hang Zhao, Joshua B. Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10475–10484. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01049. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Gan_Music_Gesture_for_Visual_Sound_Separation_CVPR_2020_paper.html.

[22] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B. Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pages 9701–9707. IEEE, 2020. doi: 10.1109/ICRA40945.2020.9197008. URL https://doi.org/10.1109/ICRA40945.2020.9197008.

[23] Chuang Gan, Yi Gu, Siyuan Zhou, Jeremy Schwartz, Seth Alter, James Traer, Dan Gutfreund, Joshua B. Tenenbaum, Josh H. McDermott, and Antonio Torralba. Finding fallen objects via asynchronous audio-visual integration. *CoRR*, abs/2207.03483, 2022. doi: 10.48550/arXiv.2207.03483. URL https://doi.org/10.48550/arXiv.2207.03483.

[24] Ruohan Gao, Changan Chen, Ziad Al-Halah, Carl Schissler, and Kristen Grauman. Visualechoes: Spatial image representation learning through echolocation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IX*, volume 12354 of *Lecture Notes in Computer Science*, pages 658–676. Springer, 2020. doi: 10.1007/978-3-030-58545-7_38. URL https://doi.org/10.1007/978-3-030-58545-7_38.

[25] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. IQA: visual question answering in interactive environments. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4089–4098. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018. 00430. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Gordon_IQA_Visual_Question_CVPR_2018_paper.html.

[26] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7272–7281. IEEE Computer Society, 2017. doi: 10.1109/CVPR. 2017.769. URL https://doi.org/10.1109/CVPR.2017.769.

[27] Dinesh Jayaraman and Kristen Grauman. Learning to look around: Intelligently exploring unseen environments for unknown tasks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1238–1247. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018. 00135. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Jayaraman_Learning_to_Look_CVPR_2018_paper.html.

[28] Péter Karkus, Shaojun Cai, and David Hsu. Differentiable slam-net: Learning particle SLAM for visual navigation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2815–2825. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Karkus_Differentiable_SLAM-Net_Learning_Particle_SLAM_for_Visual_Navigation_CVPR_2021_paper.html.

[29] Martin Lohmann, Jordi Salvador, Aniruddha Kembhavi, and Roozbeh Mottaghi. Learning about objects by learning to interact with them. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/291597a100aadd814d197af4f4bab3a7-Abstract.html.

[30] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 275–285. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00034. URL https://doi.org/10.1109/ICCV48922.2021.00034.

[31] Sagnik Majumder, Changan Chen, Ziad Al-Halah, and Kristen Grauman. Few-shot audio-visual learning of environment acoustics. *CoRR*, abs/2206.04006, 2022. doi: 10. 48550/arXiv.2206.04006. URL https://doi.org/10.48550/arXiv.2206.04006.

[32] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

[33] Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andy Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, Dharshan Kumaran, and Raia Hadsell. Learning to navigate in complex environments. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=SJMGPrcle.

[34] Steven D. Morad, Roberto Mecca, Rudra P. K. Poudel, Stephan Liwicki, and Roberto Cipolla. Embodied visual navigation with automatic curriculum learning in real environments. *IEEE Robotics Autom. Lett.*, 6(2):683–690, 2021. doi: 10.1109/LRA.2020. 3048662. URL https://doi.org/10.1109/LRA.2020.3048662.

[35] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/15825aee15eb335cc13f9b559f166ee8-Abstract.html.

[36] Kranti Kumar Parida, Siddharth Srivastava, and Gaurav Sharma. Beyond image to depth: Improving depth prediction using echoes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8268–8277. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Parida_Beyond_Image_to_Depth_Improving_Depth_Prediction_Using_Echoes_CVPR_2021_paper.html.

[37] Senthil Purushwalkam, Sebastia Vicenc Amengual Gari, Vamsi Krishna Ithapu, Carl Schissler, Philip Robinson, Abhinav Gupta, and Kristen Grauman. Audio-visual floorplan reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1163–1172. IEEE, 2021. doi: 10.1109/ICCV48922.2021.00122. URL https://doi.org/10.1109/ICCV48922.2021.00122.

[38] Santhosh K. Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 400–418. Springer, 2020. doi: 10.1007/978-3-030-58558-7_24. URL https://doi.org/10.1007/978-3-030-58558-7_24.

[39] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances*

in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 3856–3866, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2cad8fa47bbef282badbb8de5374b894-Abstract.html.

[40] Manolis Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, and Vladlen Koltun. Habitat: A platform for embodied AI research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9338–9346. IEEE, 2019. doi: 10.1109/ICCV.2019.00943. URL https://doi.org/10.1109/ICCV.2019.00943.

[41] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

[42] 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 A. Newcombe. The replica dataset: A digital replica of indoor spaces. *CoRR*, abs/1906.05797, 2019. URL http://arxiv.org/abs/1906.05797.

[43] Xiaoqiang Teng, Deke Guo, Yulan Guo, Xiaolei Zhou, and Zhong Liu. Cloudnavi: Toward ubiquitous indoor navigation service with 3d point clouds. *ACM Trans. Sens. Networks*, 15(1):1:1–1:28, 2019. doi: 10.1145/3216722. URL https://doi.org/10.1145/3216722.

[44] Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5601–5611. Computer Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Tian_Can_Audio-Visual_Integration_Strengthen_Robustness_Under_Multimodal_Attacks_CVPR_2021_paper.html.

[45] Ethan Tracy and Navinda Kottege. Catchatter: Acoustic perception for mobile robots. *IEEE Robotics Autom. Lett.*, 6(4):7209–7216, 2021. doi: 10.1109/LRA.2021.3094492. URL https://doi.org/10.1109/LRA.2021.3094492.

[46] Abhinav Valada and Wolfram Burgard. Deep spatiotemporal models for robust proprioceptive terrain classification. *Int. J. Robotics Res.*, 36(13-14):1521–1539, 2017. doi: 10.1177/0278364917727062. URL https://doi.org/10.1177/0278364917727062.

[47] Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11612–11621. Computer

Vision Foundation / IEEE, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Valverde_There_Is_More_Than_Meets_the_Eye_Self-Supervised_Multi-Object_Detection_CVPR_2021_paper.html.

[48] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: learning near-perfect pointgoal navigators from 2.5 billion frames. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=H1gX8C4YPr.

[49] Teresa Wilcox, Rebecca Woods, Catherine Chapa, and Sarah McCurry. Multisensory exploration and object individuation in infancy. *Developmental psychology*, 43(2):479, 2007.

[50] Abdelrahman Younes, Daniel Honerkamp, Tim Welschehold, and Abhinav Valada. Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds. *CoRR*, abs/2111.14843, 2021. URL https://arxiv.org/abs/2111.14843.

[51] Yinfeng Yu, Lele Cao, Fuchun Sun, Xiaohong Liu, and Liejun Wang. Pay self-attention to audio-visual navigation. *arXiv preprint arXiv:2210.01353*, 2022. doi: 10.48550/arXiv.2210.01353. URL https://arxiv.org/abs/2210.01353.

[52] Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, and Xiaohong Liu. Sound adversarial audio-visual navigation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.