Towards Generalisable Audio Representations for Audio-Visual Navigation

Shunqi Mao Chaoyi Zhang Heng Wang Weidong Cai
School of Computer Science, University of Sydney, Australia
{smao7434, czha5168, hwan9147}@uni.sydney.edu.au, tom.cai@sydney.edu.au

Abstract

In audio-visual navigation (AVN), an intelligent agent needs to navigate to a constantly sound-making object in complex 3D environments based on its audio and visual perceptions. While existing methods attempt to improve the navigation performance with preciously designed path planning or intricate task settings, none has improved the model generalisation on unheard sounds with task settings unchanged. We thus propose a contrastive learning-based method to tackle this challenge by regularising the audio encoder, where the sound-agnostic goal-driven latent representations can be learnt from various audio signals of different classes. In addition, we consider two data augmentation strategies to enrich the training sounds. We demonstrate that our designs can be easily equipped to existing AVN frameworks to obtain an immediate performance gain (13.4%↑ in SPL on Replica and 12.2%↑ in SPL on MP3D). Our project is available at https://AV-GeN.github.io/.

1. Introduction

Intelligent robots ought to be capable of perceiving, interpreting, and acting, given the multi-sensory signals from the surrounding environment. The recent SoundSpaces challenge [3] provides a desired testbed for such ability, where an embodied agent explores complex 3D environments and finds the positions of sounding objects based on visual and auditory perceptions, together with a feasible baseline AV-NAV to learn step-by-step actions for navigation with multi-modality inputs. In [7], the authors decomposed AVN into predicting the relative position of the sound source and navigating to the position given visual inputs. The AV-WaN [4] further enabled agents to navigate through intermediate waypoints with an occupancy map. However, these attempts generalise poorly to unheard sounds. Although learning in complex scenarios with multiple non-stationary audio sources could enhance the generalisation [11], it requires complicated reformulations of the environment. More specifically, several works have been proposed to design more complicated tasks to facilitate navigation, such as distractor attacks [12], short-duration sounds [2], exploration [6], and sound separation [8], while none of them deal with the overfitting problem on training sounds.

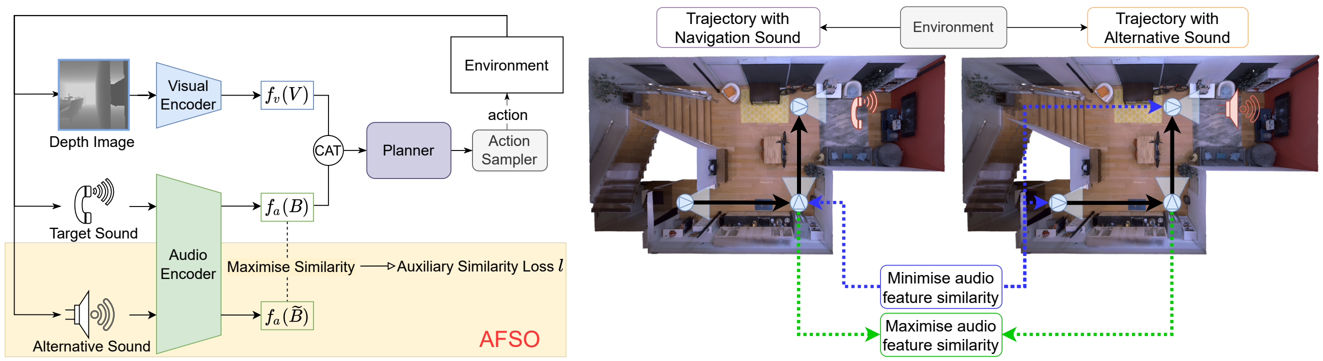
To reduce the generalisation errors, we propose a novel Audio Feature Similarity Optimisation (AFSO) method, where the audio features of distinct sounds, which carry identical goal-driven information, will be optimised to be closer in latent space. Moreover, we augment the sounds to prevent learning biased audio encoders. Our method can be conveniently applied to any learning-based AVN paradigms without substantially altering the task definition. Evaluated on two large-scale benchmarking datasets Replica [9] and MP3D [1], our method outperforms the existing frameworks by a large margin on the SoundSpaces challenge.

2. Method

2.1. Audio Feature Similarity Optimisation

Modern AVN frameworks encode the acoustic signals on the binaural audio spectrograms. While aiming to learn the source-receiver spatial relationships (*i.e.*, the positions of sounding objects w.r.t. the agent), they are easily subject to the distinction of different sound types (*e.g.*, telephone or speaker). Inspired by recent contrastive learning strategy [5], we propose the Audio Feature Similarity Optimisation (AFSO) method to alleviate the effects of sound types and focus on learning spatial relationships. Guiding the audio encoder with auxiliary similarity loss as depicted in Fig. 1a, we explicitly maximise the feature similarity between two audio observations that are sourced from different sounds but at the same relative position to the agent, and minimise the one between two audio observations at different relative sounding positions. In this way, the audio encoder can now focus on goal-driven patterns that are indicative of the target position without overfitting to specific sounds.

Specifically, a pair of binaural audio signals are considered similar only if they are sourced from the identical emitting-receiver positions within the same scene. Such two audio signals will form a positive pair, whereas all other pairs will be reckoned as negative samples. We illustrate an example in Fig. 1b. To efficiently form the training



(a) (b)

Figure 1. Our proposed audio feature similarity optimisation (AFSO) method. (a) Schematic illustration of how our AFSO method is plugged into a generic AVN framework. (b) The graph shows two identical trajectories with telephone and speaker as source sounds. The AFSO method maximises the feature similarity of positive audio-pairs of equivalent source-receiver spatial relationships and minimises the similarity between the negative pairs.

pairs, we directly simulate the second element in the pair instead of searching for the matched ones from collected trajectories. For each binaural audio signals b_k in a training batch, we simulate positive pairing acoustic signal $\widetilde{b}_k$ by convolving the room impulse response at the current step with an alternative type of source sound. The audio data is then transformed to binaural spectrograms as inputs to the audio encoder. However, such a formulation could potentially introduce false-negative (FN) pairs, where identical or similar audio observations in the trajectories may be treated as negative samples. To reduce the occurrence of FN pairs, we only compute the alternative audio and the similarity loss for a randomly sampled subset of N audio observations from the trajectories. Therefore, we obtain $2N$ data samples per batch including N original sounds and N corresponding simulated audio signals. Following [5], for a positive pair of audio observation (b_i, b_j), we calculate the auxiliary similarity loss using InfoNCE loss [10] as $\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(f_a(b_i), f_a(b_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{k \neq i} \exp(\mathrm{sim}(f_a(b_i), f_a(b_k)/\tau)}$, where f_a represents the audio encoder, sim denotes the cosine similarity function $\mathrm{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$, and τ is a temperature parameter. We apply a weight factor w to the similarity loss $L = \sum_B l_{i,j}$ and combine it with standard AVN losses for optimisation.

2.2. Source Sound Augmentation

To enrich the training sounds distribution, we also apply two sound augmentation strategies. At each episode, we augment the source sounds with the following techniques: 1) Sound Reverse: We reverse the input audio signals with a probability of p; 2) Sound Mix-up: We sample two audios signals (potentially reversed) and mix them as $x_{mix} = \lambda x_1 + (1 - \lambda) x_2$, where λ is a scalar sampled from the symmetric Beta distribution $\lambda \sim Beta(\alpha, \alpha)$.

3. Experiments

We evaluate our method by deploying it to two state-of-the-art audio-visual models, AV-NAV [3] and AV-WaN [4], on two benchmarking datasets, Replica [9] and MP3D [1], following their official dataset split [3]. We reproduce two baselines with default hyperparameters specified in papers. For our AFSO and sound augmentation strategies, we set the weight factor w as 0.1, the temperature τ as 0.07, the audio batch N as 256, the reverse probability p as 0.5, and the mix-up factor α as 1. We report SPL, SR and SNA as evaluation metrics [3, 4].

We evaluate all methods on test splits with *unseen* scenes and *unheard* sounds. As presented in Tab. 1, our method significantly improves the performance of the baseline models on both datasets with around 12% growth in SPL, which demonstrates our methods' ability in improving the generalisation of the audio encoder and preventing overfitting.

	Replica			*MP3D**		
	SPL↑	SR↑	SNA↑	SPL↑	SR↑	SNA↑
AV-NAV	38.2	45.2	21.5	26.3	43.6	11.8
AV-NAV + Ours	**51.4**	**64.4**	**30.4**	**37.1**	**55.8**	**19.1**
AV-WaN	35.7	48.4	28.5	36.2	57.4	27.4
AV-WaN + Ours	**49.1**	**69.8**	**38.6**	**48.4**	**73.9**	**36.9**

Table 1. Quantitative comparisons with AVN SOTAs in test splits on Replica and MP3D (* denote the SoundSpaces challenge [3]).

4. Conclusion

We introduce a contrastive learning-based approach AFSO to regularise the audio encoder to distill goal-driven representations for AVN. Our method can be easily deployed to existing frameworks, and substantially enlarges their generalisation ability on unfamiliar sound sources.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *Proceedings of the International Conference on 3D Vision (3DV)*, pages 667–676, 2017. 1, 2

[2] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15511–15520, 2021. 1

[3] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 17–36, 2020. 1, 2

[4] Changan Chen, Sagnik Majumder, Al-Halah Ziad, Ruohan Gao, Santhosh Kumar Ramakrishnan, and Kristen Grauman. Learning to set waypoints for audio-visual navigation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 1, 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020. 1, 2

[6] Victoria Dean, Shubham Tulsiani, and Abhinav Gupta. See, hear, explore: Curiosity via audio-visual association. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 14961–14972, 2020. 1

[7] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B. Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707, 2020. 1

[8] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. Move2hear: Active audio-visual source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 275–285, 2021. 1

[9] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1, 2

[10] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1906.05797*, 2018. 2

[11] 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. *arXiv preprint arXiv:2111.14843*, 2021. 1

[12] Yinfeng Yu, Wenbing Huang, Fuchun Sun, Changan Chen, Yikai Wang, and Xiaohong Liu. Sound adversarial audio-visual navigation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. 1