BMVC 2022

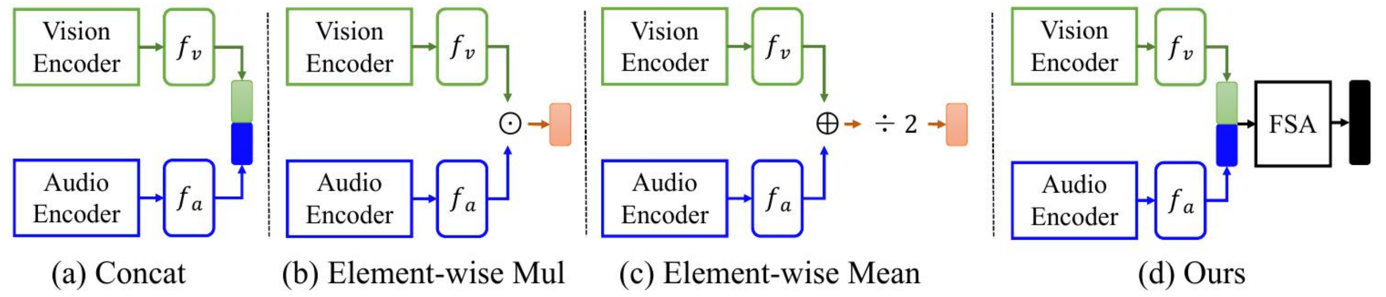
Pay Self-Attention to Audio-Visual Navigation

Yinfeng Yu, Lele Cao, Fuchun Sun, Xiaohong Liu, Liejun Wang

Tsinghua University (THU)

Motivation

- The relative contribution of each modal in real-time is dynamically different[1] at a different time step.
- The current summary for audio-visual fusion in audio-visual navigation is shown in the following figure.



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

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

Contributions

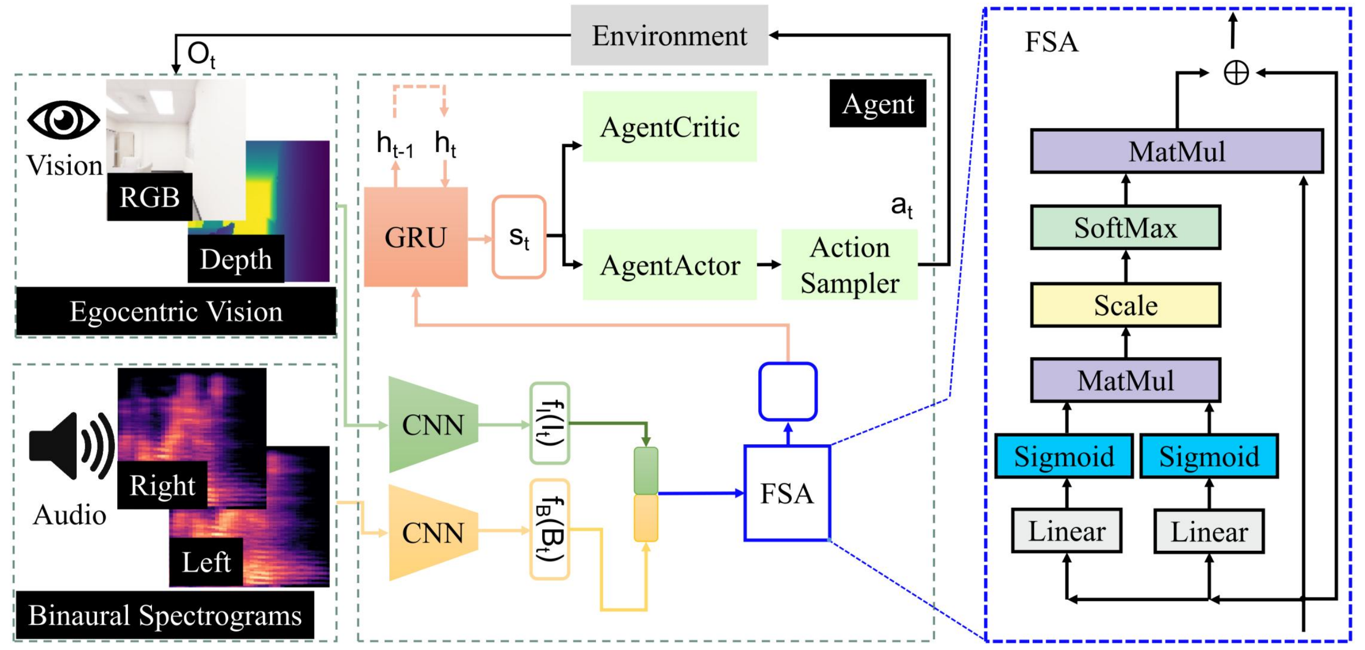
- We propose a end-to-end framework (FSAAVN) to address a currently under-researched problem: audio-visual navigation to chase a moving sound target.
- We design a novel audio-visual fusion module (FSA) to learn a context-aware strategy to determine the relative contribution of each modal in real-time.
- 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 (above Figure) and visual/audio encoder [The considered encoders: CNN (convolution neural network), ViT (vision transformer), Capsule.] provides useful insights for future practitioners in this field.

Task

Agent
Moving Audio Goal

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.

Neural network model



The overview of FSAAVN: Feature Self-Attention Audio-Visual embodied Navigation.

Results

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

Table 1: shows that FSAAVN using FSA fusion constantly performs the best (in boldface) on both datasets in all splitting settings.

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

From Table 2, we can concluded that:

1) among all tested modalities, depth alone achieves the best performance;

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.

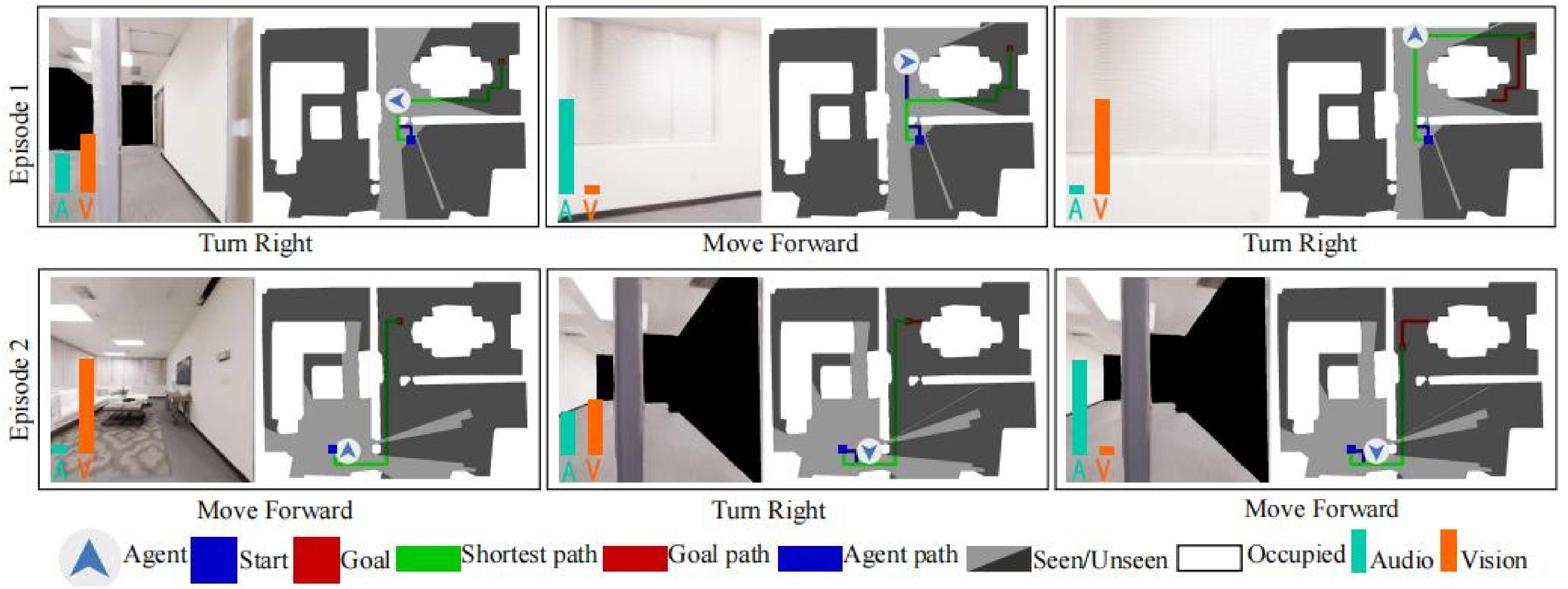
Table 3: Performance Comparison (STDEV≤0.01) of different visual/audio encoder backbones.

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

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.

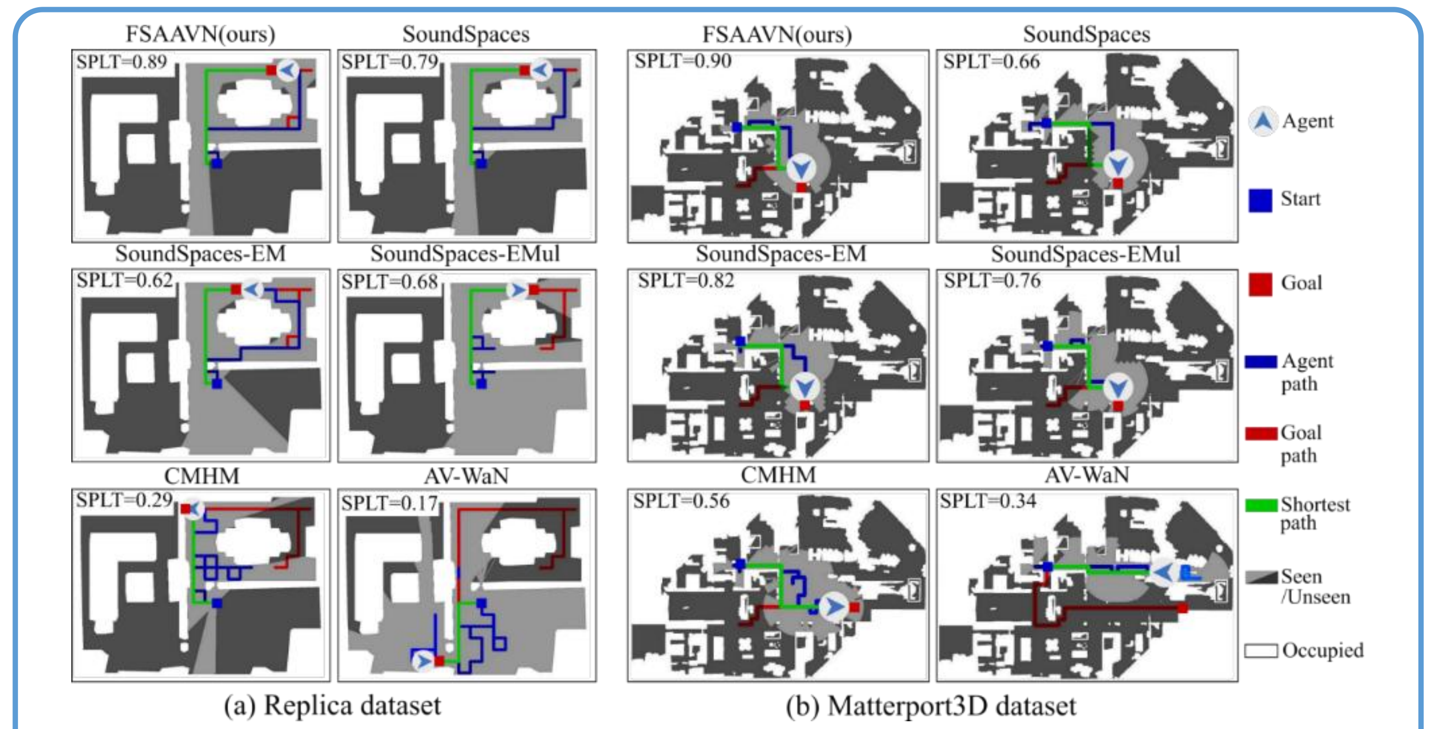


Dynamic visual and echo impact for two episodes. Columns corresponds to three sampled time steps. The green and orange bars represent the importance of audio and vision, respectively.

The left figure shows the impact scores (for two episodes in two rows) on the egocentric robot view at different time steps.

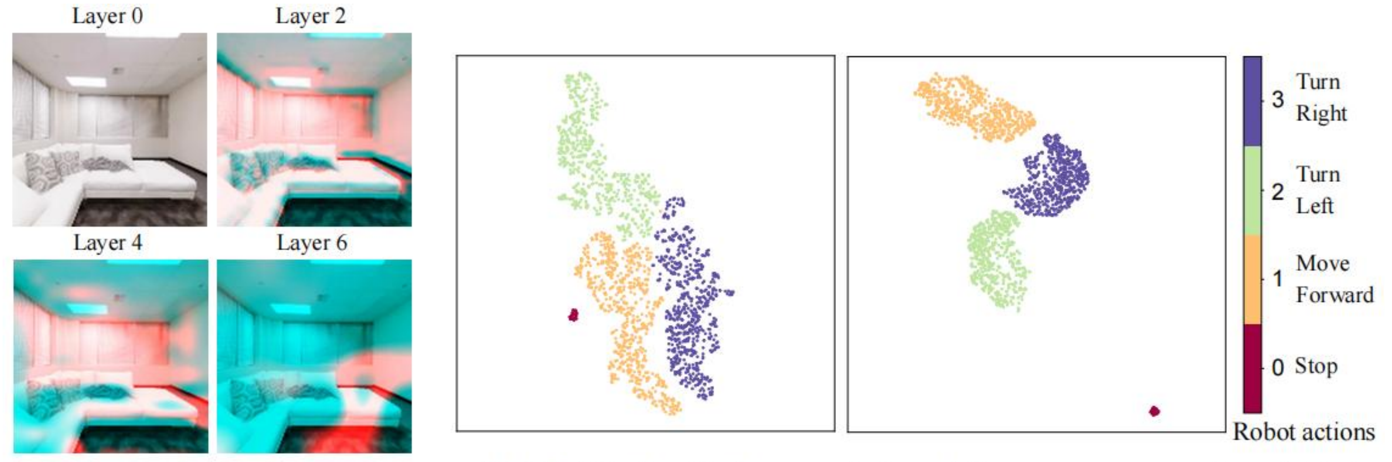
We can see that FSAAVN dynamically re-weight the modalities (according to the current surroundings)

while chasing after the moving audio target.



(a) Replica dataset (b) Matterport3D dataset

Navigation trajectories by the end of a particular episode from (a) Replica and (b) Matterport3D dataset. Higher SPLT values and shorter blue paths indicate better performances.



(a) Learned visual features (b) UMAP of audio features (left) and the state representations (right)

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

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.

Reference

[1] C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Project & Code

https://yyf17.github.io/FSAAVN