AV-GeN

Generalisable Audio-Visual Navigation Framework

Presented by

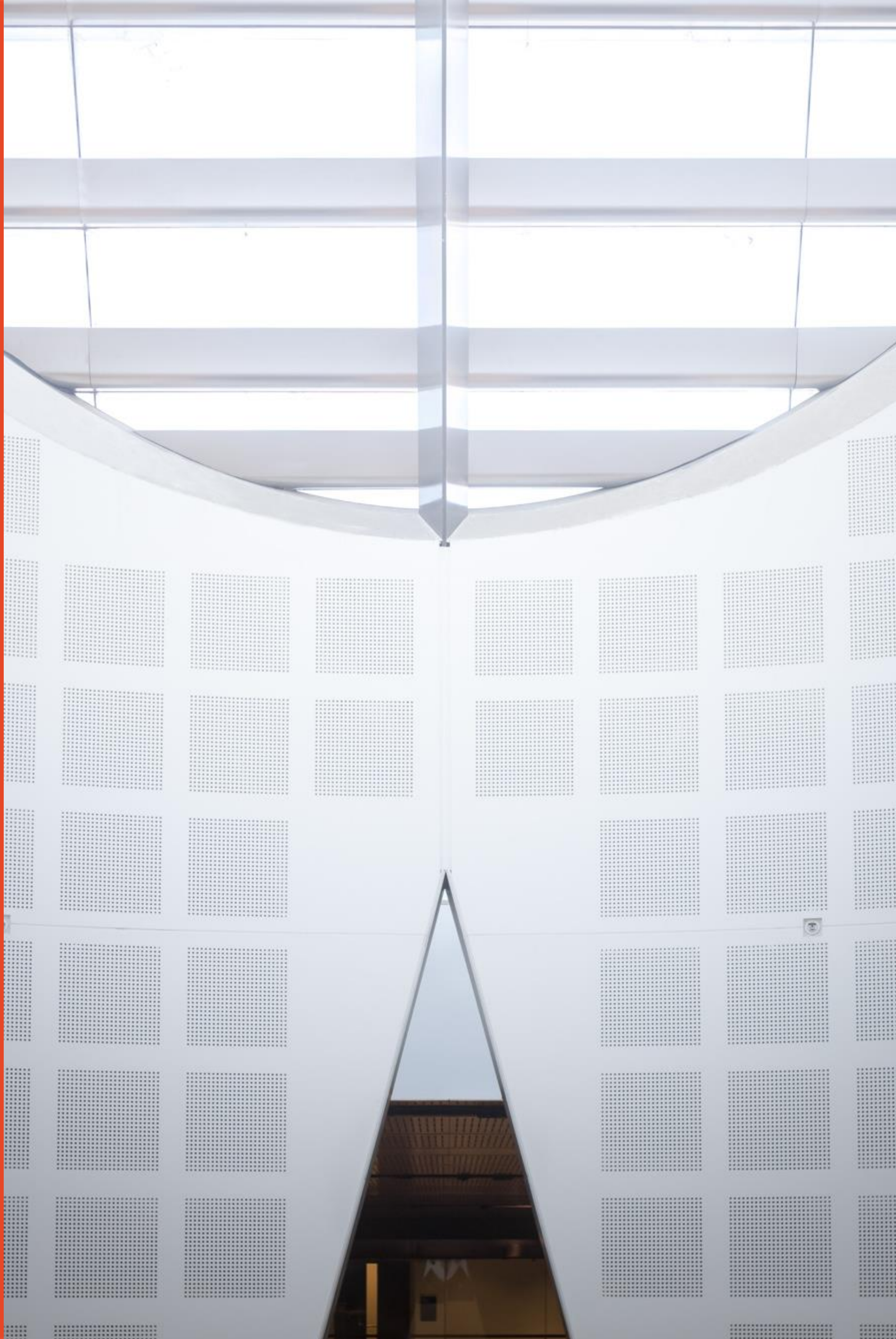
Shunqi Mao, BCST (Advanced) (Hons)

School of Computer Science

Supervised by

A/Prof Weidong (Tom) Cai

THE UNIVERSITY OF SYDNEY

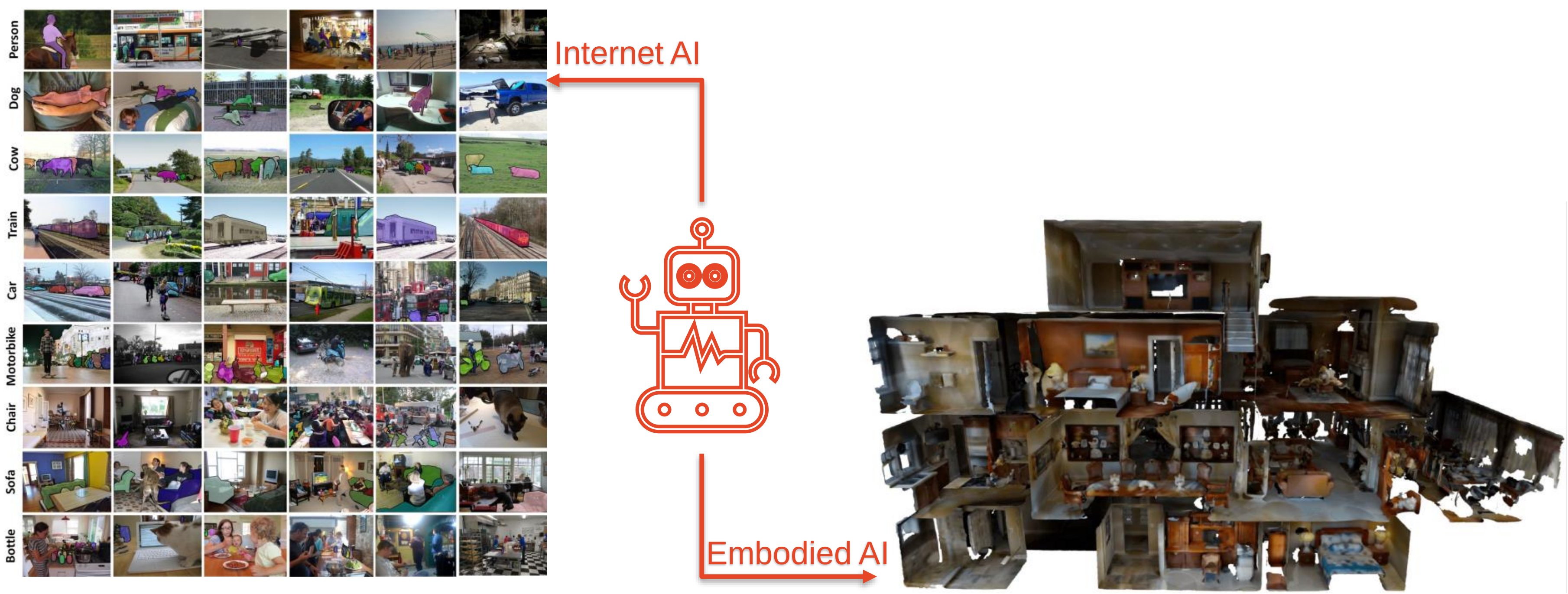


Outline

- Motivation
- Background
- Methods
- Results
- Conclusion and Future Work

Motivation - Embodied AI

– Learn from environments instead of randomized datasets.
– Experienced based on interactions instead of fixed inputs/targets.



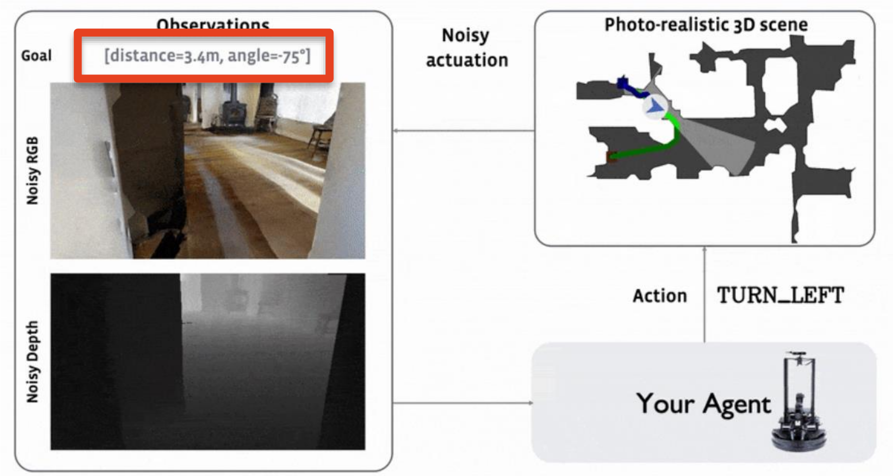
Motivation - Goal Oriented Navigation

– Navigate to **goal** positions with motion commands

Motivation - Goal Oriented Navigation

- Navigate to **goal** positions with motion commands

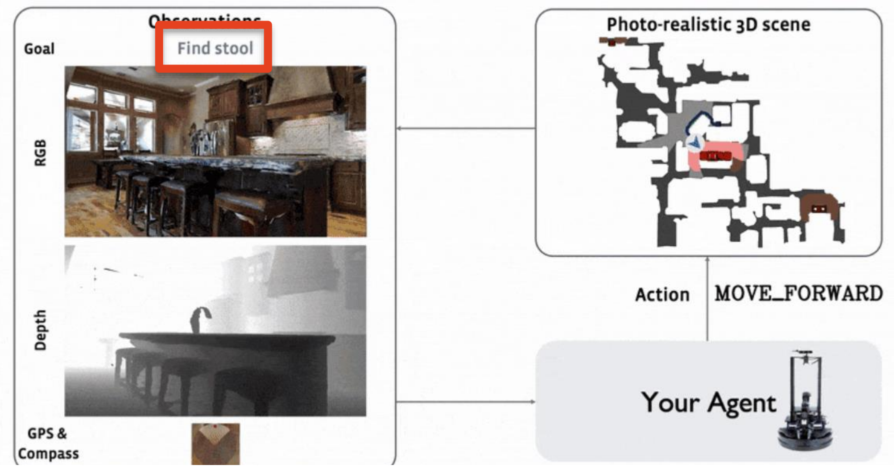
- In different tasks, **goal** could be defined differently
  - Point
  - Object
  - Image
  - Language
  - Audio
  - …

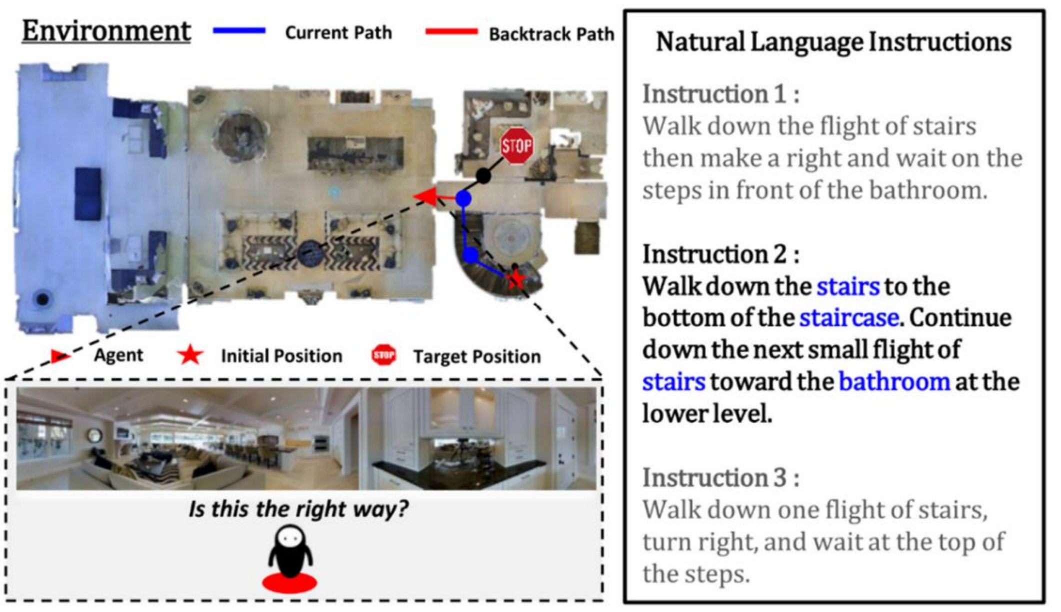


Motivation - Goal Oriented Navigation

- Navigate to **goal** positions with motion commands

- In different tasks, **goal** could be defined differently
  - Point
  - Object
  - Image
  - Language
  - Audio
  - …





Audio-Visual Navigation (AVN)

Intelligent agent should also be able to hear!
- Multi-sensory inputs: vision + acoustic signals

Actions:
- Move Forward 0.5m
- Turn Left
- Turn Right
- Stop

Criteria
- Accurate Stop
- Short Path

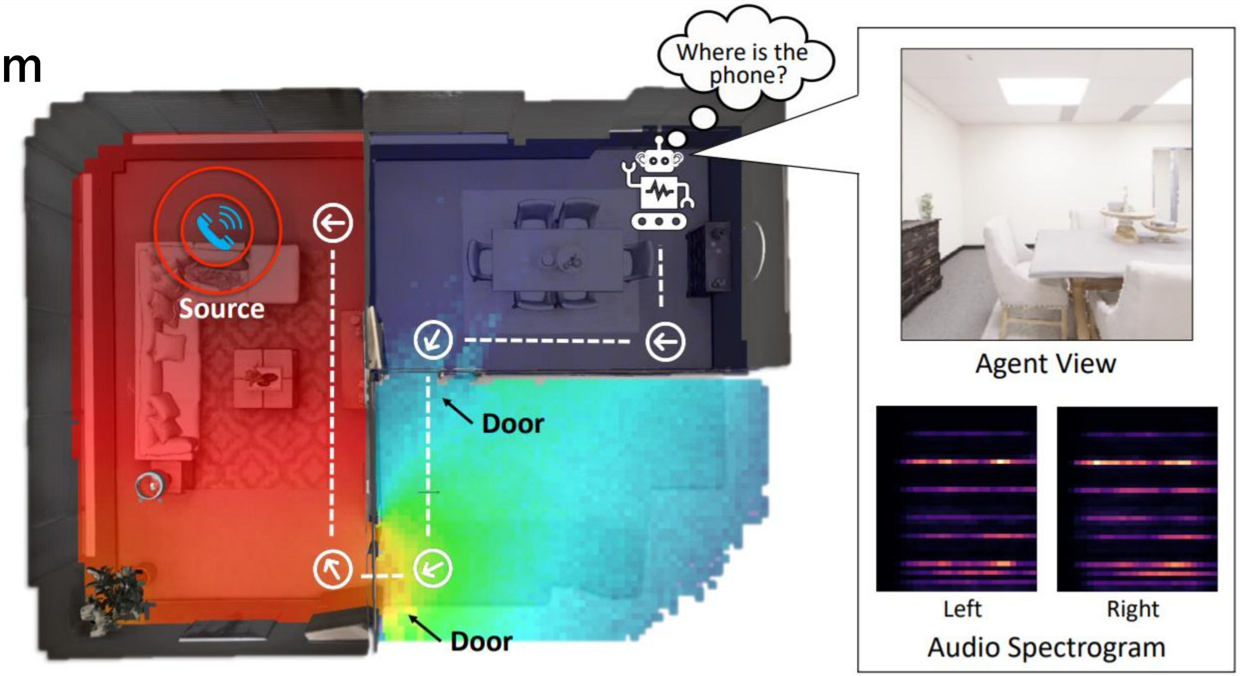
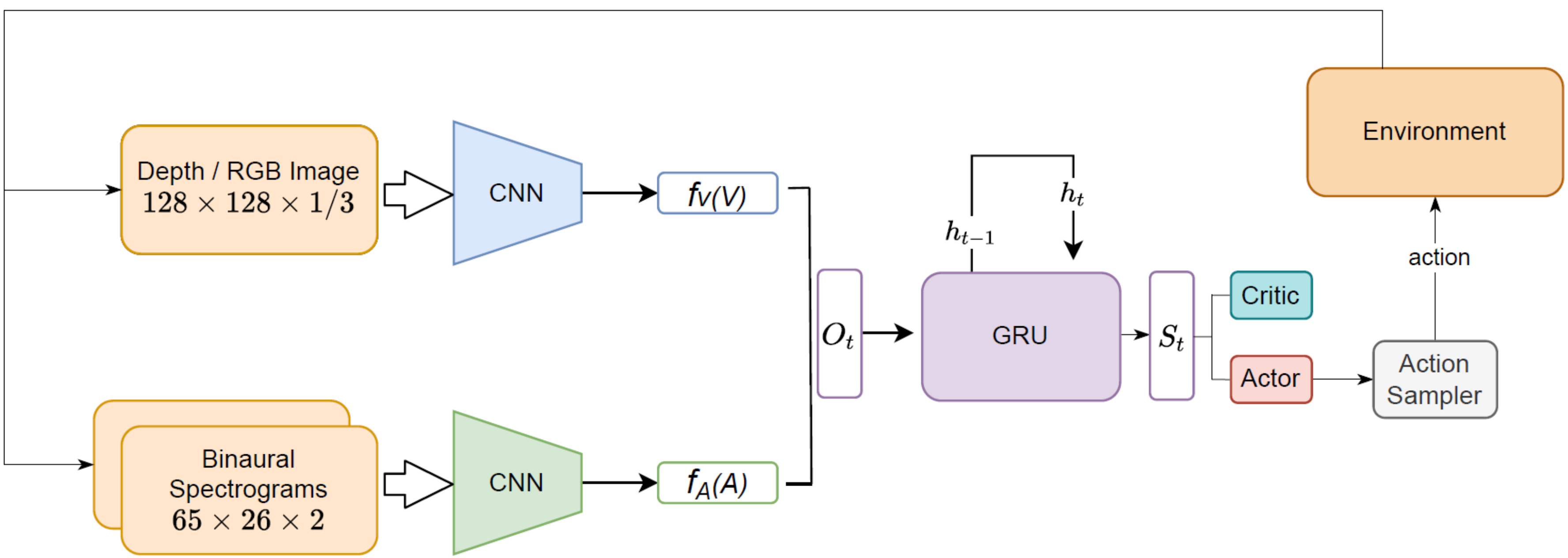


Image adapted from Chen et al., 2020

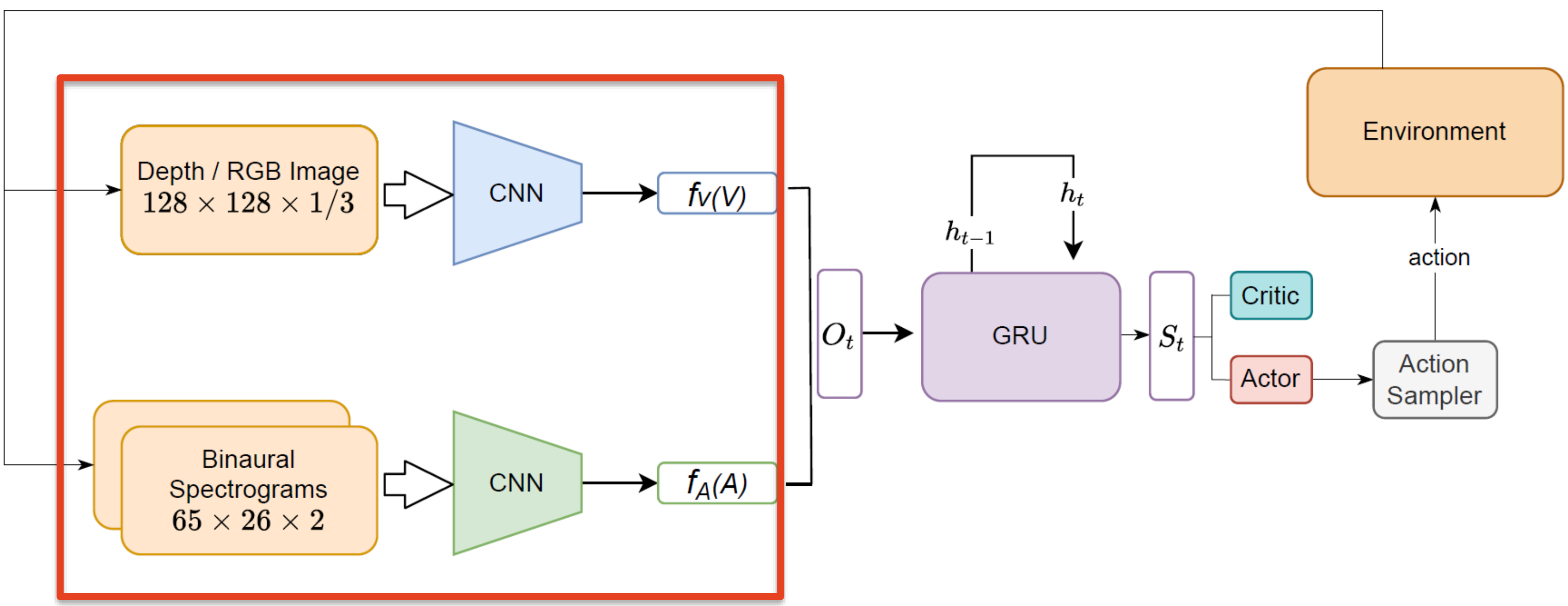
Related Work

- Audio-Visual Navigation (**AV-NAV**) Framework
  - CNN for feature extraction
  - GRU for agent memory
  - Actor-critics for reinforcement learning



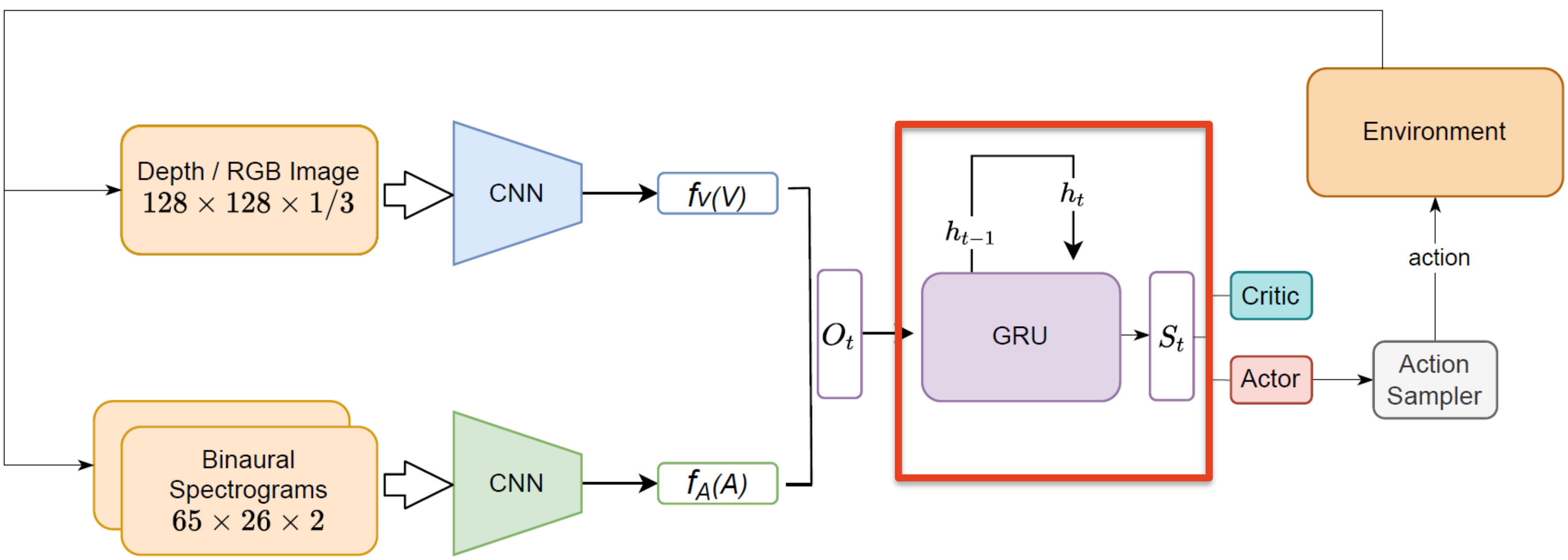
Related Work

- Audio-Visual Navigation (**AV-NAV**) Framework
  - CNN for feature extraction
  - GRU for agent memory
  - Actor-critics for reinforcement learning



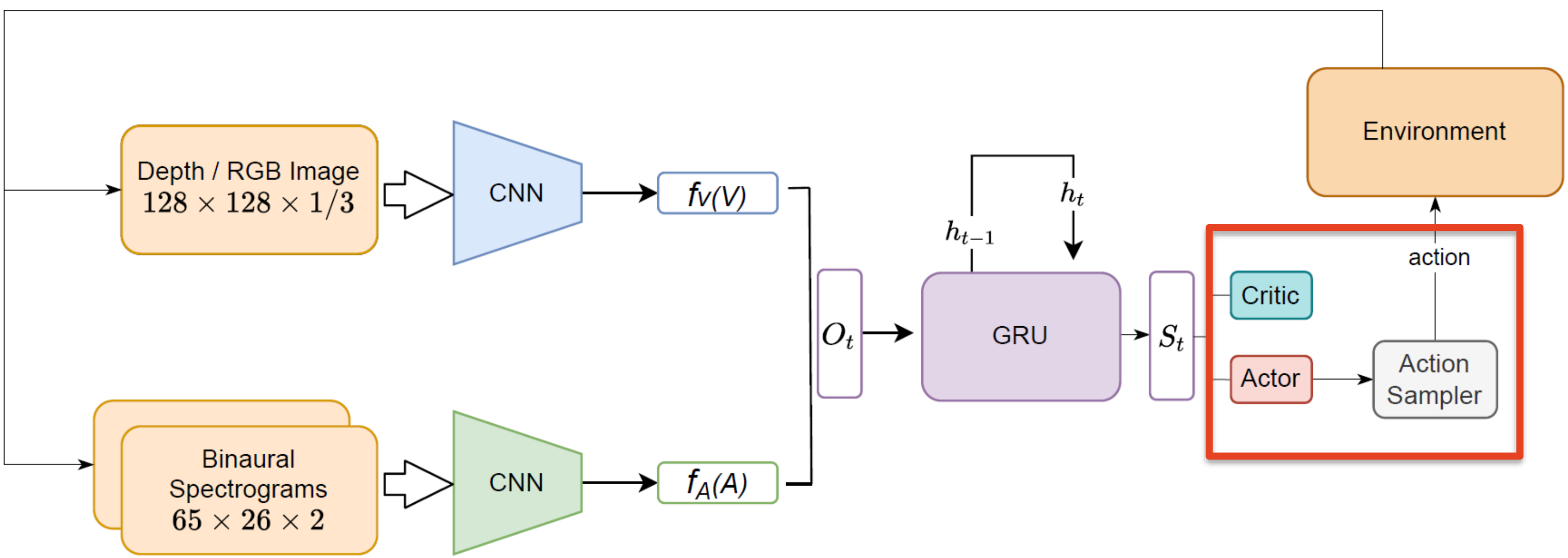
Related Work

- Audio-Visual Navigation (**AV-NAV**) Framework
  - CNN for feature extraction
  - GRU for agent memory
  - Actor-critics for reinforcement learning



Related Work

- Audio-Visual Navigation (**AV-NAV**) Framework
  - CNN for feature extraction
  - GRU for agent memory
  - Actor-critics for reinforcement learning



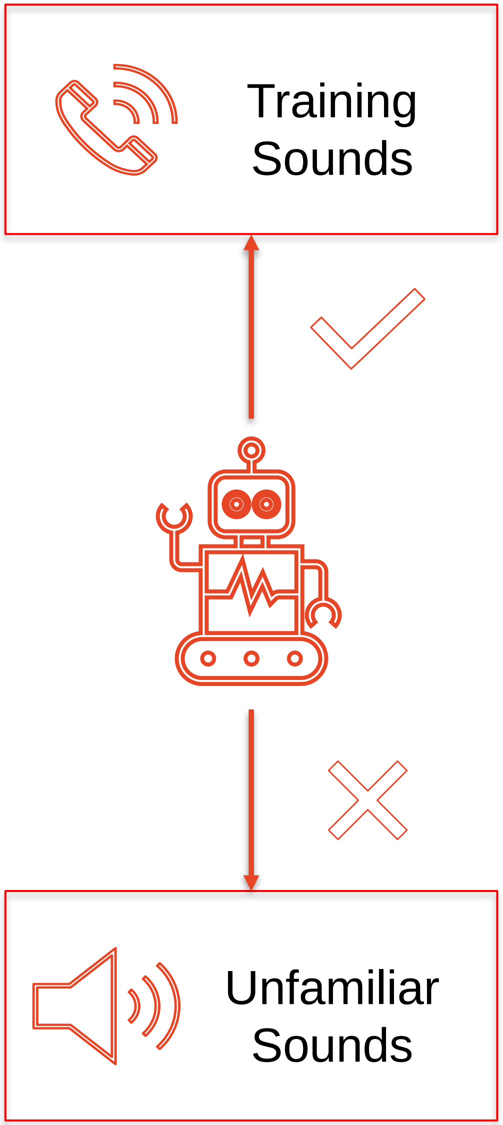
Related Work

- Occupancy Map and Dynamic Path Planner
- Acoustic Mapping
- **AV-WaN:** Waypoint Navigation
- Transformer-Based Navigation Memory (Semantic AVN)
- Distracting Sound (Adversarial AVN)

Existing Limitations

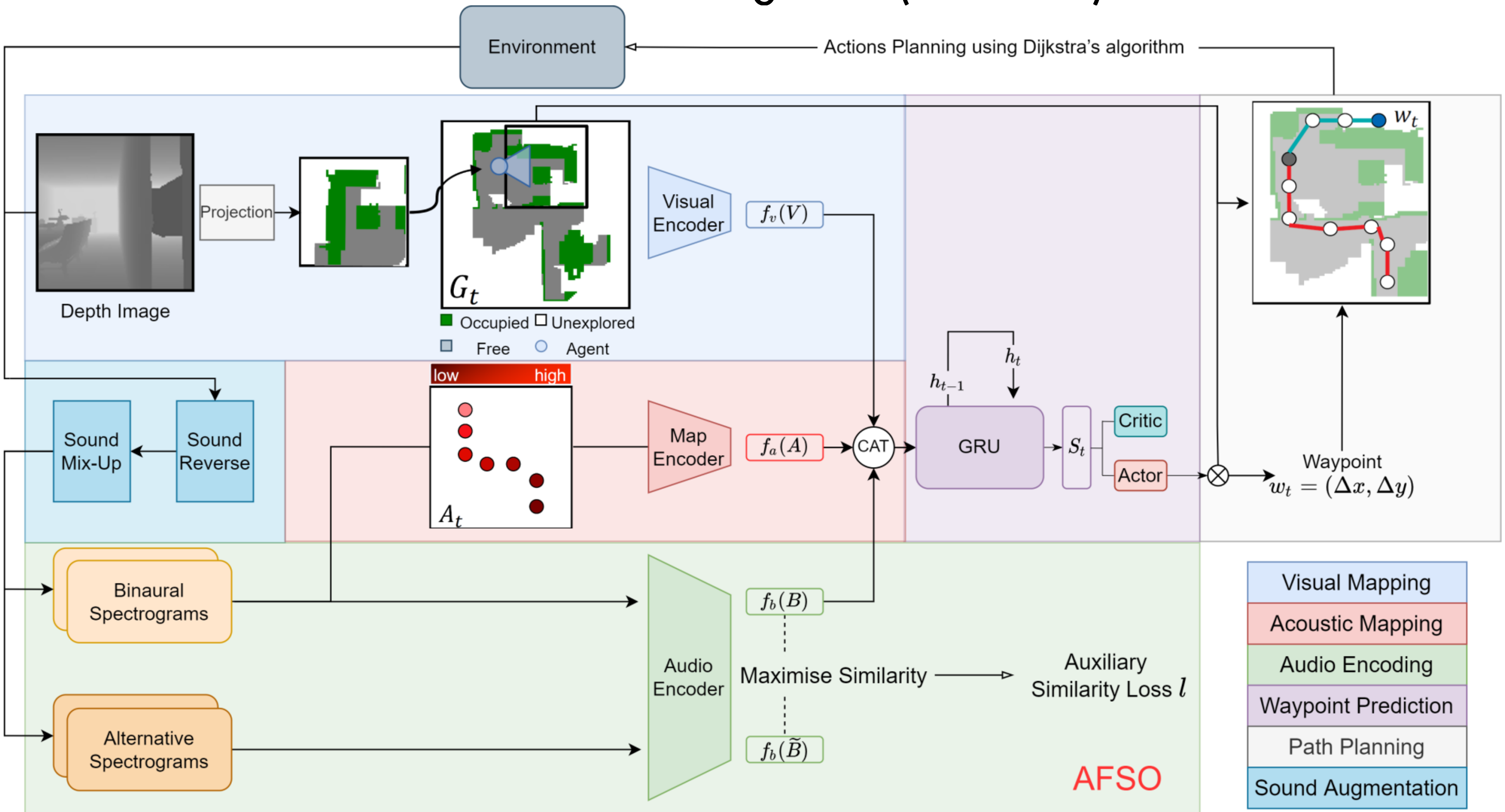
Existing frameworks performs poorly at navigating towards un-familiar audio goals.

When evaluated on unfamiliar target sounds, performance drops for a half compared to evaluated on training sounds



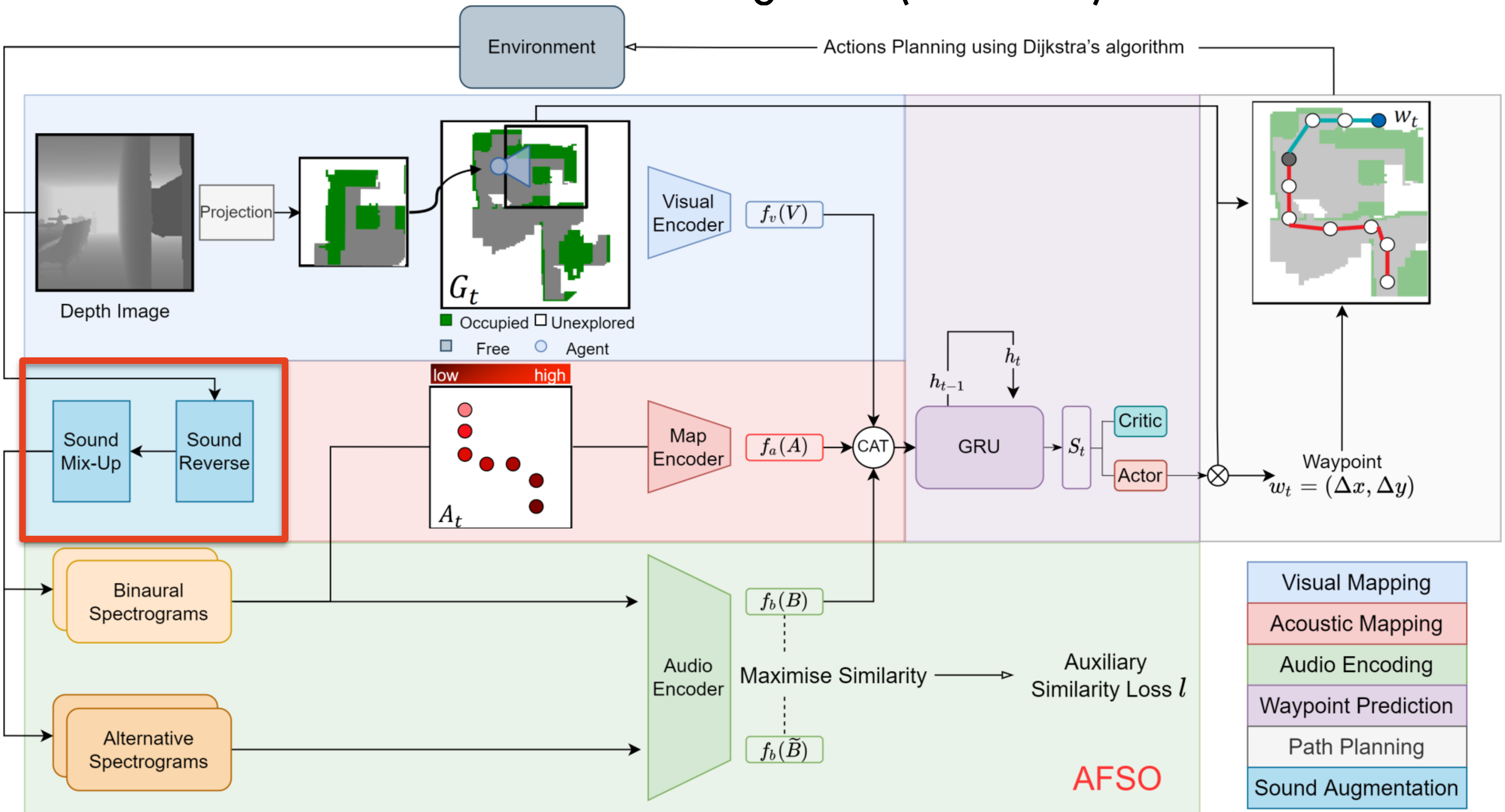
Method - AV-GeN

Generalisable Audio-Visual Navigation (**AV-GeN**) Framework



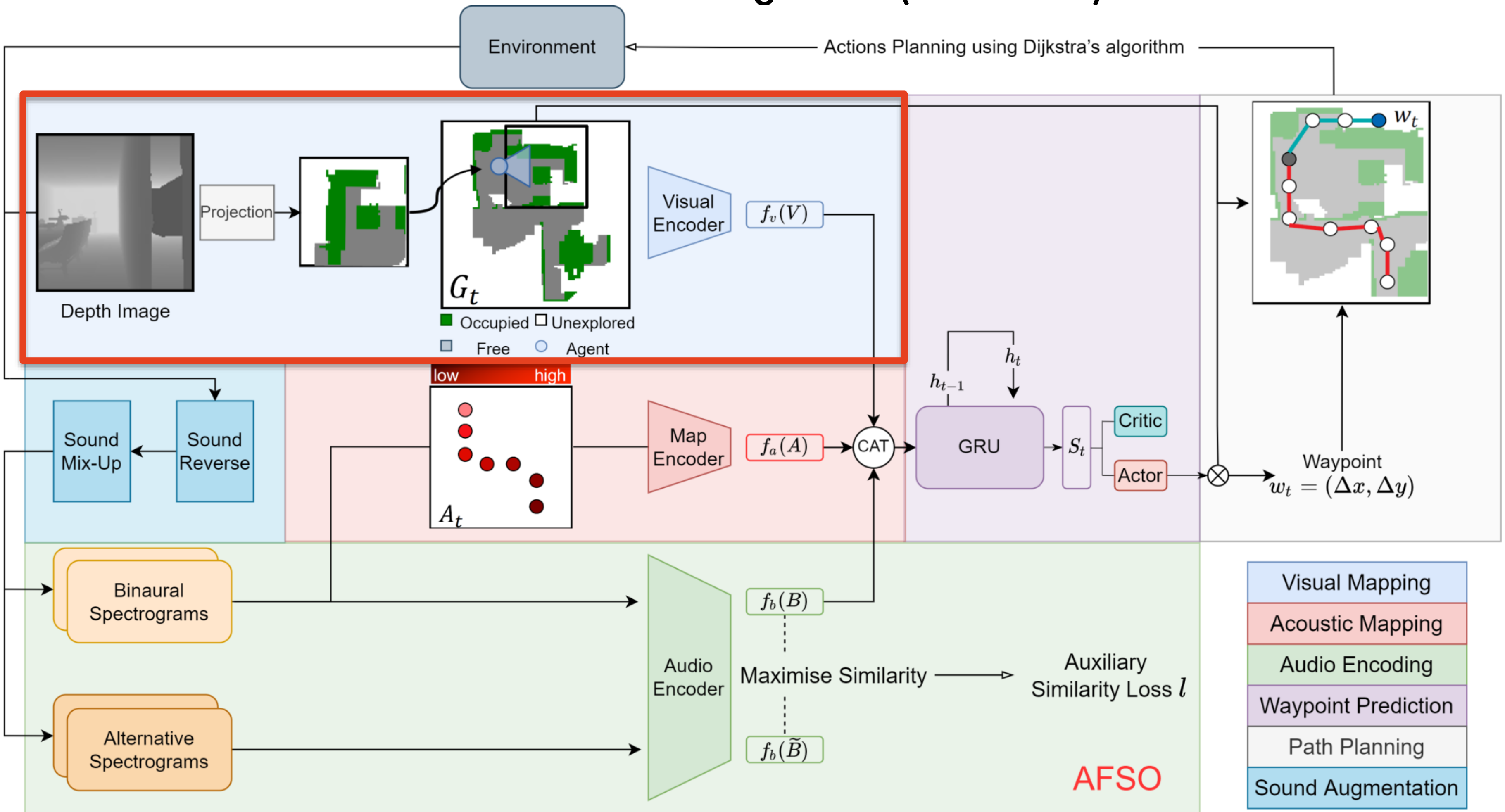
Method - AV-GeN

Generalisable Audio-Visual Navigation (**AV-GeN**) Framework



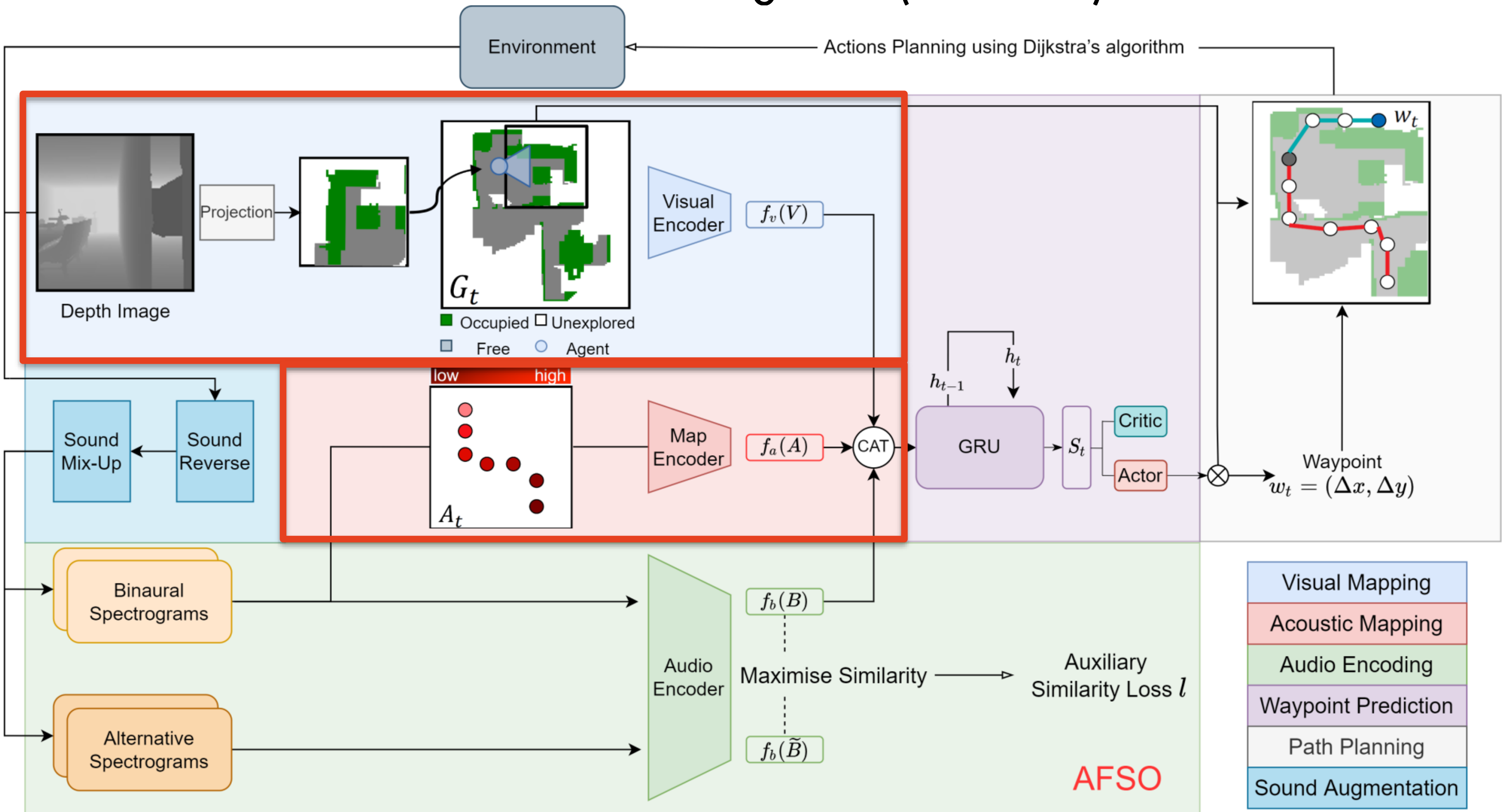
Method - AV-GeN

Generalisable Audio-Visual Navigation (**AV-GeN**) Framework



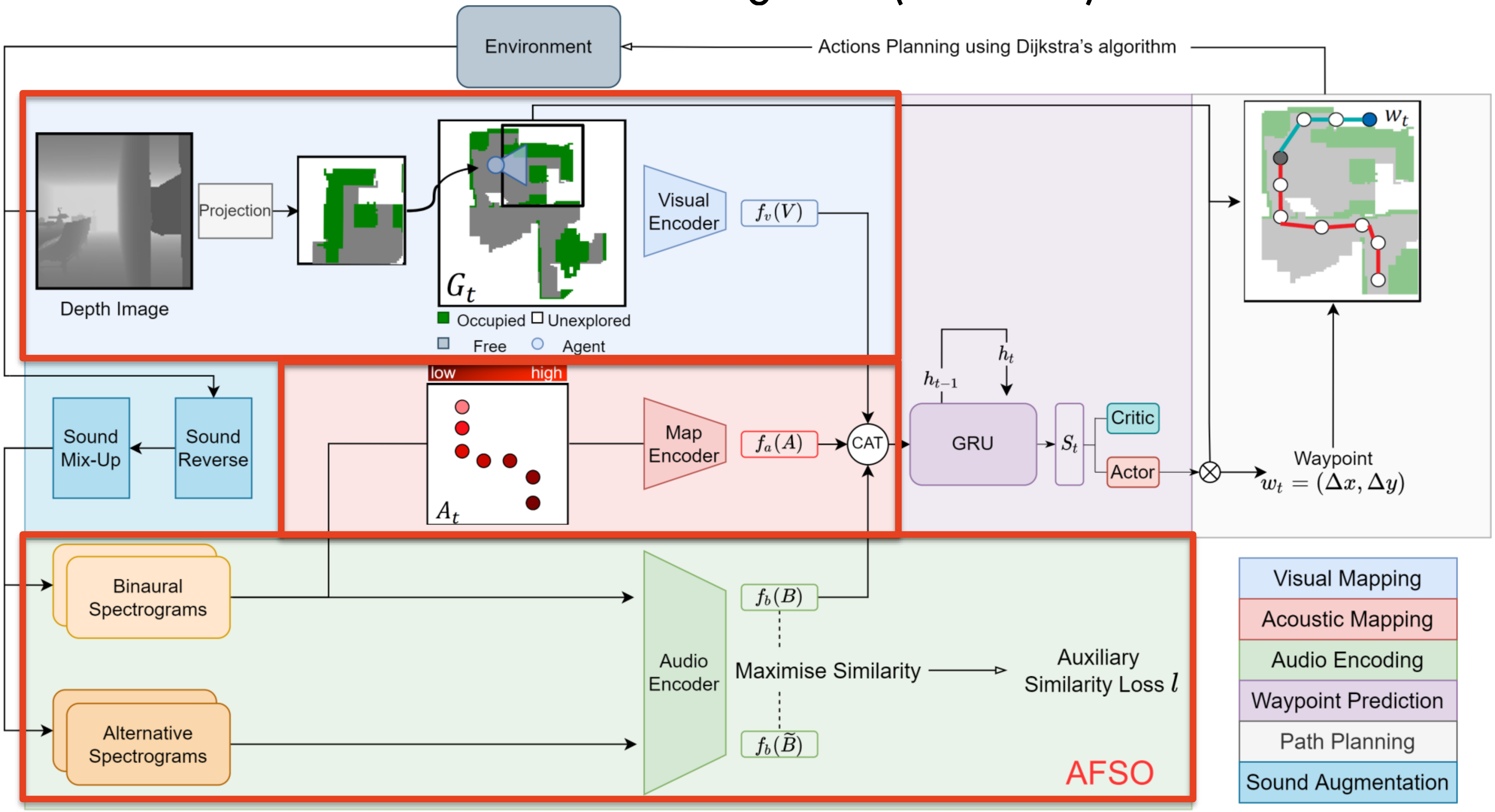
Method - AV-GeN

Generalisable Audio-Visual Navigation (**AV-GeN**) Framework



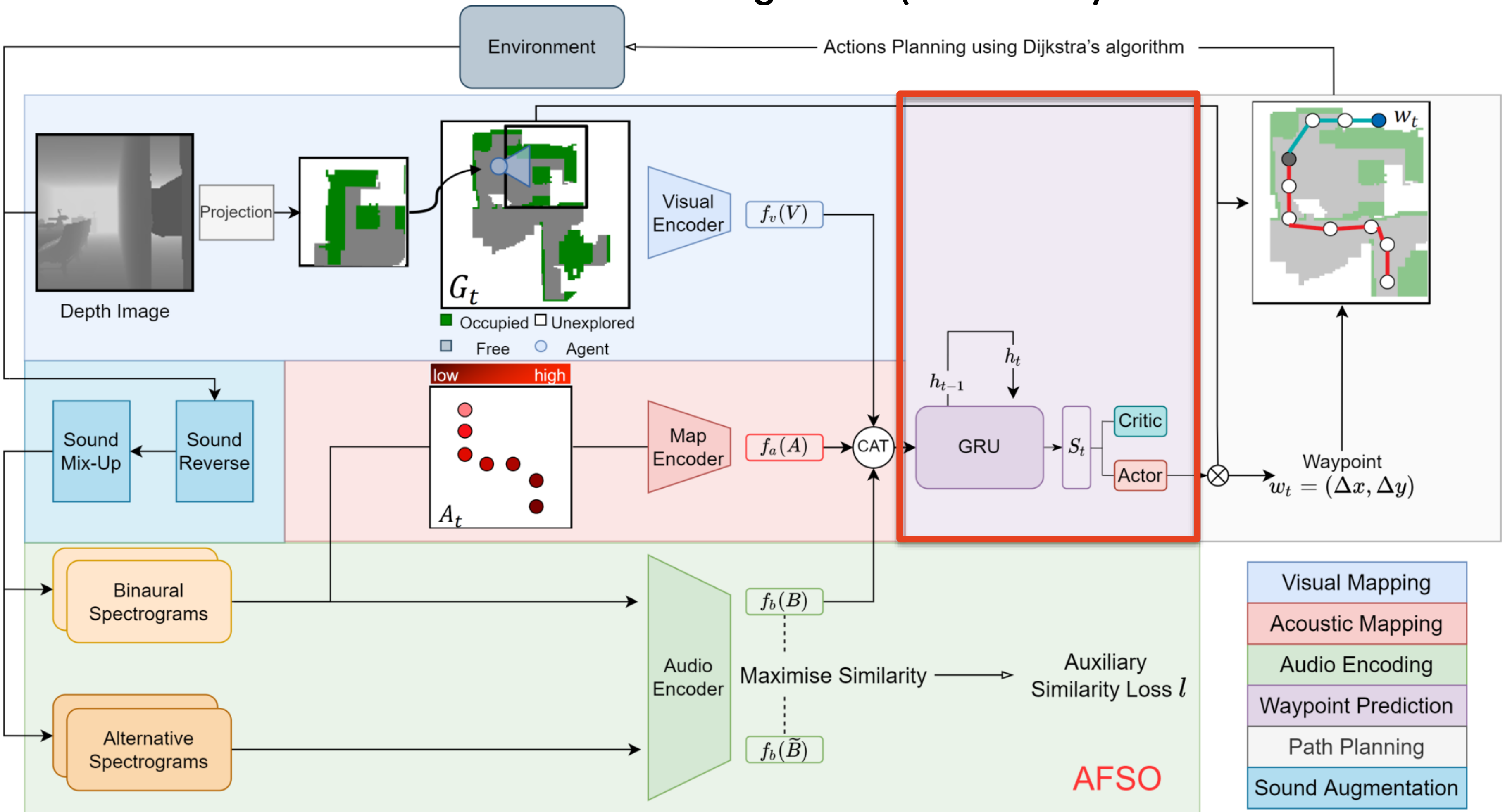
Method - AV-GeN

Generalisable Audio-Visual Navigation (**AV-GeN**) Framework



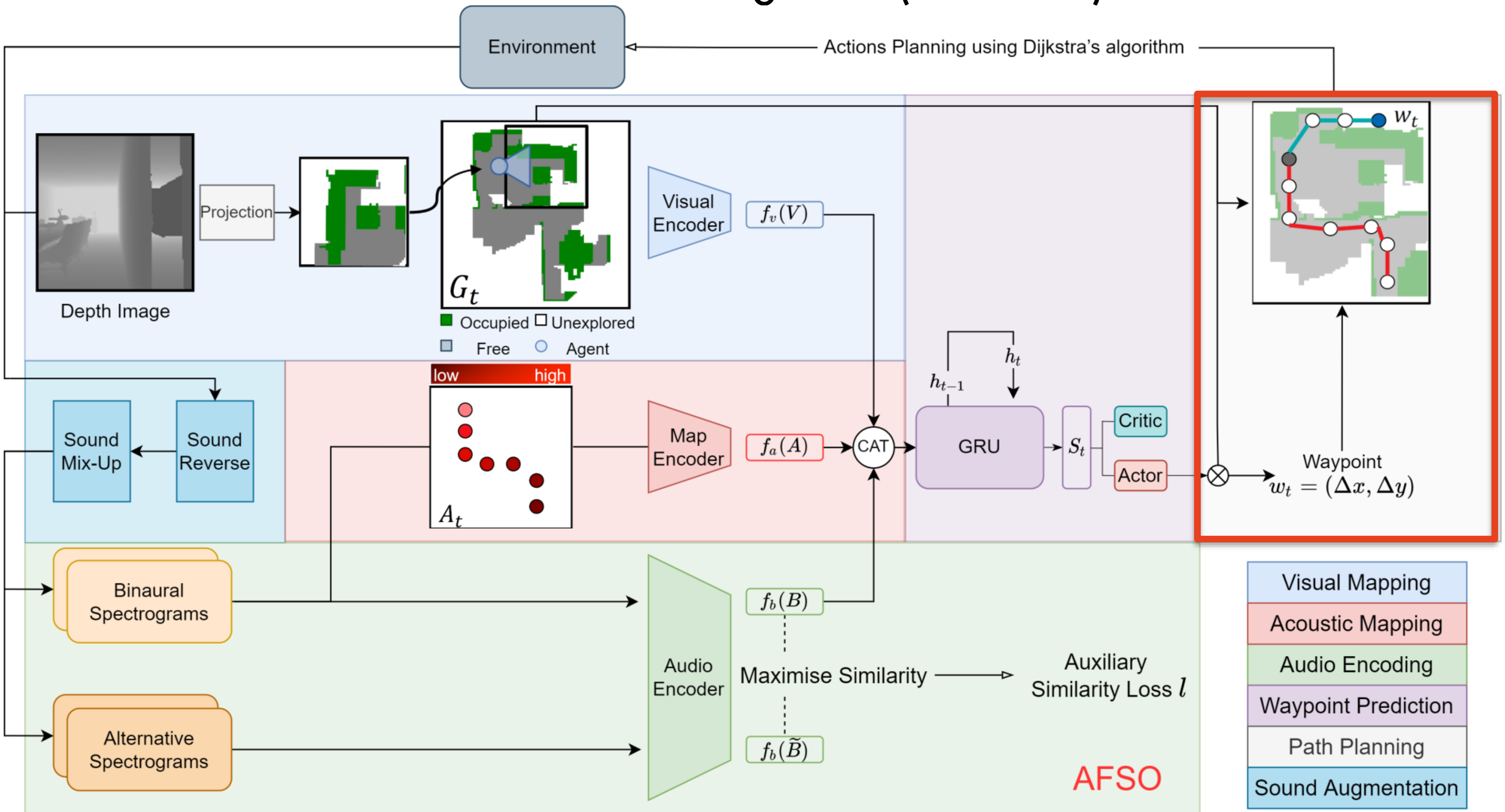
Method - AV-GeN

Generalisable Audio-Visual Navigation (**AV-GeN**) Framework



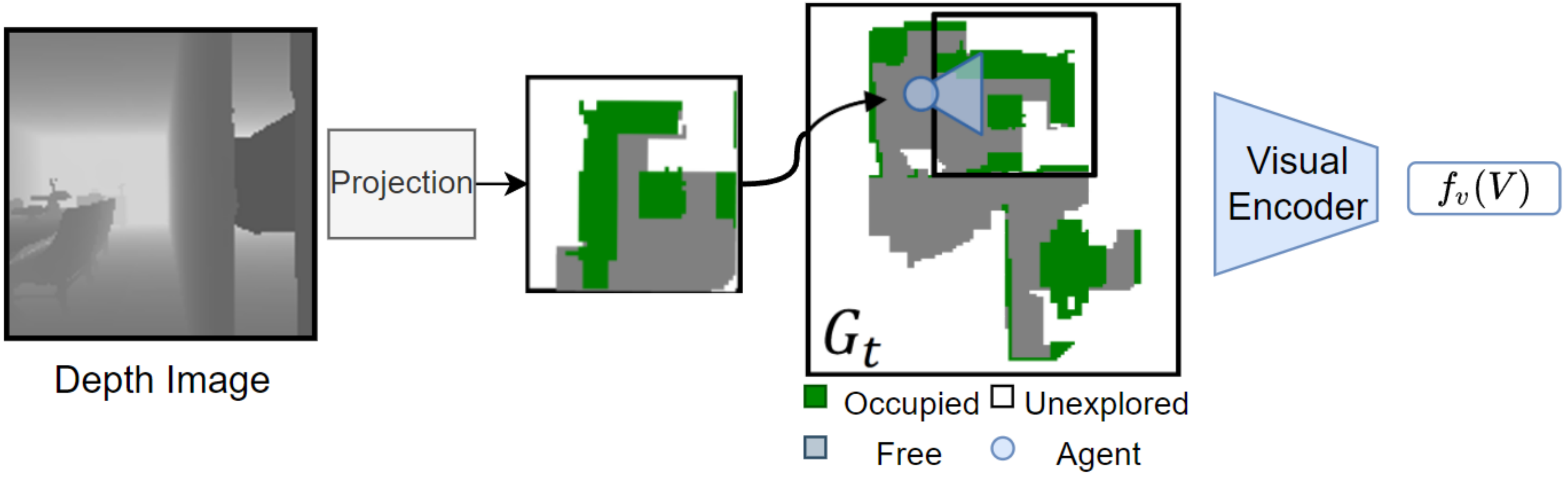
Method - AV-GeN

Generalisable Audio-Visual Navigation (**AV-GeN**) Framework



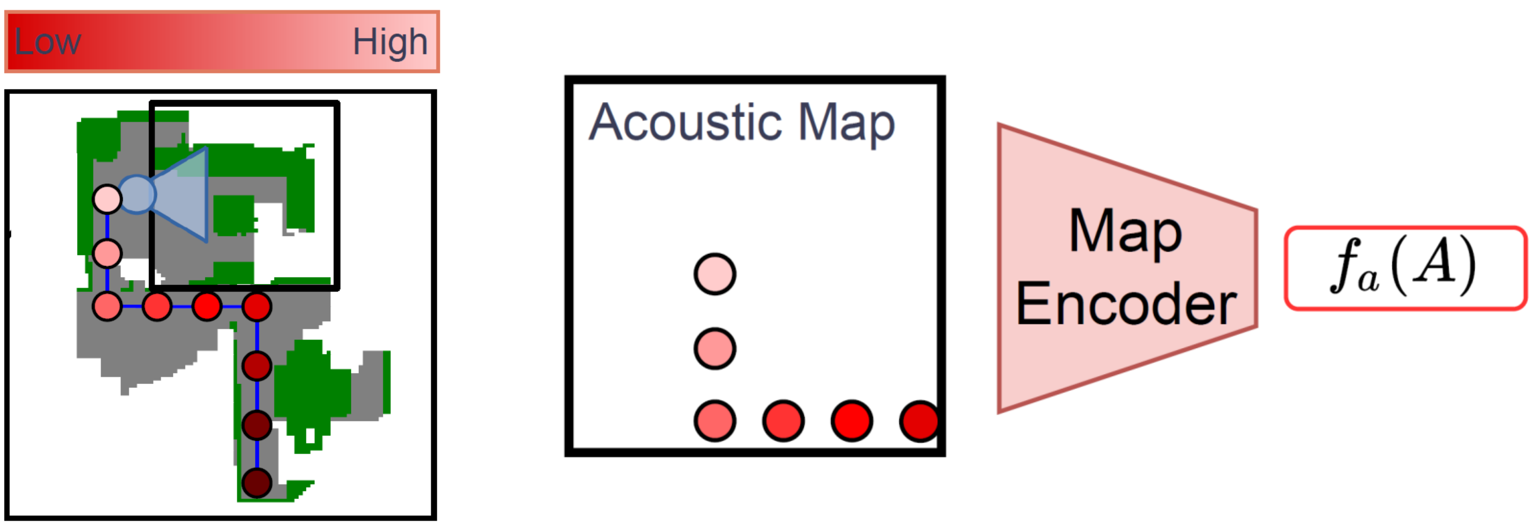
Visual Mapping

- Project depth image to a local top-down occupancy map
- Maintain a global occupancy map in an egocentric view.
- Learn geometrical mapping features with a CNN.



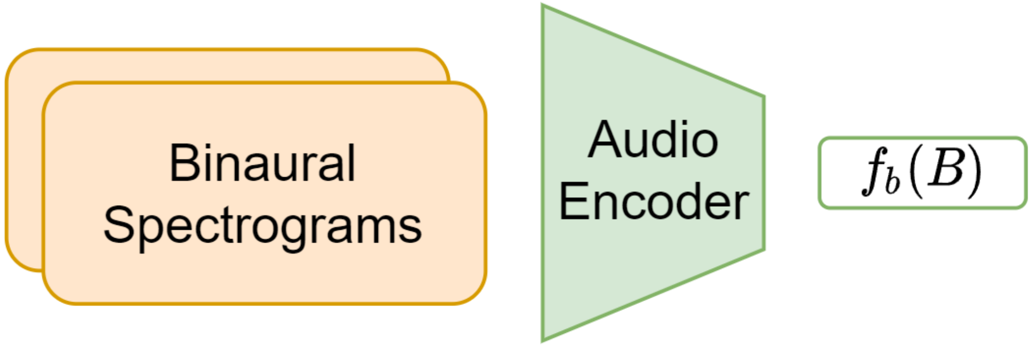
Acoustic Mapping

- Maintain a global map of the intensity values of audio signals.
- Crop a local acoustic map from the global map.
- Learn acoustic mapping features with a CNN.



AFSO-Based Audio Encoding

CNN-based audio feature extractors are prune to overfitting.



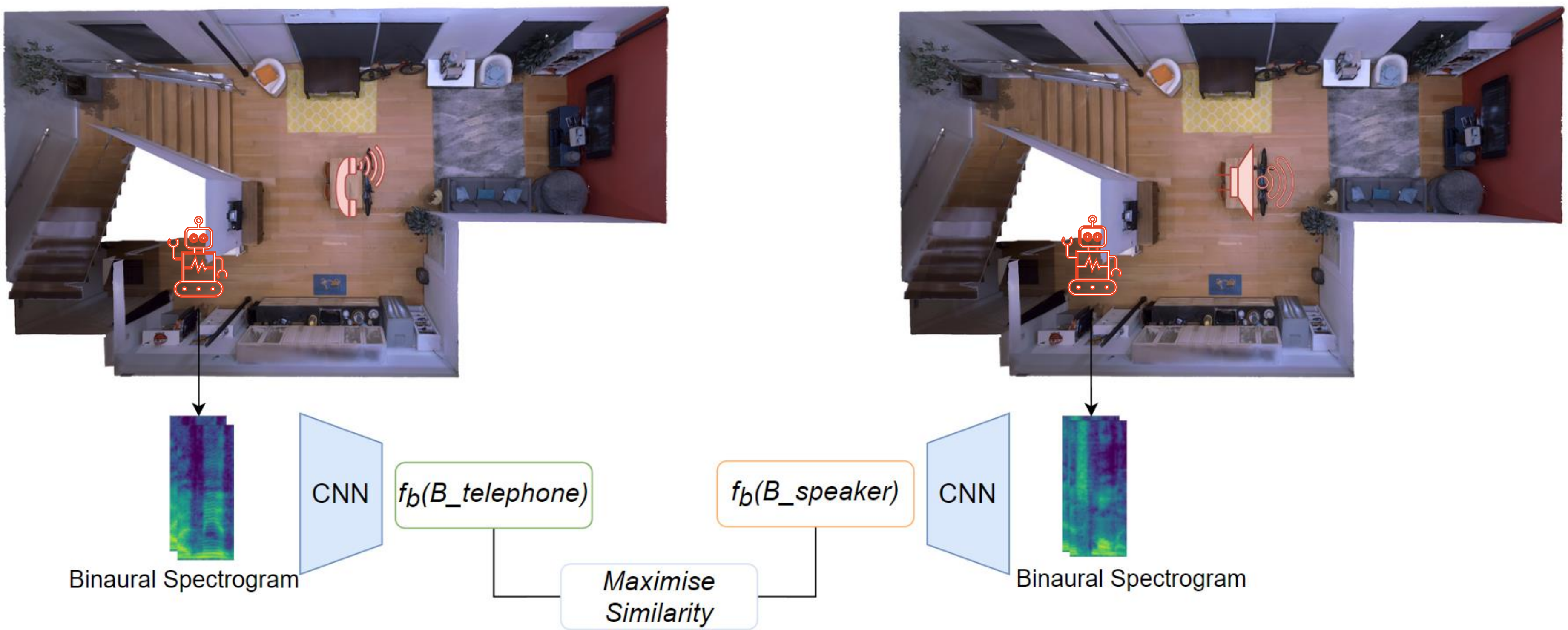
We propose Audio Feature Similarity Optimisation (**AFSO**) to regularise the audio encoder, where the sound-agnostic goal-driven latent representations can be learnt.

Intuition

The audio encoder does not need to learn the semantic class of sounds, but only need to focus on the **source-receiver spatial relationships** implied by the audio signals.

Audio Feature Similarity Optimisation

The feature similarity between audio features should be **maximised** if they imply the same audio goal position, even if they are emitted from different audio sources.

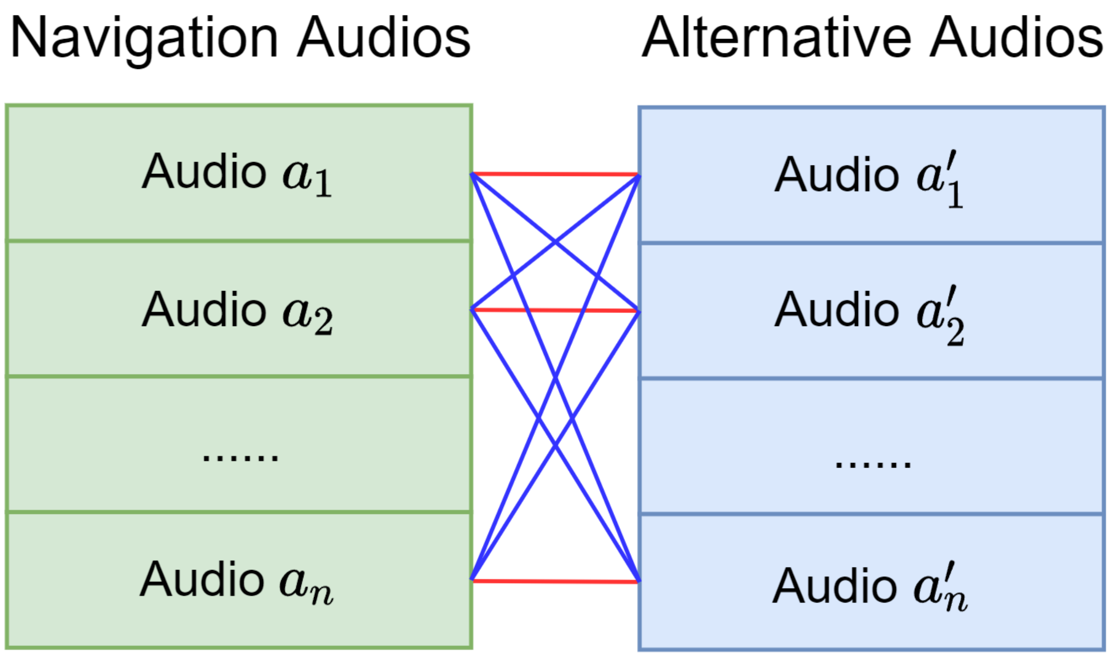


Audio Feature Similarity Optimisation

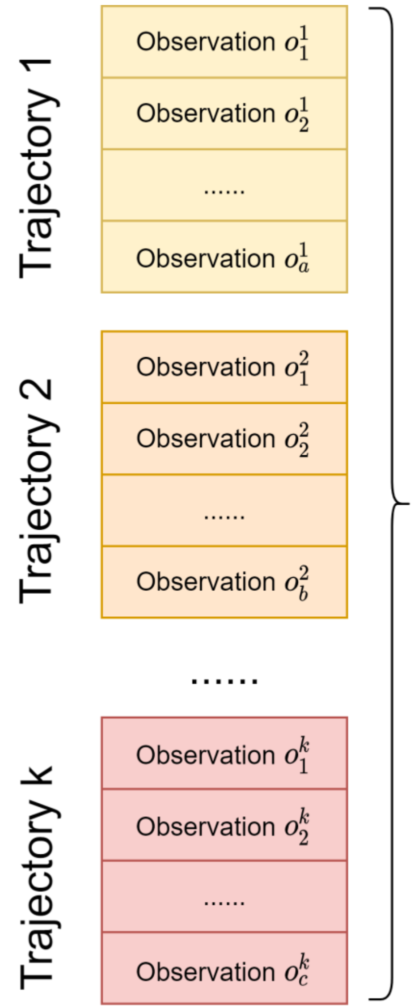
The feature similarity between audio features should be **minimised** if they imply the different audio goal position, even if they are emitted from the same audio sources.

Audio Feature Similarity Optimisation

- The pair of audio observations is considered positive only if the audios are sourced from the same scene, audio source position, and receiver position.
- We directly simulate the positive pairing elements.

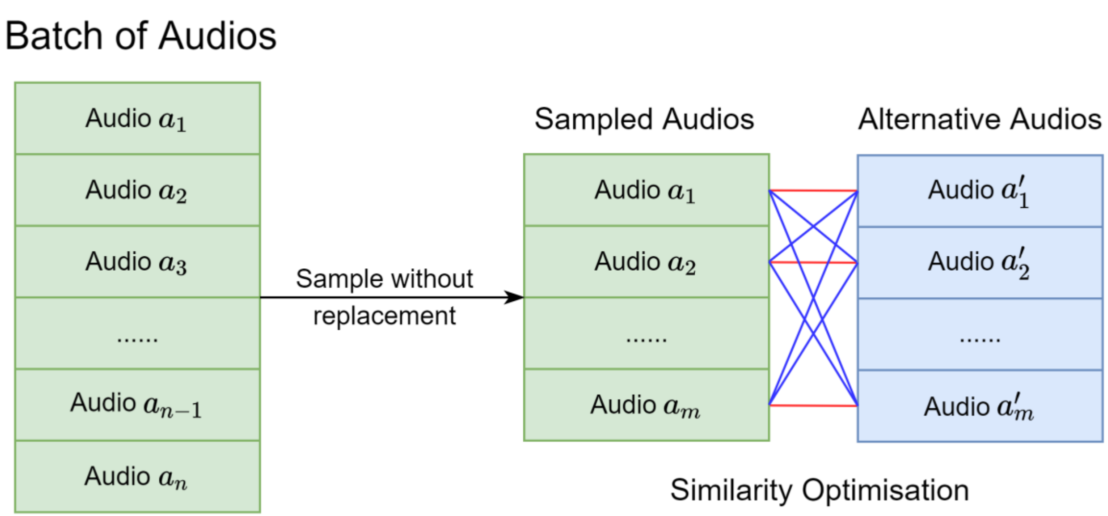


Batch Sampling



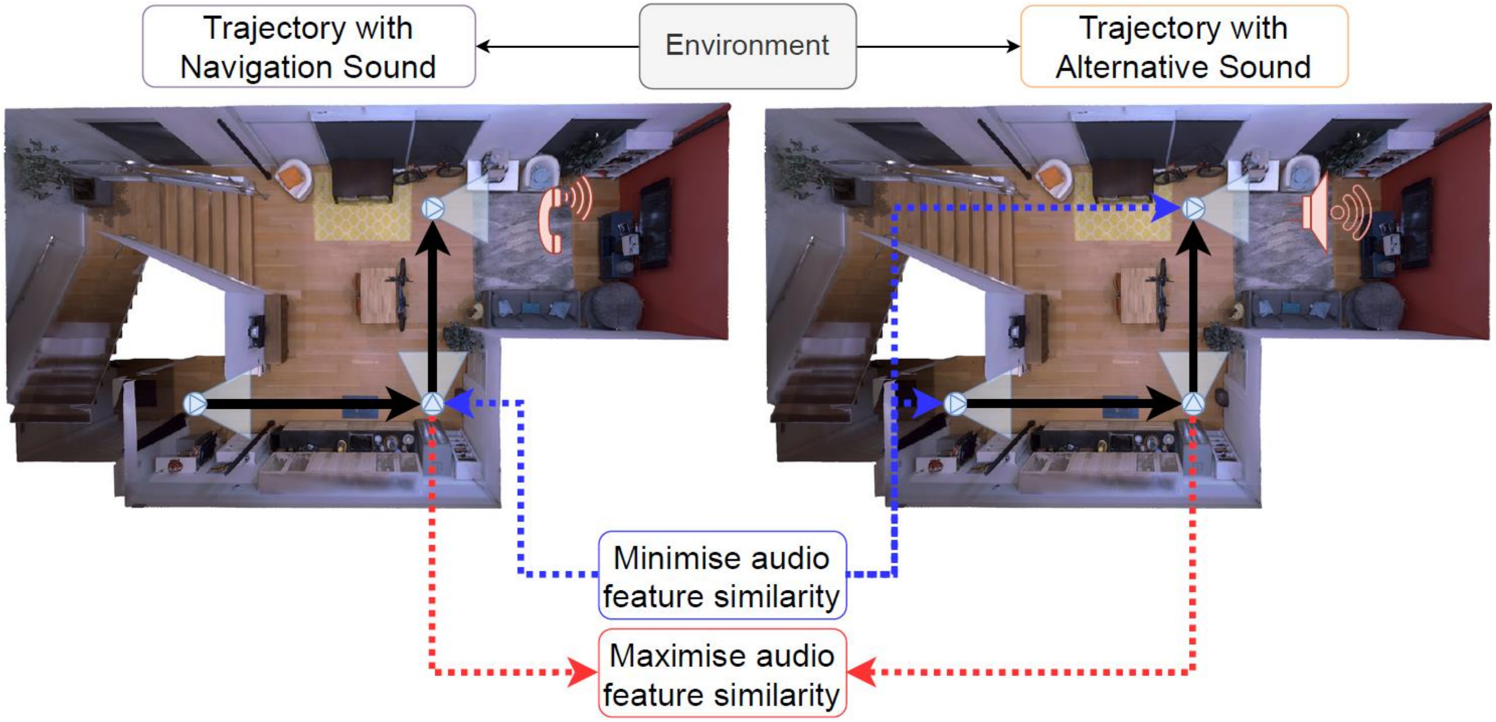
Some audio pairs with similar relative position information might be treated as negative pairs.

We randomly sample a mini-batch of audio observations to reduce the false-negative pairs in the contrastive optimisation.



Audio Feature Similarity Optimisation

- The pair of audio observations is considered positive only if the audios are sourced from the same scene, audio source position, and receiver position.
- We directly simulate the positive pairing elements.

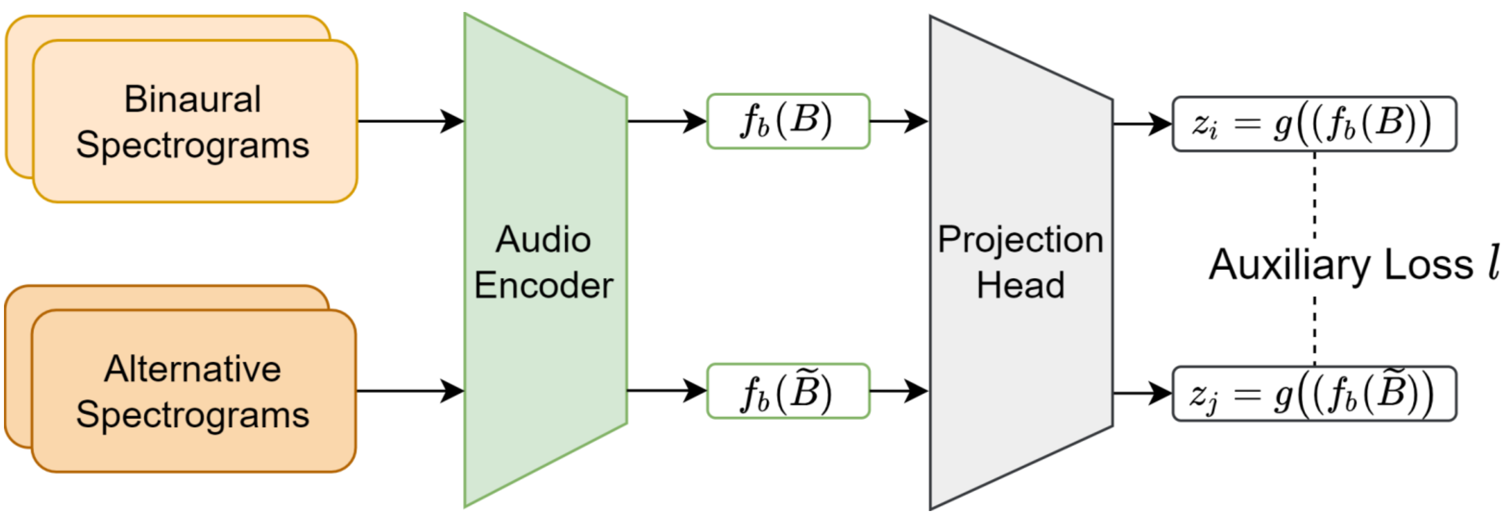


Contrastive Optimisation

For a positive pair of audio signal (i, j), the loss function is defined as:

$$l_{i,j} = -log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} 1_{k \neq i} exp(sim(z_i, z_j)/\tau)},$$

where sim denotes the cosine similarity $sim(u, v) = \frac{u^T v}{\|u\| \|v\|}$, and τ denotes a temperature parameter (InfoNCE Loss).



Sound Augmentation

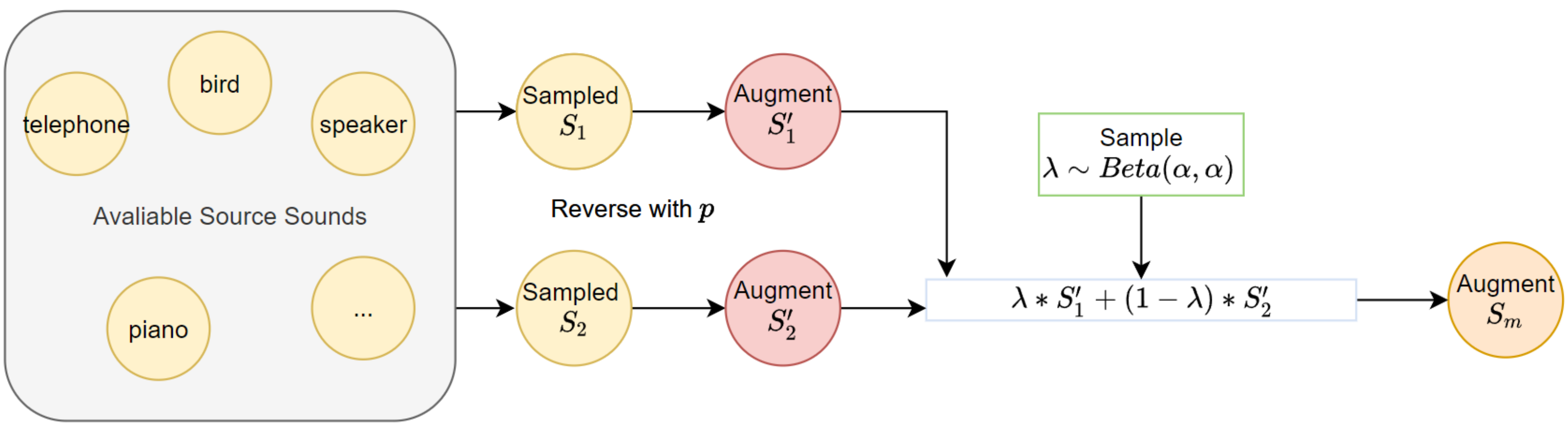
- Reverse

$$R(S[i_1, i_2, \dots, i_n]) = S[i_n, i_{n-1}, \dots, i_1]$$

- Mix-up

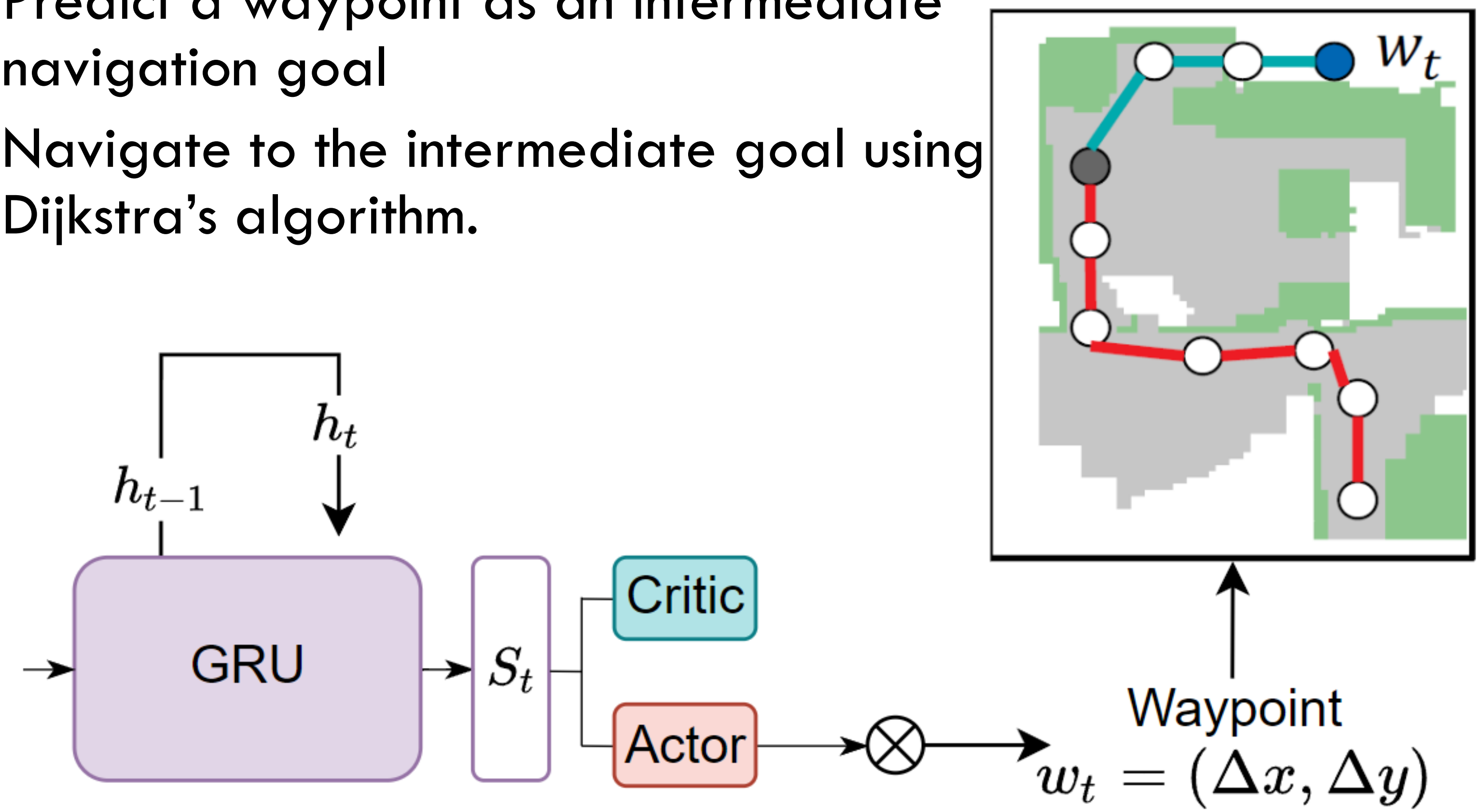
$$S_m = \lambda S_1 + (1 - \lambda) S_2$$

$$\lambda \sim Beta(\alpha, \alpha)$$



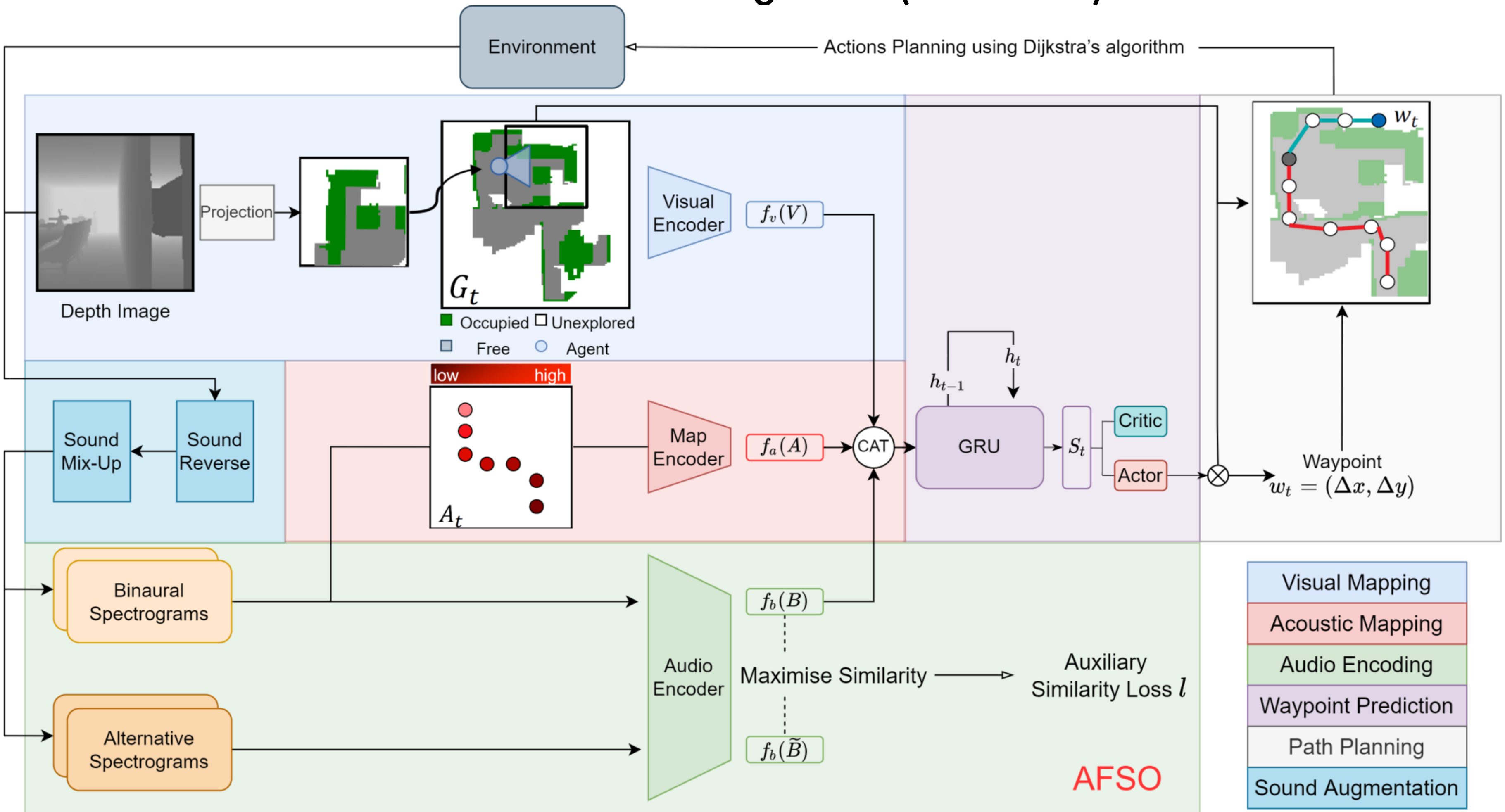
Waypoint Prediction and Path Planning

- GRU for navigation memory and actor-critic RL algorithm to optimise the networks.
- Predict a waypoint as an intermediate navigation goal
- Navigate to the intermediate goal using Dijkstra's algorithm.



Method - AV-GeN

Generalisable Audio-Visual Navigation (**AV-GeN**) Framework



Experiments

- Matterport3D dataset contains 85 real-world scans with an average floor space of $517m^2$

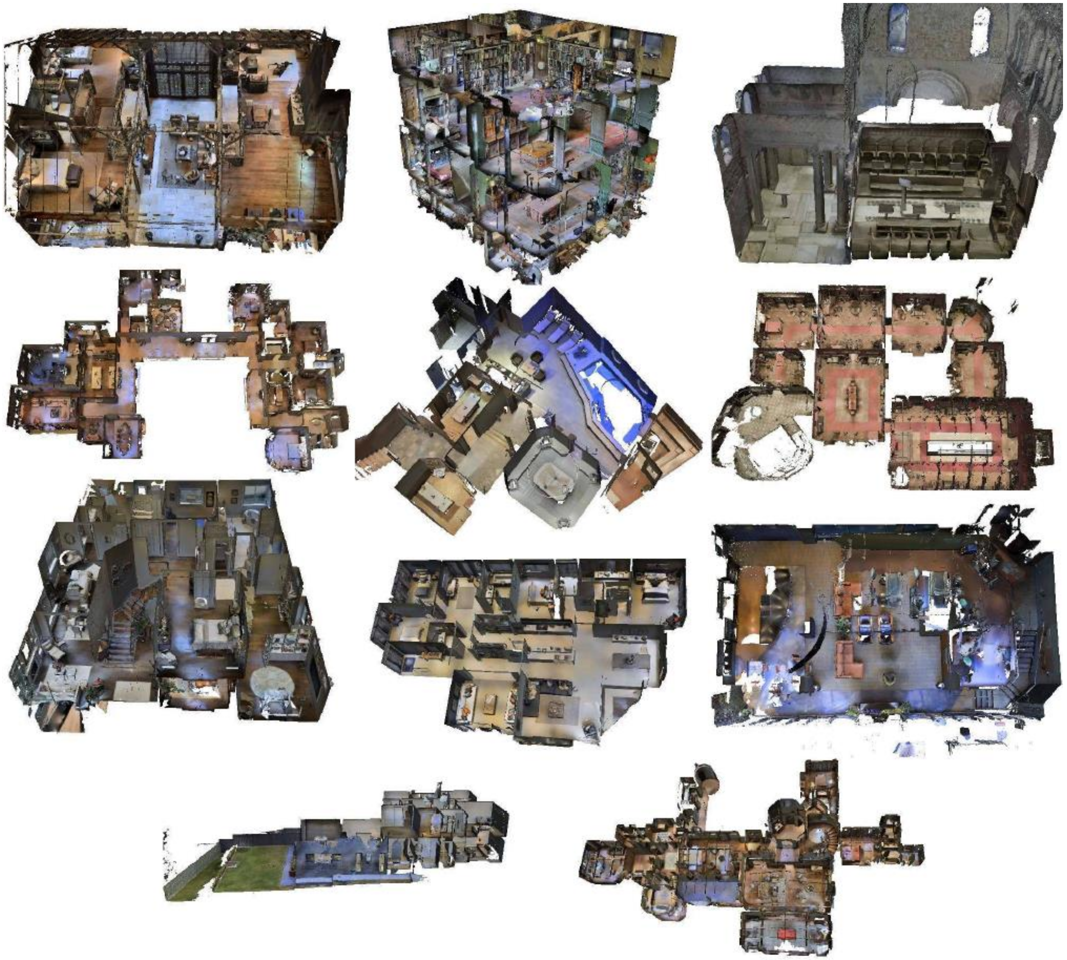


Image adapted from Chang et al., 2017

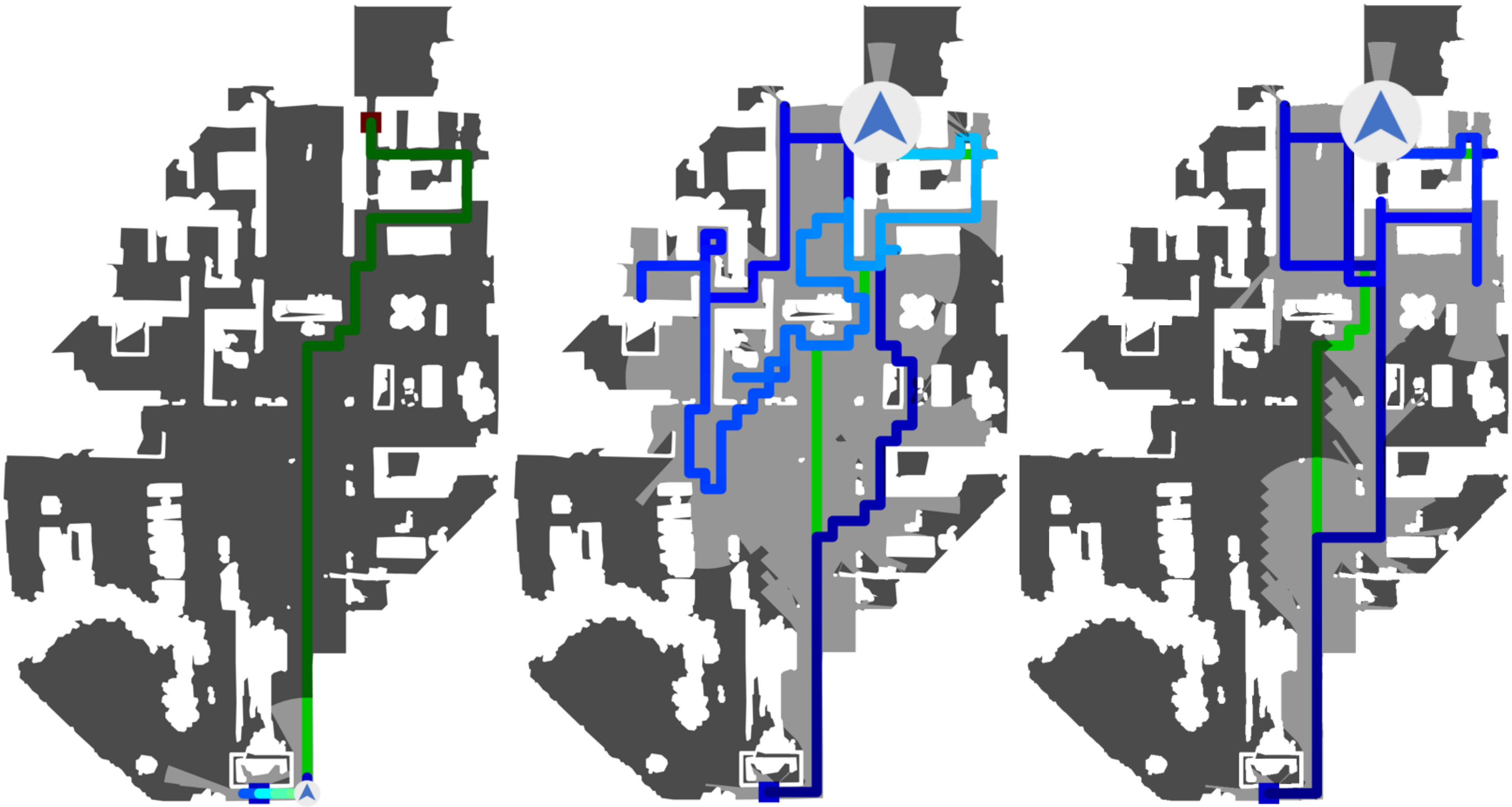
- Train/val/test split
  - 59/10/11 scenes
  - 73/11/18 sounds

Quantitative Results

- Success Rate (SR)
- Success Weighted by Number of Actions (SNA)
- Success Weighted by Path Length (SPL)

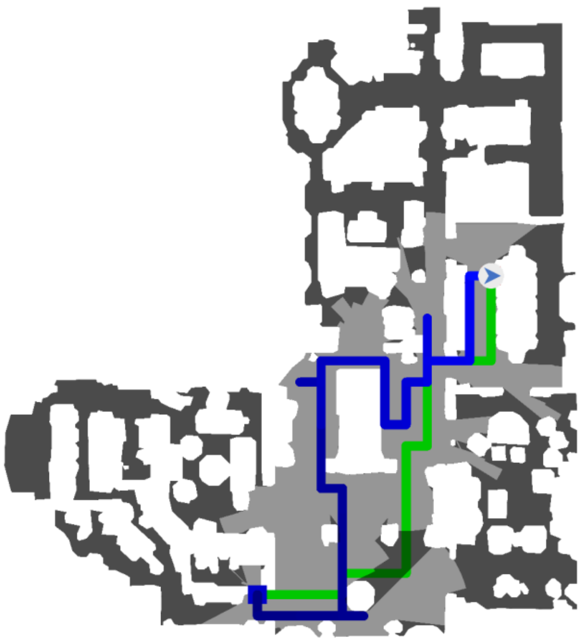
	SPL%↑	SR%↑	SNA%↑
AV-NAV	26.3	43.6	11.8
AV-WaN	36.2	57.4	27.4
AV-GeN (Ours)	**48.4**	**73.9**	**37**

Visualisations

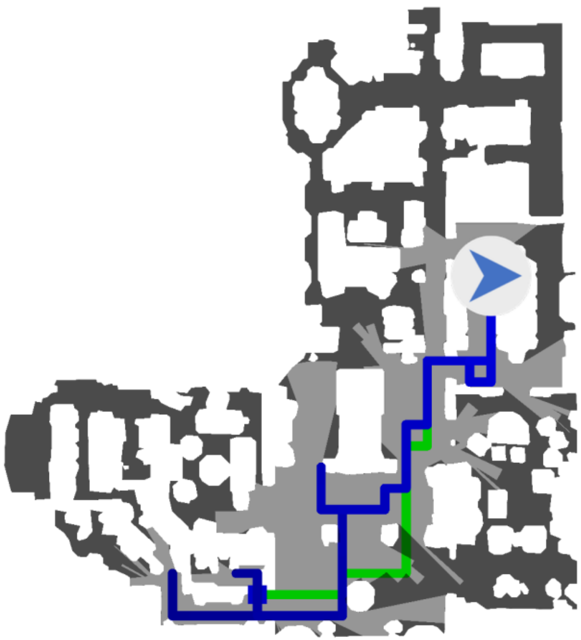


AV-NAV AV-WaN AV-GeN

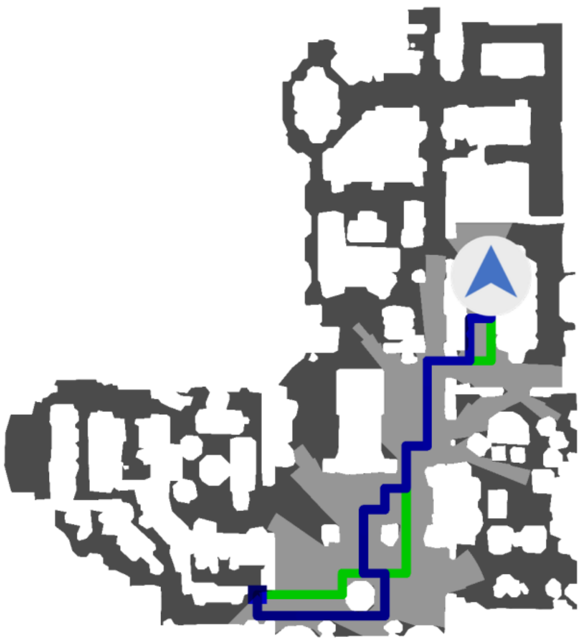
Visualisations



AV-NAV

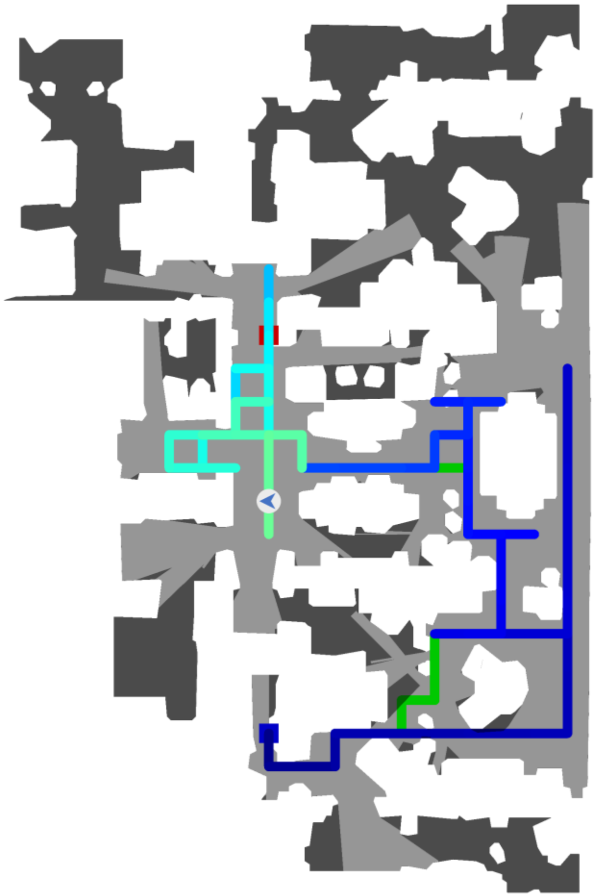


AV-WaN

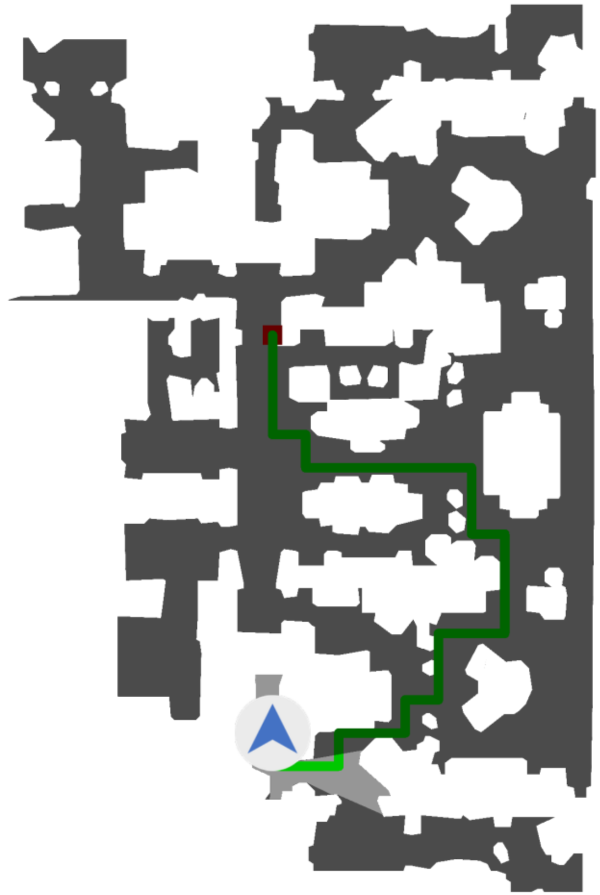


AV-GeN

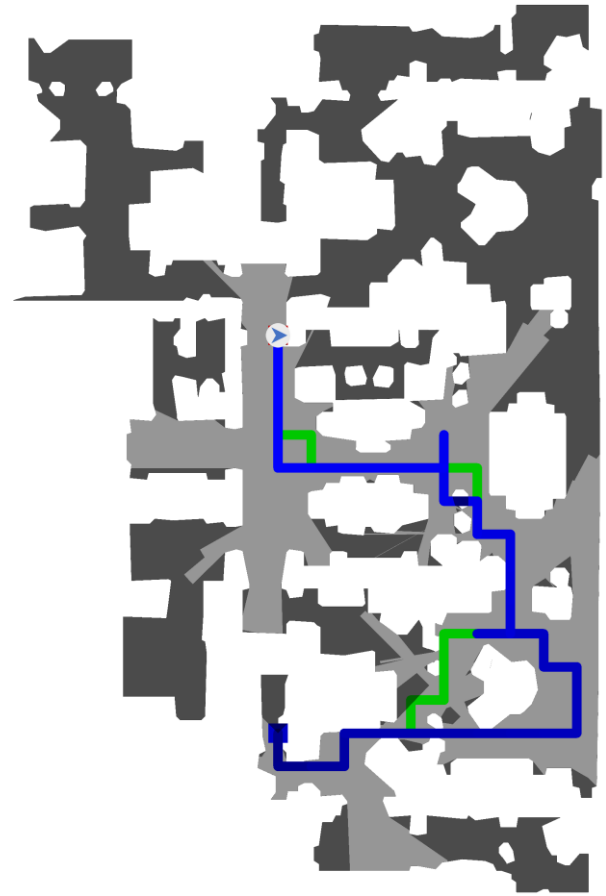
Visualisations



AV-NAV



AV-WaN



AV-GeN

Ablations

We study the importance of each designed module in the two novel methods proposed, AFSO and sound augmentation.

Method	AFSO		Augmentation		Performance		
	Sampling	Projection	Mix-up	Reverse	SPL%↑	SR%↑	SNA%↑
AV-GeN	√	×	√	√	**48.4**	**73.9**	**37.0**
w/o Aug	√	√	-	-	43.3	66.4	33.9
w/o AFSO	-	-	√	√	39.9	68	31.0
w/o both	-	-	-	-	36.7	56.4	28.1
Ablations on AFSO	√	×	-	-	41.2	67.8	32.4
	×	√	-	-	41.5	65.6	31.9
	×	×	-	-	40.8	62.4	32.7
Ablations on augmentation	-	-	×	√	37.0	66.2	28.3
	-	-	√	×	37.7	62.6	29.7

Publications

- Top-1 SR and Top-3 SPL in the SoundSpaces Challenge

SoundSpaces

- Accepted to CVPR 2022 Embodied AI Workshop

Discussion and Conclusion

Contributions

– Propose Audio Feature Similarity Optimisation (AFSO) method

– Propose Source Sound Augmentation method

– Develop the AV-GeN framework

Advantages

– Improve generalisation

– Flexible adaption

– Cheap computational cost

Limitations

– Result fluctuations

– Generalise differently to distinct sounds

Future Work

Improve the AV-GeN framework

- Validate the framework with different AVN variants
- False-negative pairs removal
- Parameter-free similarity loss estimation
- …

Towards more generalisable frameworks

- Learnable acoustic mapping
- Few-shot learning
- …

Many other fun stuffs with audio-visual environment!

Q&A

References

1. 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.

2. 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.

3. 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.

4. Jisu Hwang and Incheol Kim. Joint multimodal embedding and backtracking search in vision-and-language navigation. *Sensors*, 21:1012, 02 2021. doi: 10.3390/s21031012.

5. 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. *Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9338–9346, 2019.

6. Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*,2018.