Visual Learning of Sounds in Spaces

Changan Chen
changan.io
UT Austin
01/18/2023

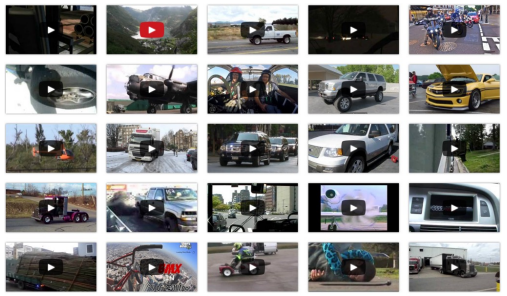
TEXAS
The University of Texas at Austin

Human perception is a multisensory experience

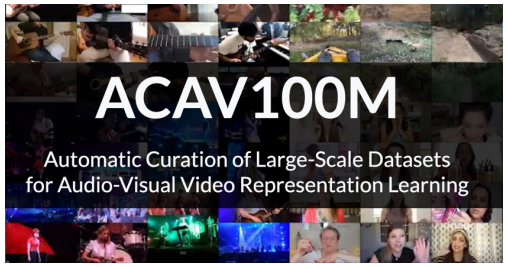
We often use *vision*, *audio*, *touch*, *smell* to sense the world

Perceptual experience based on objects

Audio-visual classification

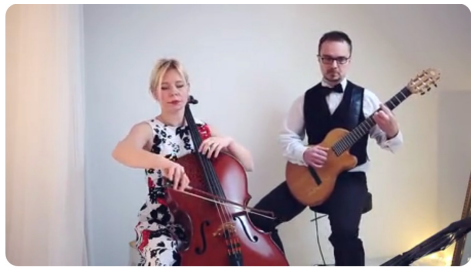


Audio Set, 2017



ACAV100M, 2021

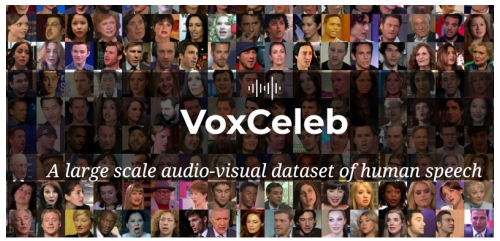
Audio-visual separation



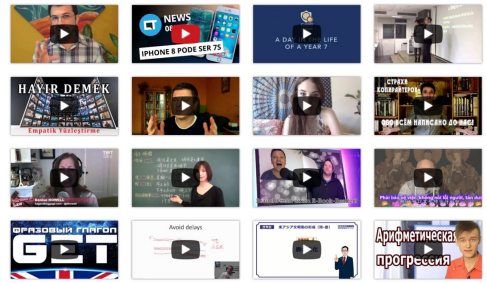
MUSIC, 2018

MUSIC-Synthetic, 2020

Speech separation

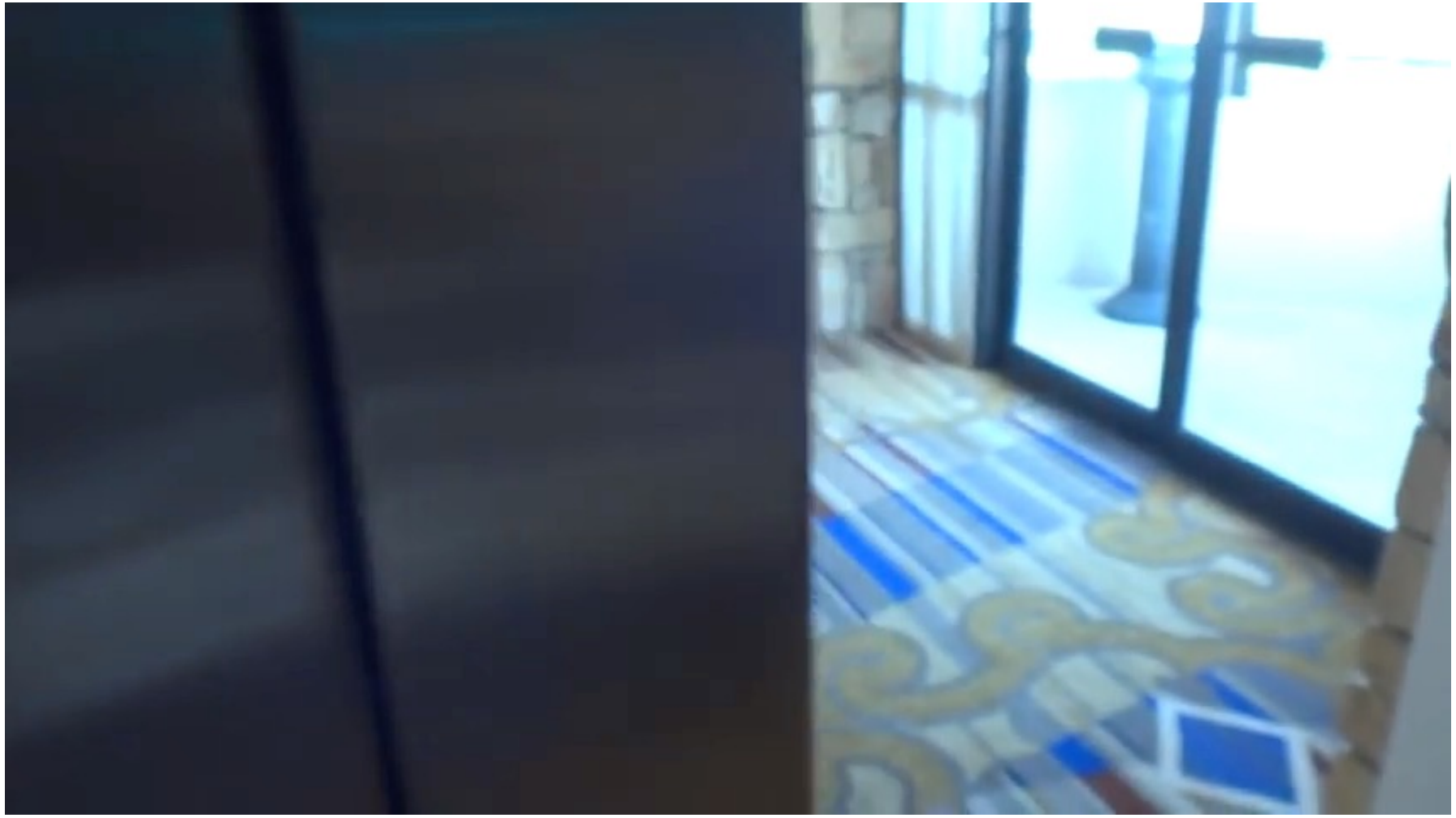


VoxCeleb, 2019



AVSpeech, 2018

Perceptual experience embodied in spaces



Why visual learning of sounds in spaces - Robot learning

Home assistance robot

Rescue robot

Robots that can navigate and localize sounding objects by reasoning the spatial, semantic, acoustic information in the audio and visual observation

Why visual learning of sounds in spaces - Augmented reality / virtual reality

Immersive experience

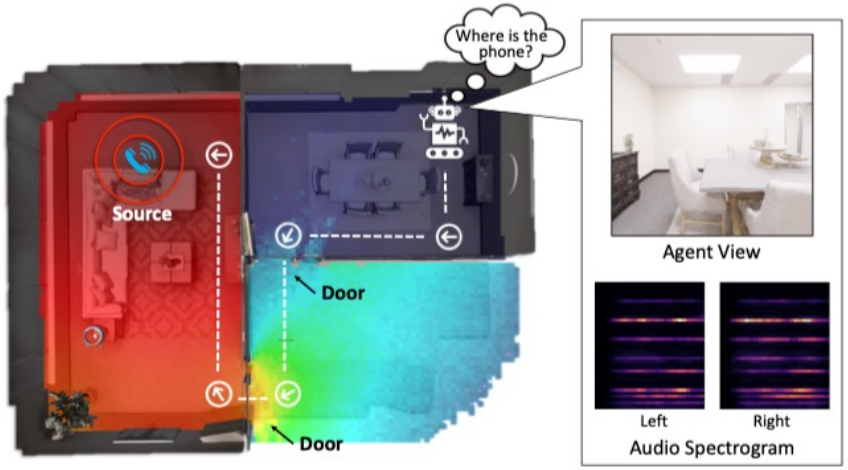
Enhanced hearing

AR/VR systems that create immersive experience for users as well as augment the hearing ability of the device wearer

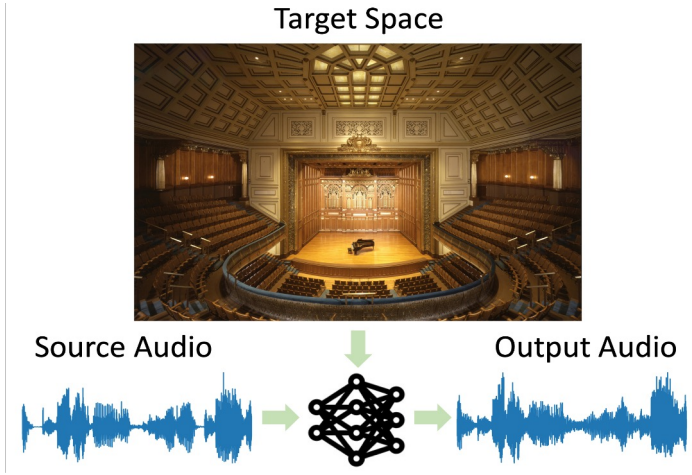
Visual learning of sounds in spaces

My research: to learn how sounds are situated, produced and transformed physically in spaces based on visual inputs

Empower robots to hear in spaces

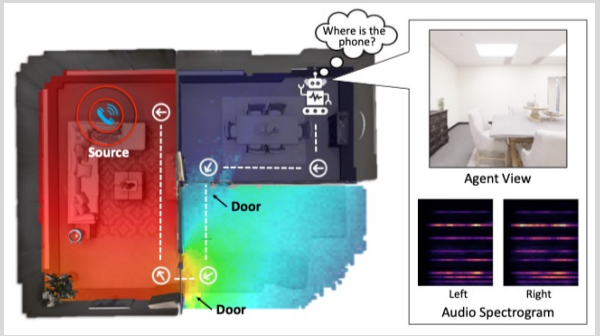


Visual understanding of room acoustics



Empower robots to hear in spaces

Robot Learning



SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020 (Spotlight)
Changan Chen, Unna Jain, Carl Schissler, Sebastia V. Amengual Gari, Ziad Al-Halah, Vamsi K. Ithapu, Philip Robinson, Kristen Grauman

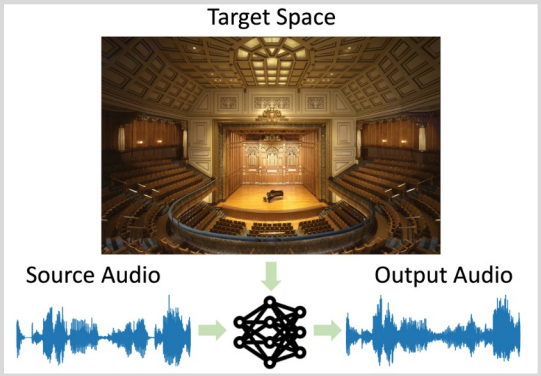
Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021
Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh K. Ramakrishnan, Kristen Grauman

Semantic Audio-Visual Navigation, CVPR 2021
Changan Chen, Ziad Al-Halah, Kristen Grauman

Sound Adversarial Audio-Visual Navigation, ICLR 2022
Yinfeng Yu, Wenbing Huang, Fuhun Sun, Changan Chen, Yikai Wang, Xiaohong Liu

SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, NeurIPS 2022
Changan Chen*, Carl Schissler*, Sanchit Garg*, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, Kristen Grauman

Acoustic Learning



Visual Acoustic Matching, CVPR 2022 (Oral)
Changan Chen, Ruohan Gao, Paul Calamia, Kristen Grauman

VisualEchoes: Spatial Image Representation Learning through Echolocation, ECCV 2020
Ruohan Gao, Changan Chen, Carl Schissler, Ziad Al-Halah, Kristen Grauman

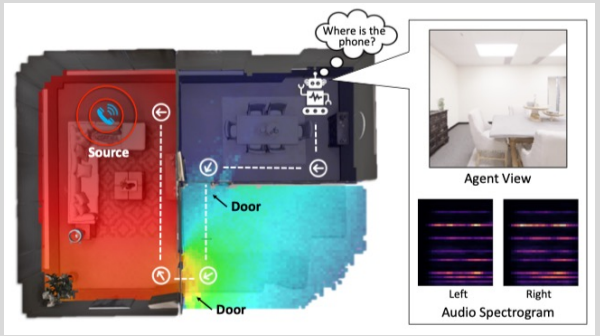
Few-Shot Audio-Visual Learning of Environment Acoustics, NeurIPS 2022
Sagnik Majumder, Changan Chen*, Ziad Al-Halah*, Kristen Grauman

Learning Audio-Visual Dereverberation, Under Review
Changan Chen, Wei Sun, David Harwath, Kristen Grauman

Novel-view Acoustic Synthesis, Under Review
Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Ithapu, Natalia Neverova, Kristen Grauman, Andrea Vedaldi

Visual understanding of room acoustics

Robot Learning



SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020 (Spotlight)
Changan Chen, Unna Jain, Carl Schissler, Sebastia V. Amengual Gari, Ziad Al-Halah, Vamsi K. Ithapu, Philip Robinson, Kristen Grauman

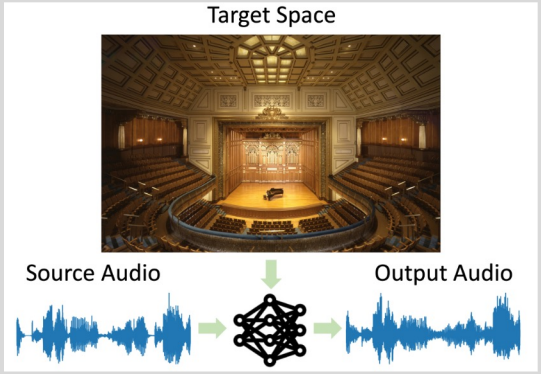
Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021
Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh K. Ramakrishnan, Kristen Grauman

Semantic Audio-Visual Navigation, CVPR 2021
Changan Chen, Ziad Al-Halah, Kristen Grauman

Sound Adversarial Audio-Visual Navigation, ICLR 2022
Yinfeng Yu, Wenbing Huang, Fuhun Sun, Changan Chen, Yikai Wang, Xiaohong Liu

SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, NeurIPS 2022
Changan Chen*, Carl Schissler*, Sanchit Garg*, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, Kristen Grauman

Acoustic Learning



Visual Acoustic Matching, CVPR 2022 (Oral)
Changan Chen, Ruohan Gao, Paul Calamia, Kristen Grauman

VisualEchoes: Spatial Image Representation Learning through Echolocation, ECCV 2020
Ruohan Gao, Changan Chen, Carl Schissler, Ziad Al-Halah, Kristen Grauman

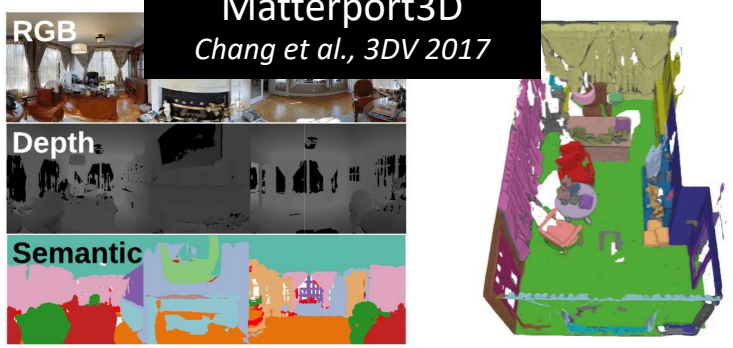
Few-Shot Audio-Visual Learning of Environment Acoustics, NeurIPS 2022
Sagnik Majumder, Changan Chen*, Ziad Al-Halah*, Kristen Grauman

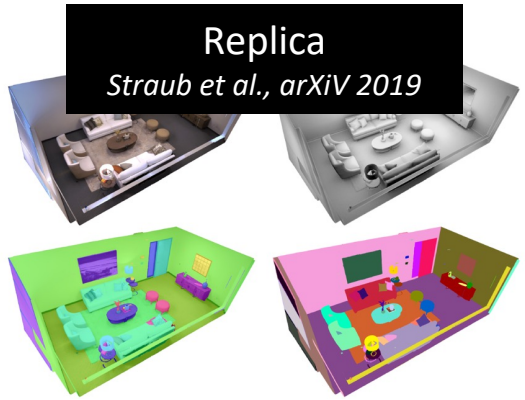
Learning Audio-Visual Dereverberation, Under Review
Changan Chen, Wei Sun, David Harwath, Kristen Grauman

Novel-view Acoustic Synthesis, Under Review
Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Ithapu, Natalia Neverova, Kristen Grauman, Andrea Vedaldi

Simulating embodiment in 3D scenes

Datasets





Simulators

Habitat

Savva et al., ICCV 2019

iGibson

Xia et al., ICRA 2020

AI2THOR

Kovle et al., arXiV 2017

Advantages: Large-scale training, fast experimentation, consistent benchmarking and replicable research

Sim2Real

Sim2Real Predictivity: Does Evaluation in Simulation Predict Real-World Performance, Kadian et al., IRAL 2020
Sim-to-Real Transfer for Vision-and-Language Navigation, Anderson et al., CoRL 2020
RoboThor: An Open Simulation-to-Real Embodied AI Platform, Deitke et al., CVPR 2020

Today's embodied agents (robots) are deaf

- We want robots that can hear and react in the environment

Vision-Only

Gupta et al., 2017

Zhu et al., 2017

Sava et al., 2019

...

Vision-Language

Anderson et al., 2018

Wang et al., 2018

Wang et al., 2019

...

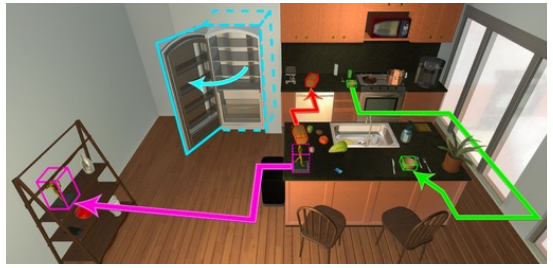
Vision-Interaction

Zhu et al., 2017

Gordon et al., 2018

Wortsman et al all, 2019

...

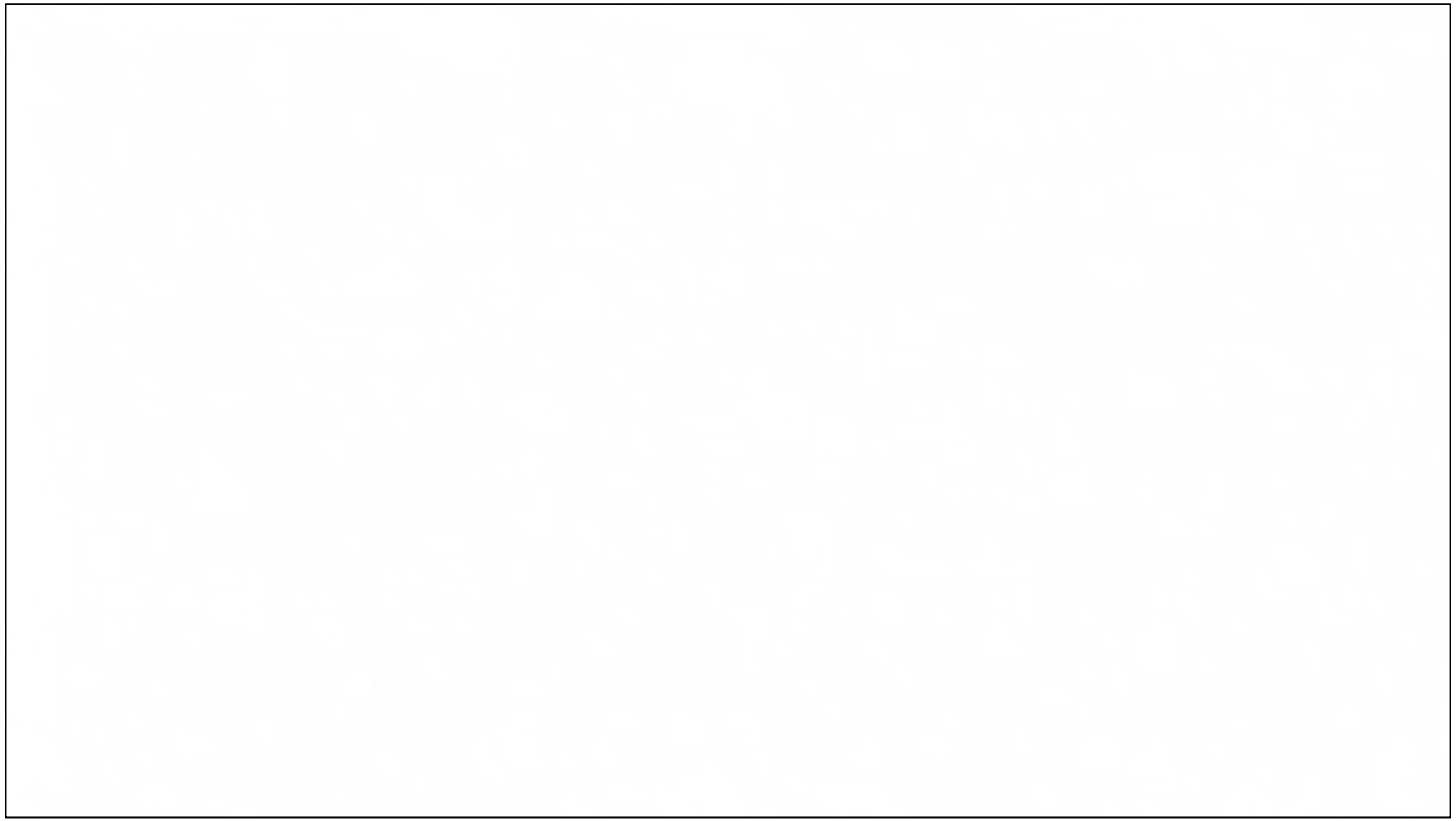


Vision-Audio

Chen and Jain et al., 2020
(this work)

- No existing simulation supports audio rendering
- No existing formulation for audio-visual navigation

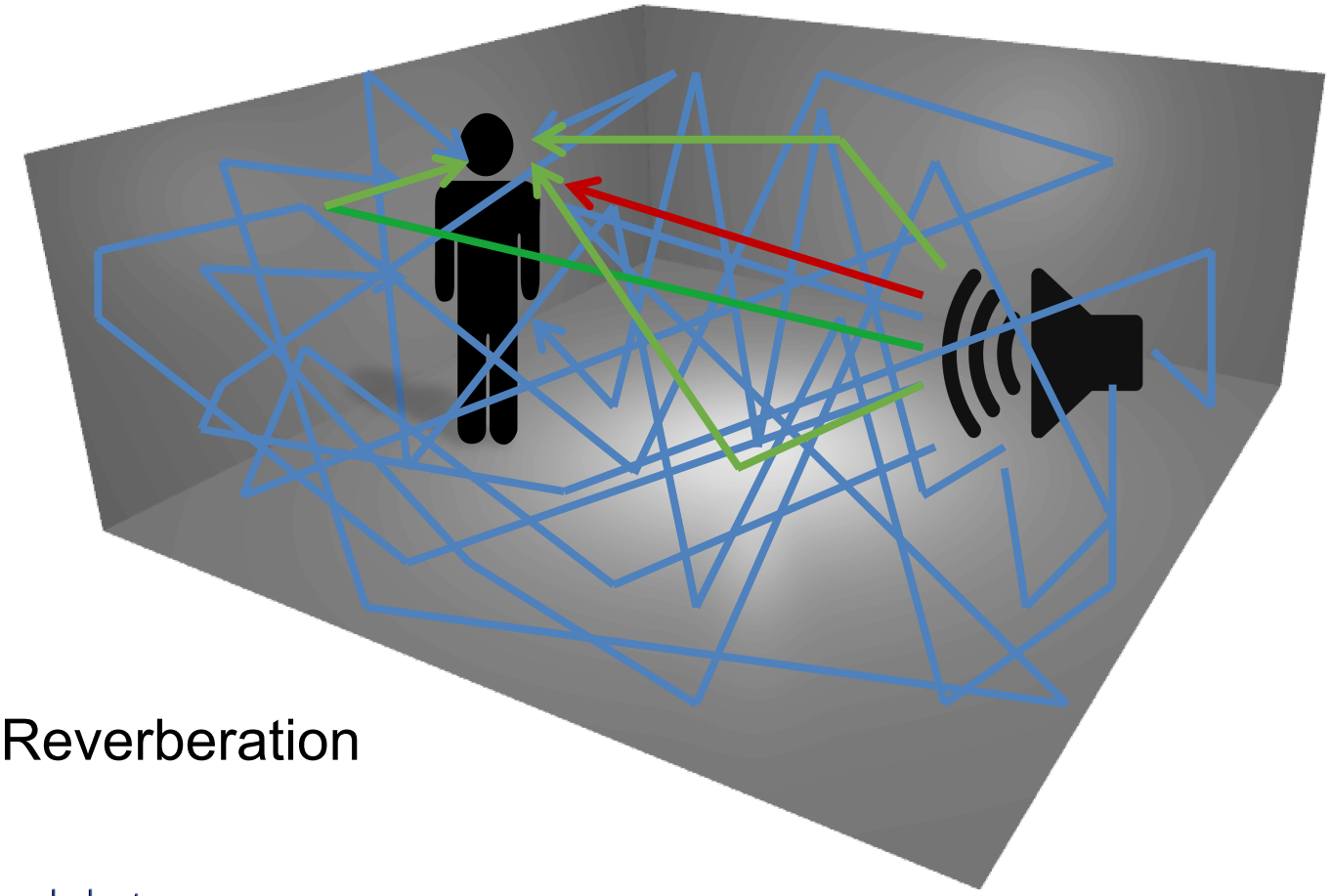
SoundSpaces demo



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

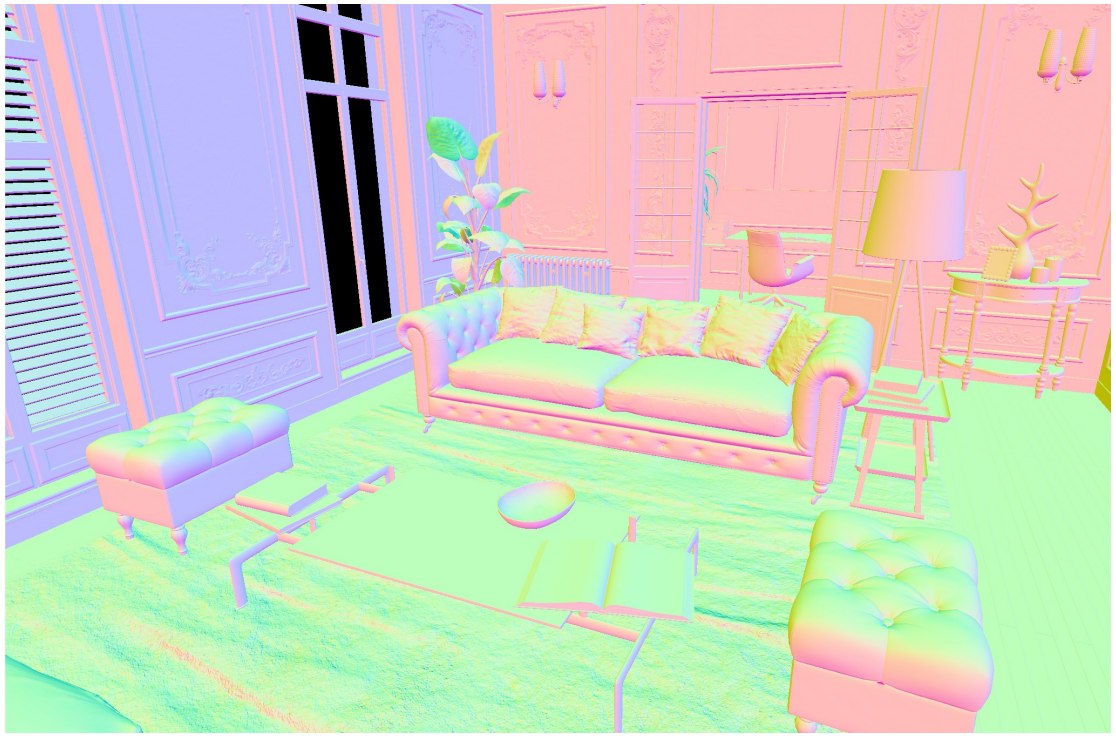
Background: acoustic simulation

Goal: simulate a perceptually-valid approximation of the room impulse response (RIR)

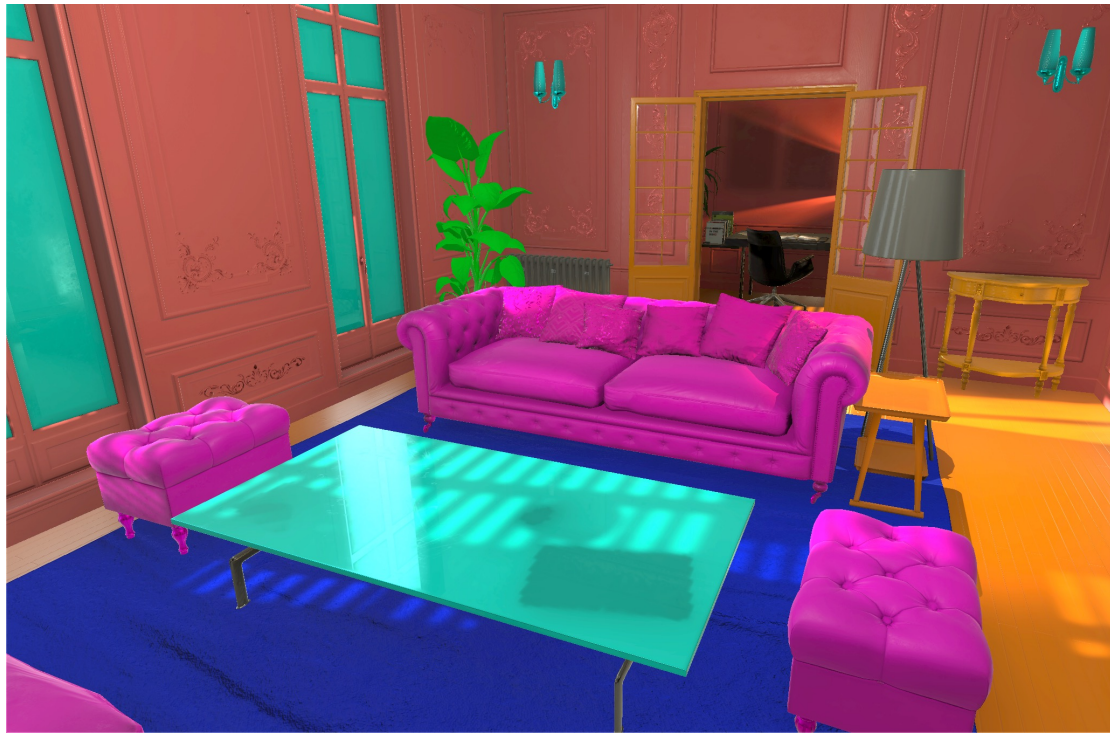


Physics-based audio rendering

3D Geometry



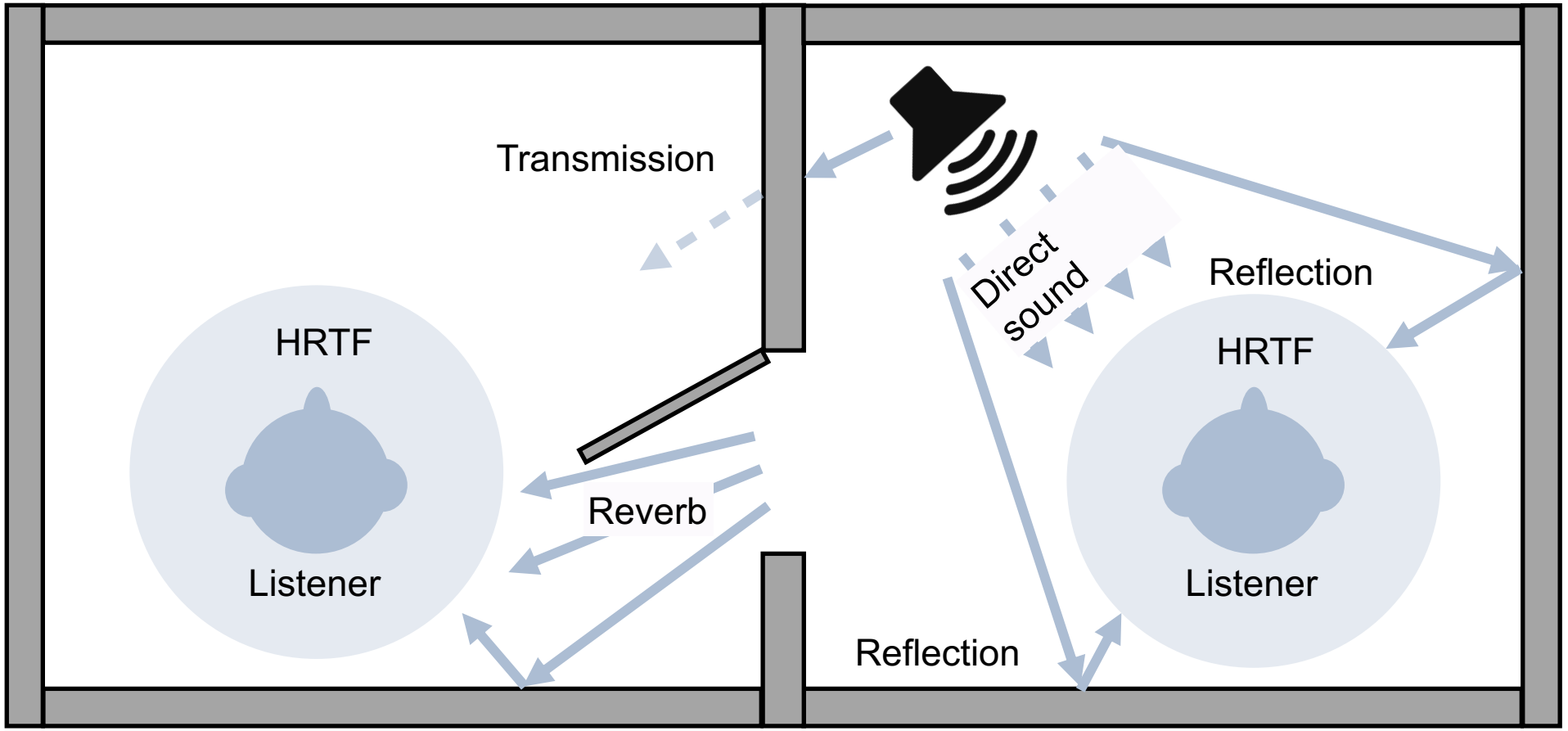
Material Properties



Simulate the sound received by the listener from a source location

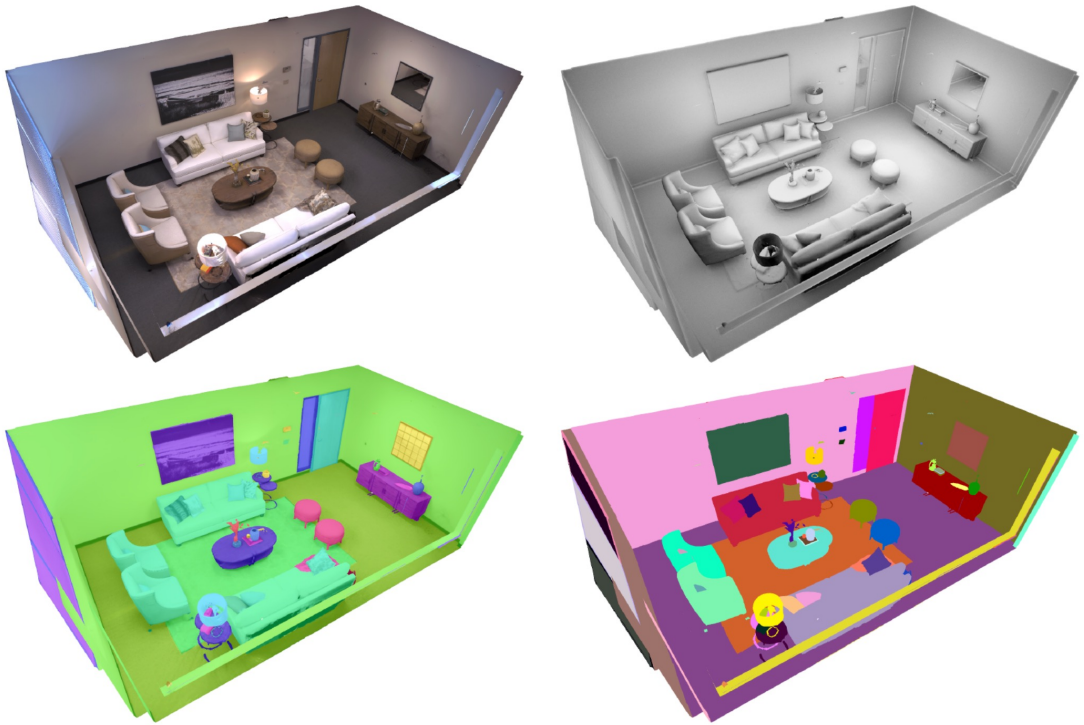
Sound propagation system

3D spatial audio for reflections and reverb with realistic acoustics based on bidirectional ray tracing

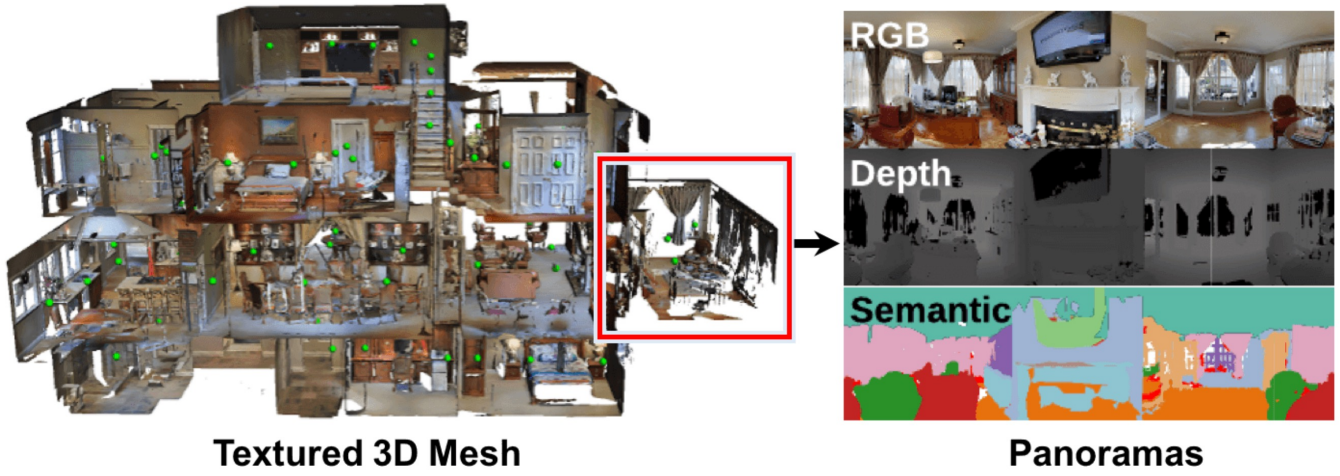


Real-scan environments

Replica[1] dataset



Matterport3D[2] dataset

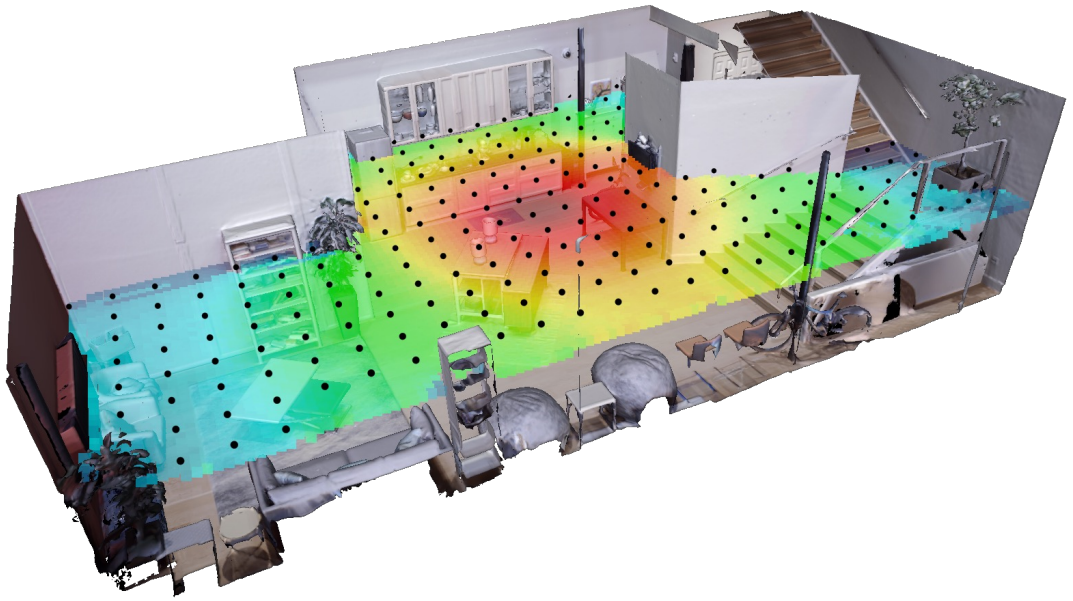


[1]The Replica Dataset: A Digital Replica of Indoor Spaces, Straub et al., arXiv, 2019
[2]Matterport3D: Learning from RGB-D Data in Indoor Environments, Chang et al., 3DV, 2017

SoundSpaces: our audio simulator

SoundSpaces produces realistic audio rendering based on the room geometry, materials, and sound source location by precomputing the room impulse response function (RIR)

Users can insert any sound of their choice at runtime. The received sound is obtained by convolving the RIR with the source sound.



	# Scenes	Avg. Area	# RIRs
Replica	18	47.24 m^2	0.9M
Matterport3D	85	517.34 m^2	16.7M

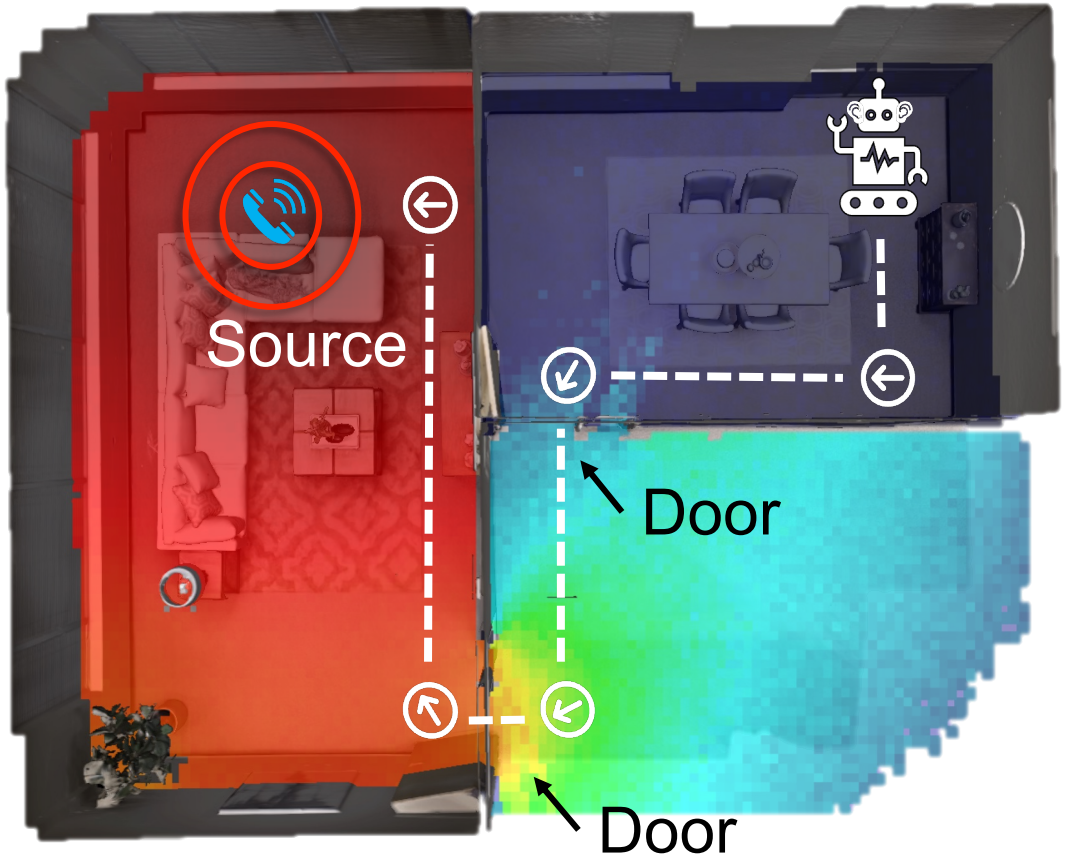
Table: Summary of dataset statistics

Visit soundspaces.org for more information!

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

Audio-visual navigation in 3D environments

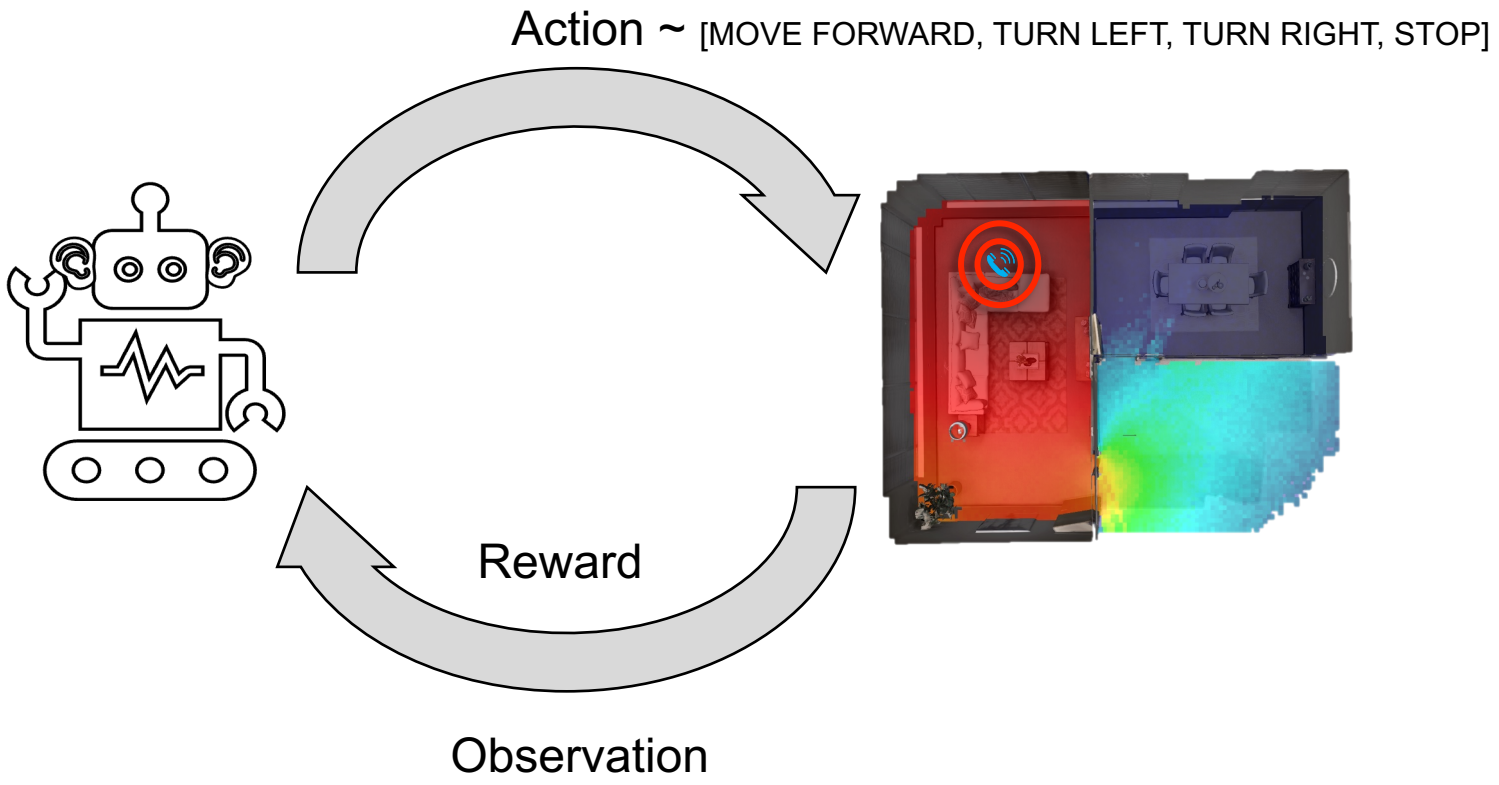
An agent navigates to a sounding object with vision and audio perception



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

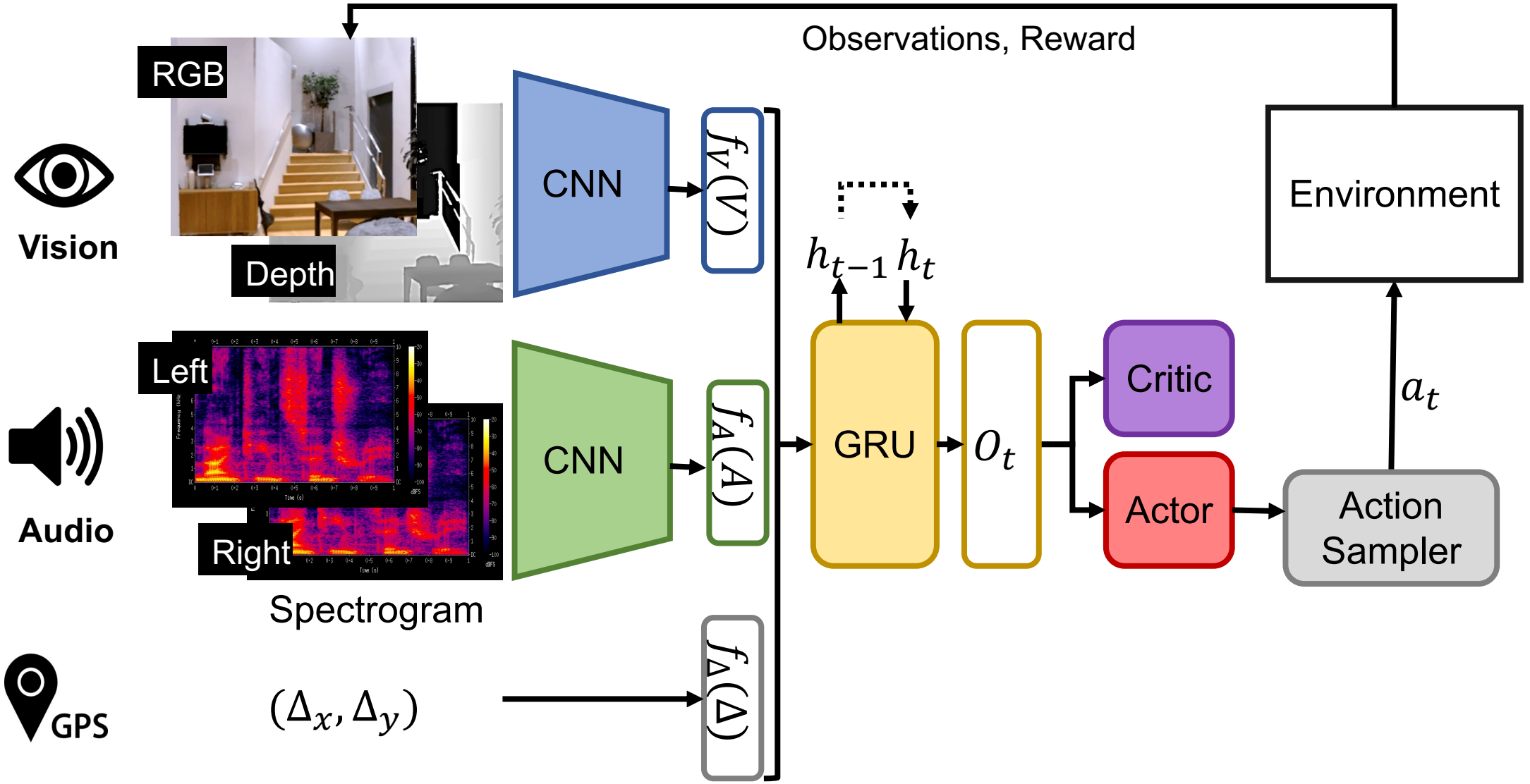
Learning with deep reinforcement learning

- The robot learns to navigate in simulation via trials and errors.
- Rewarded +1 for getting close and +10 for reaching the goal



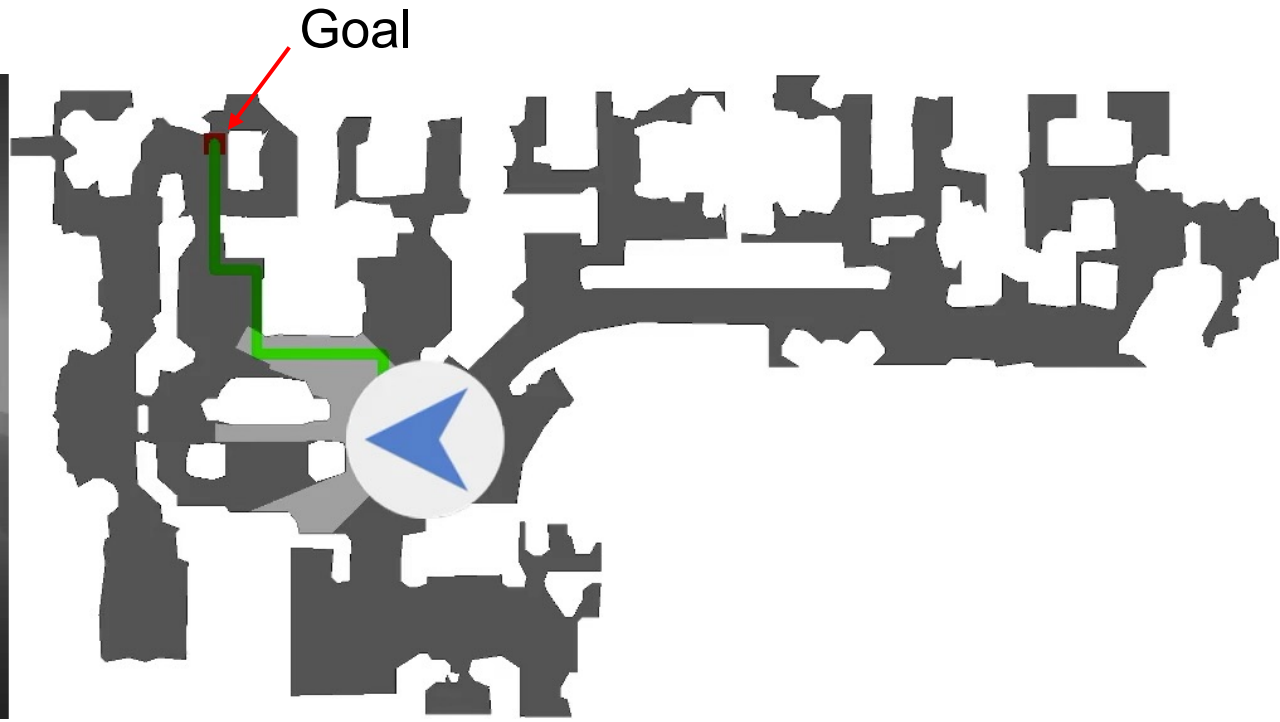
C. Chen*, U. Jain*, et al., SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020

Navigation policy



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

Navigation demo

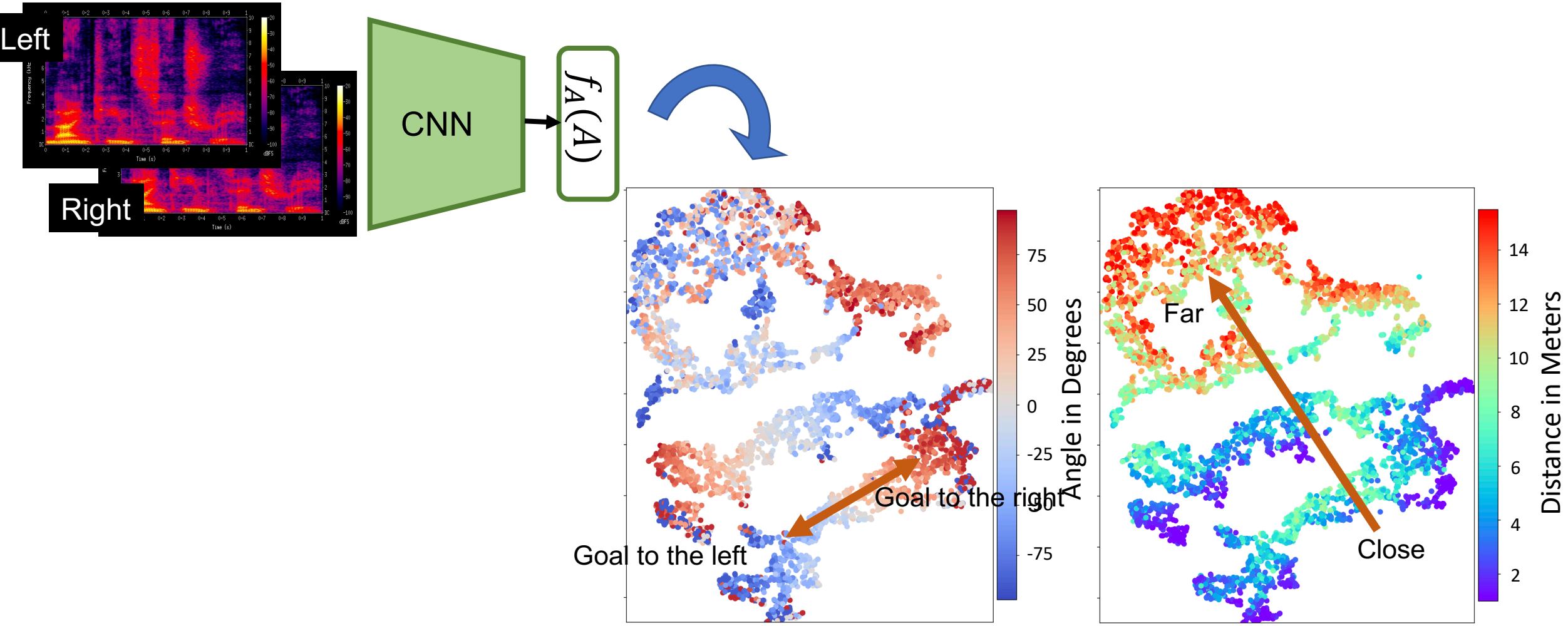


The agent leverages the complementary spatial information in audio and vision, and navigates to the goal efficiently

Agent Goal Start Shortest path Agent path Seen/Unseen area Occupied area Red Frame: Collision

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

What do the learned audio features capture?

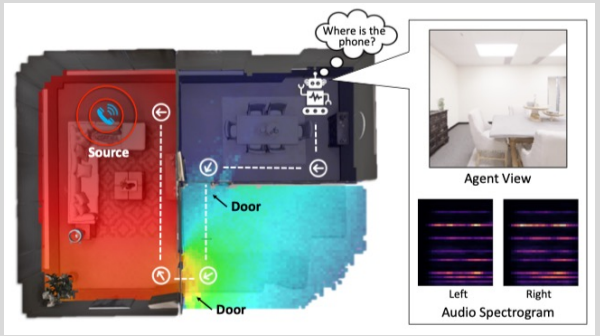


T-SNE of audio features from an AudioGoal agent

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

Empower robots to hear in spaces

Robot Learning



SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020 (Spotlight)
Changan Chen, Unna Jain, Carl Schissler, Sebastia V. Amengual Gari, Ziad Al-Halah, Vamsi K. Ithapu, Philip Robinson, Kristen Grauman

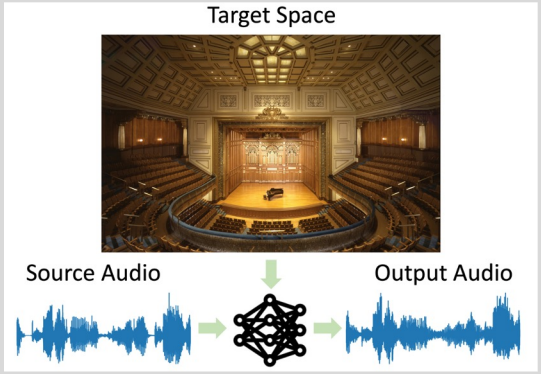
Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021
Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh K. Ramakrishnan, Kristen Grauman

Semantic Audio-Visual Navigation, CVPR 2021
Changan Chen, Ziad Al-Halah, Kristen Grauman

Sound Adversarial Audio-Visual Navigation, ICLR 2022
Yinfeng Yu, Wenbing Huang, Fuhun Sun, Changan Chen, Yikai Wang, Xiaohong Liu

SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, NeurIPS 2022
Changan Chen*, Carl Schissler*, Sanchit Garg*, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, Kristen Grauman

Acoustic Learning



Visual Acoustic Matching, CVPR 2022 (Oral)
Changan Chen, Ruohan Gao, Paul Calamia, Kristen Grauman

VisualEchoes: Spatial Image Representation Learning through Echolocation, ECCV 2020
Ruohan Gao, Changan Chen, Carl Schissler, Ziad Al-Halah, Kristen Grauman

Few-Shot Audio-Visual Learning of Environment Acoustics, NeurIPS 2022
Sagnik Majumder, Changan Chen*, Ziad Al-Halah*, Kristen Grauman

Learning Audio-Visual Dereverberation, Under Review
Changan Chen, Wei Sun, David Harwath, Kristen Grauman

Novel-view Acoustic Synthesis, Under Review
Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Ithapu, Natalia Neverova, Kristen Grauman, Andrea Vedaldi

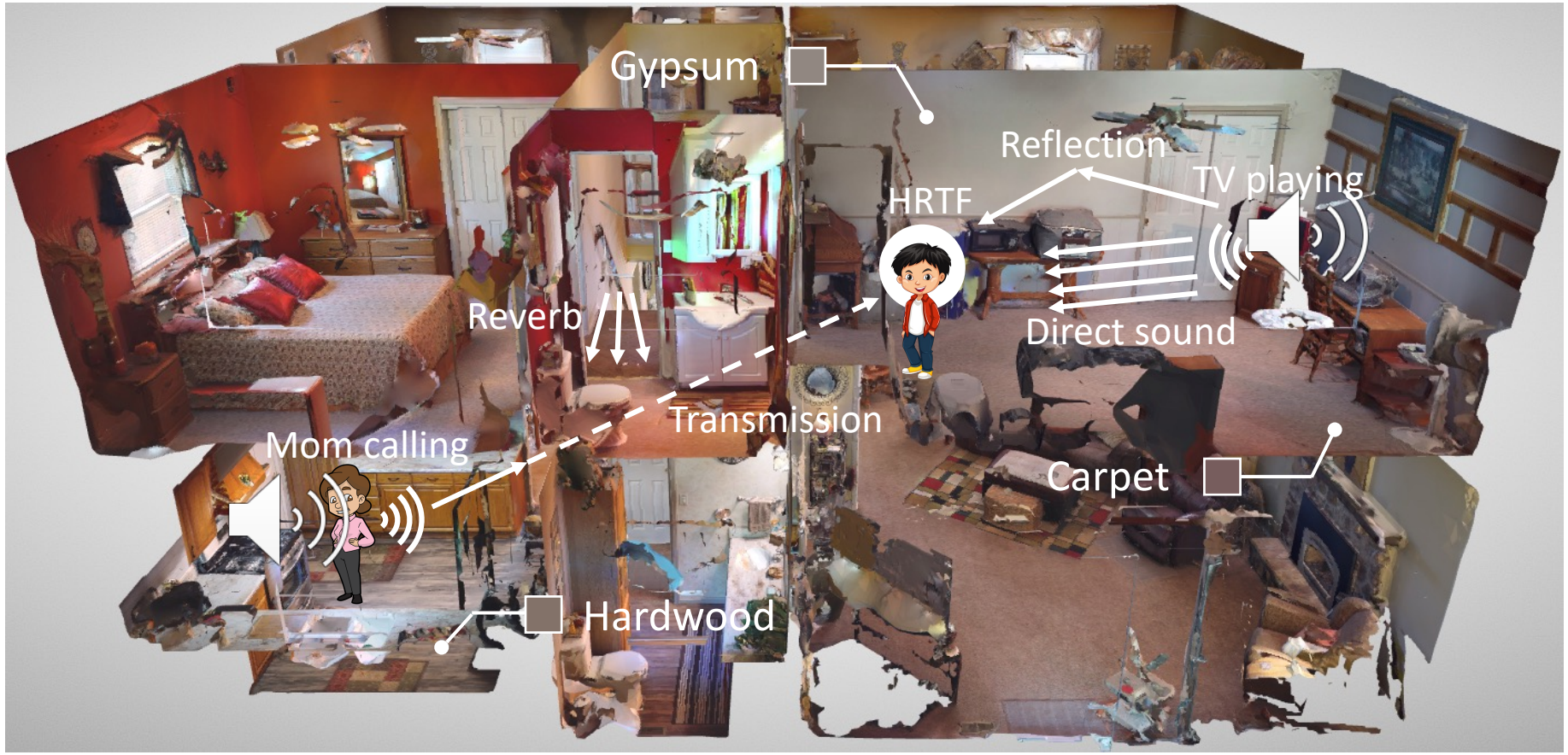
SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning

Changan Chen*[1,4], Carl Schissler*[2], Sanchit Garg*[2], Philip Kobernik[2], Alexander William Clegg[4], Paul Calamia[2], Dhruv Batra[3,4], Philip Robinson[2], Kristen Grauman[1,4]

[1]UT Austin, [2]Reality Labs at Meta, [3]Georgia Tech, [4]FAIR

NeurIPS 2022

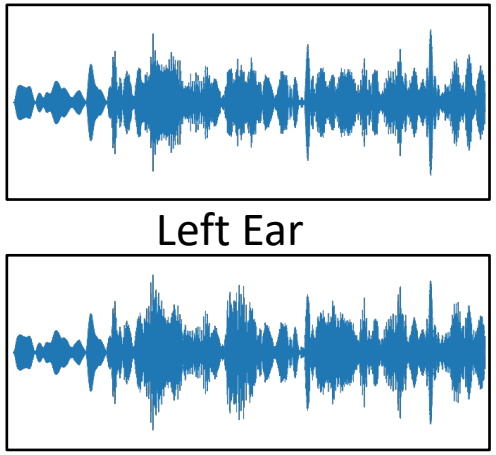
A fast, continuous, configurable and generalizable audio-visual simulation platform



Visual Render

RGB

Acoustic Render

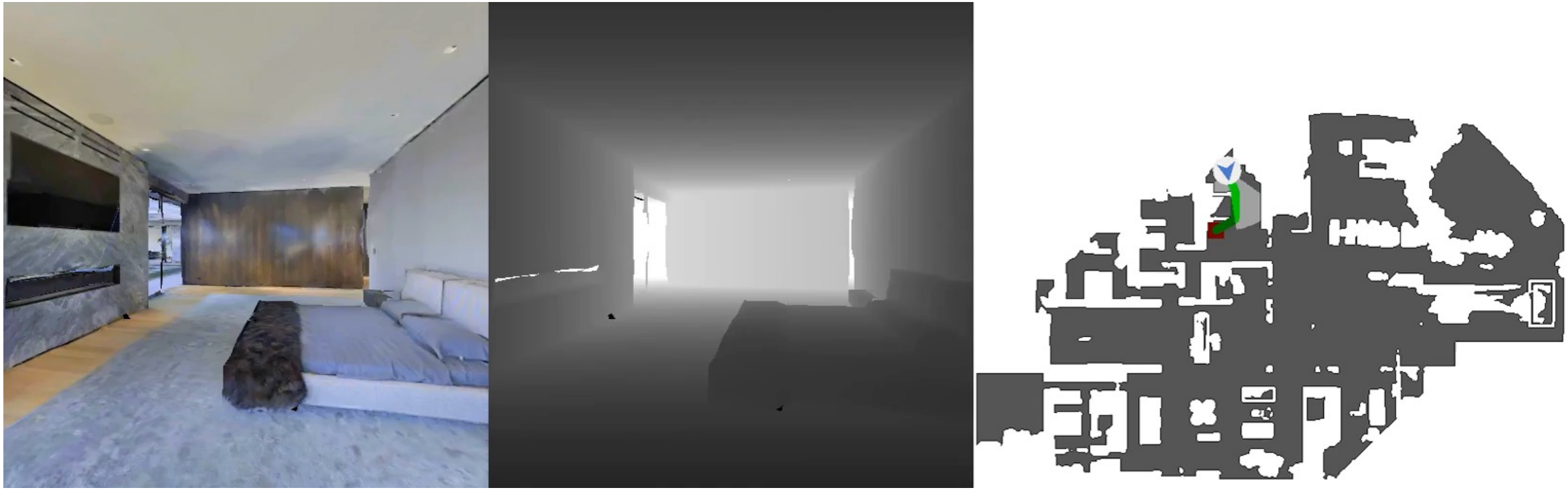


Left Ear

Right Ear

Continuous rendering

We offer both spatial and acoustic continuity.



Navigating to a sliding drawer

Configurable simulation

You can change all these parameters!

Simulation parameters

- Frequency bands
- Direct sound
- Indirect sound
- Transmission
- Diffraction
- Number of rays
- Number of threads
- Sample rate
- …

Microphone types

- Mono
- Binaural
- Stereo
- Quad
- Surround_5_1
- Surround_7_1
- Ambisonics
- Your mic array
- …

Material properties

- Absorption coefficients
- Scattering coefficients
- Transmission coefficients
- Damping coefficients
- Frequency band specs
- Instance level config
- …

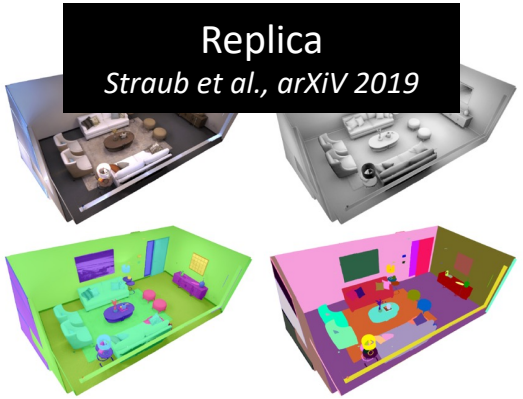
Generalizable simulation

We support arbitrary scene datasets.

Gibson
Xia et al., CVPR 2018

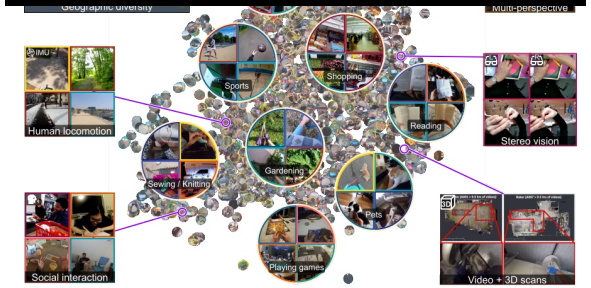
Matterport3D
Chang et al., 3DV 2017

Replica
Straub et al., arXiV 2019



HM3D
Ramakrishnan et al., NeurIPS 2021

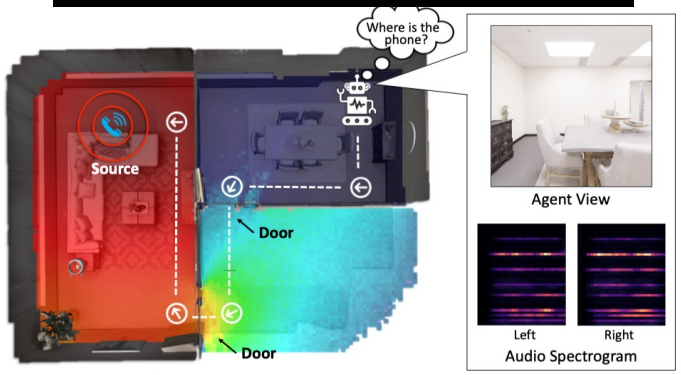
Ego4D
Grauman et al., CVPR 2022



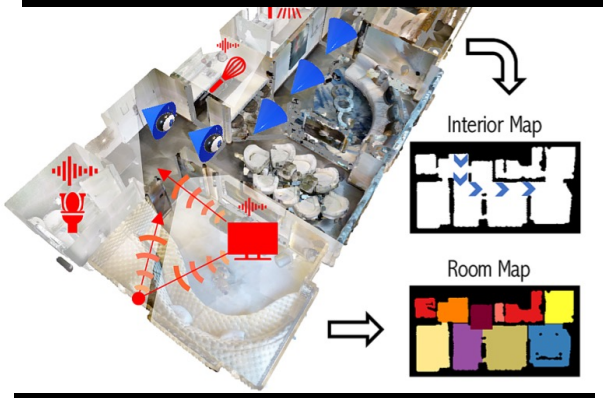
Visual-acoustic learning

We support an array of tasks!

Audio-Visual Navigation
Chen et al., ECCV 2020

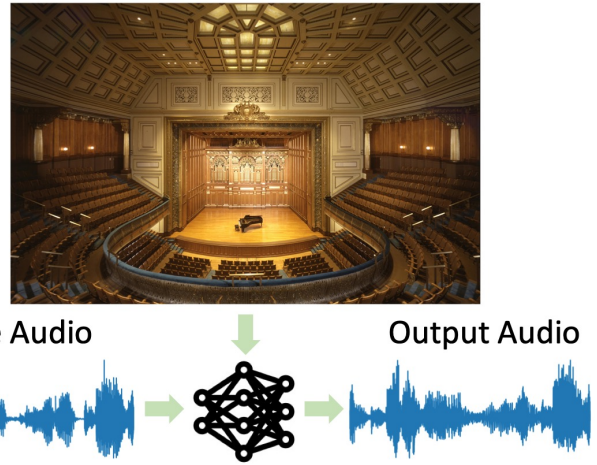


Audio-Visual Mapping
Purushwalkam et al., ICCV 2021

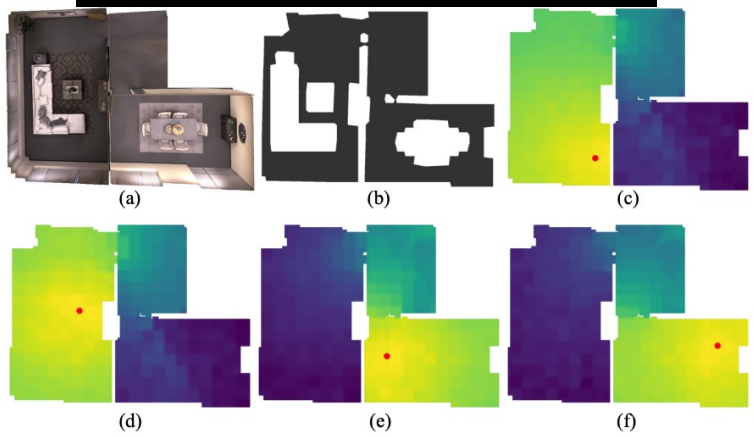


Audio-Visual Separation
Majumder et al., ICCV 2021

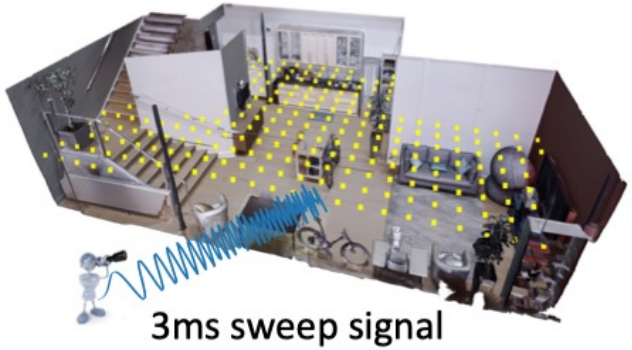
Visual Acoustic Matching
Chen et al.,CVPR 2022



Neural Acoustic Rendering
Luo et al., arXiv 2022



Echolocation Learning
Gao et al., ECCV 2020



Studying acoustics as a perception task

How the space transform the sound we hear?

What is the room geometry?

Do these glass doors lead to longer reverb?

Where are the speaker and listener?

1 drum kit 5 different spaces

Source: Shred Shed Studio

Visual acoustic learning

Can we alter the acoustic signature of the sound if we understand the acoustics of the space based on visuals?

Augmented reality

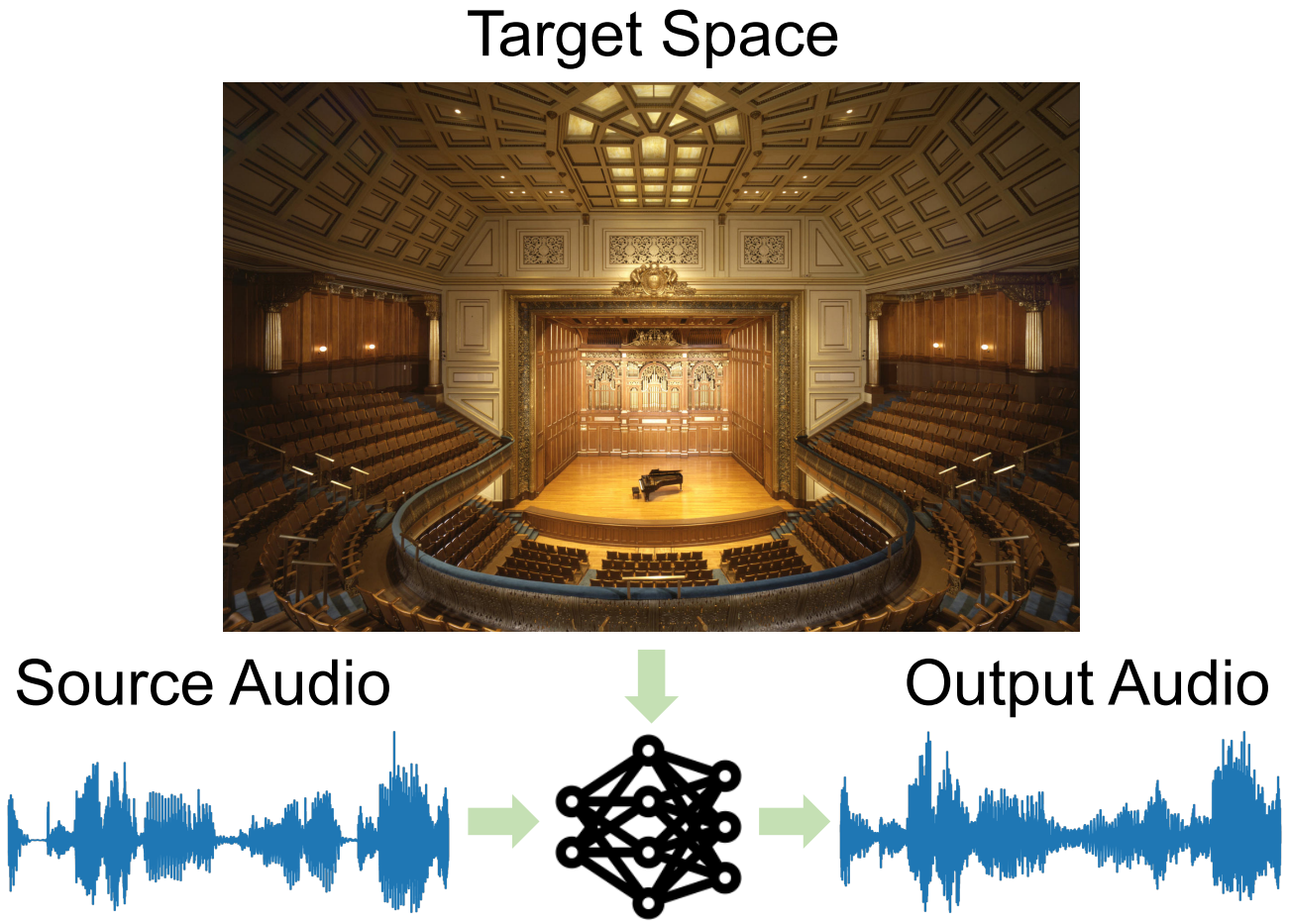
Film dubbing

Video conferencing

Chen et al., Visual Acoustic Matching, CVPR 2022

The visual acoustic matching task

We propose to transform the sound recorded in one space to another depicted in the target visual scene.

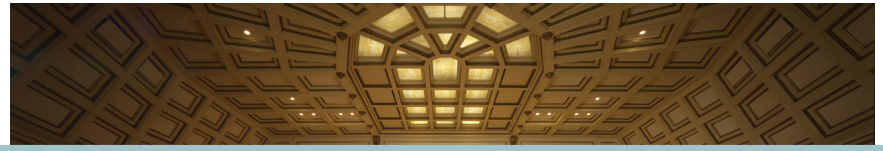


Chen et al., Visual Acoustic Matching, CVPR 2022

The visual acoustic matching task

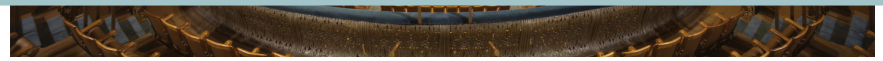
We propose to transform the sound recorded in one space to another depicted in the target visual scene.

Target Space

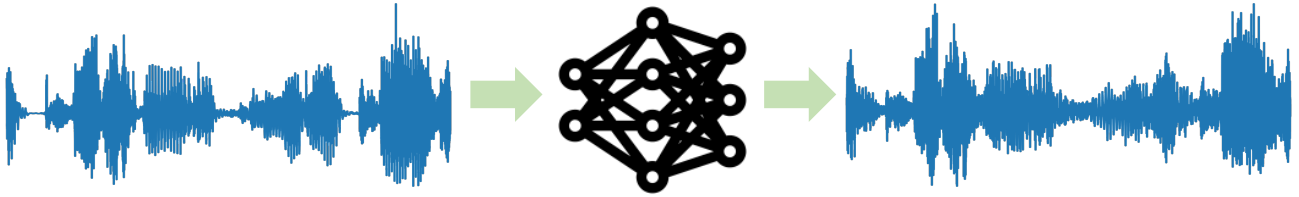


Main challenges:

1. Crossmodal (audio-visual) reasoning
2. Obtaining the right data for the task



Source Audio

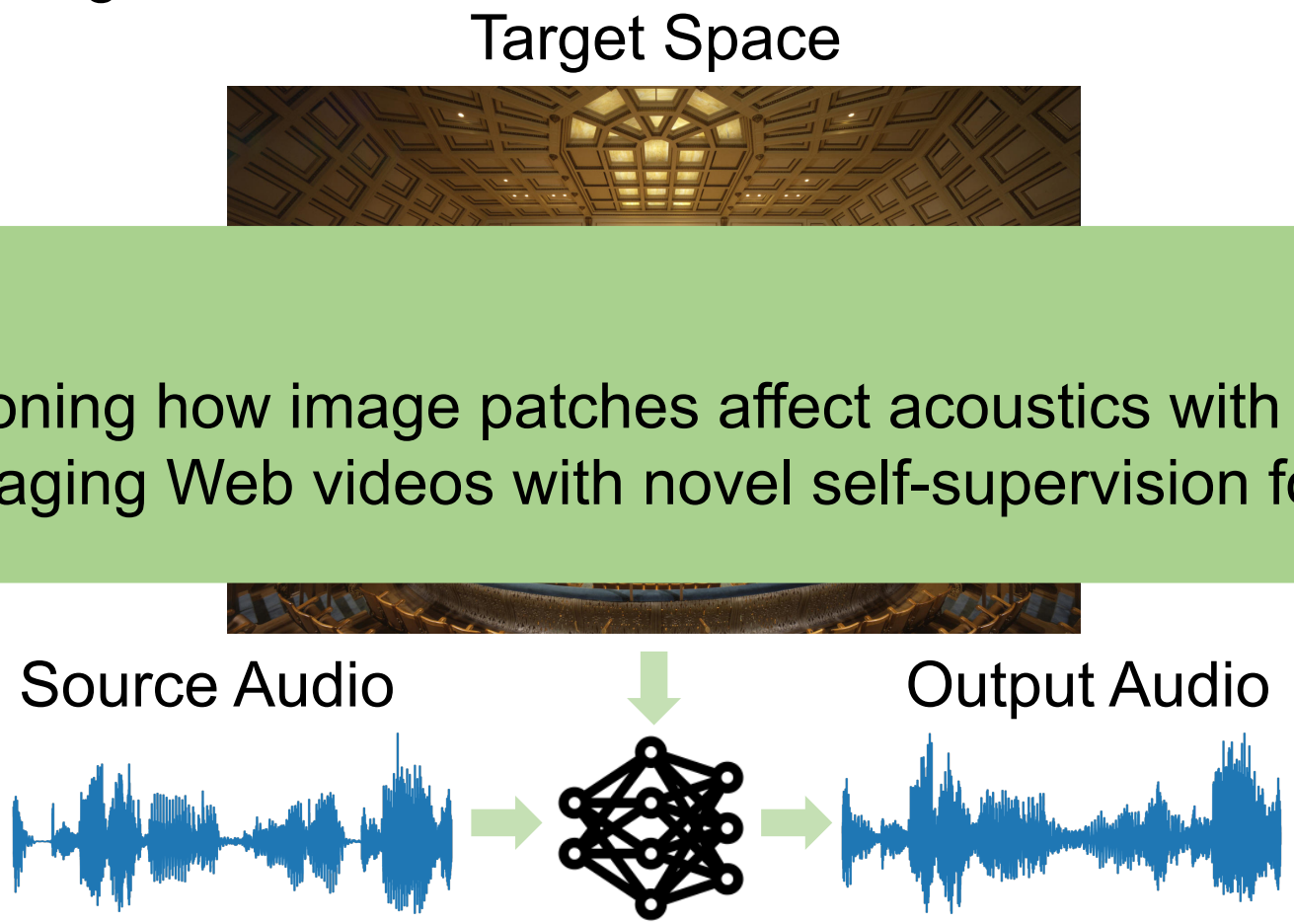


Output Audio

Chen et al., Visual Acoustic Matching, CVPR 2022

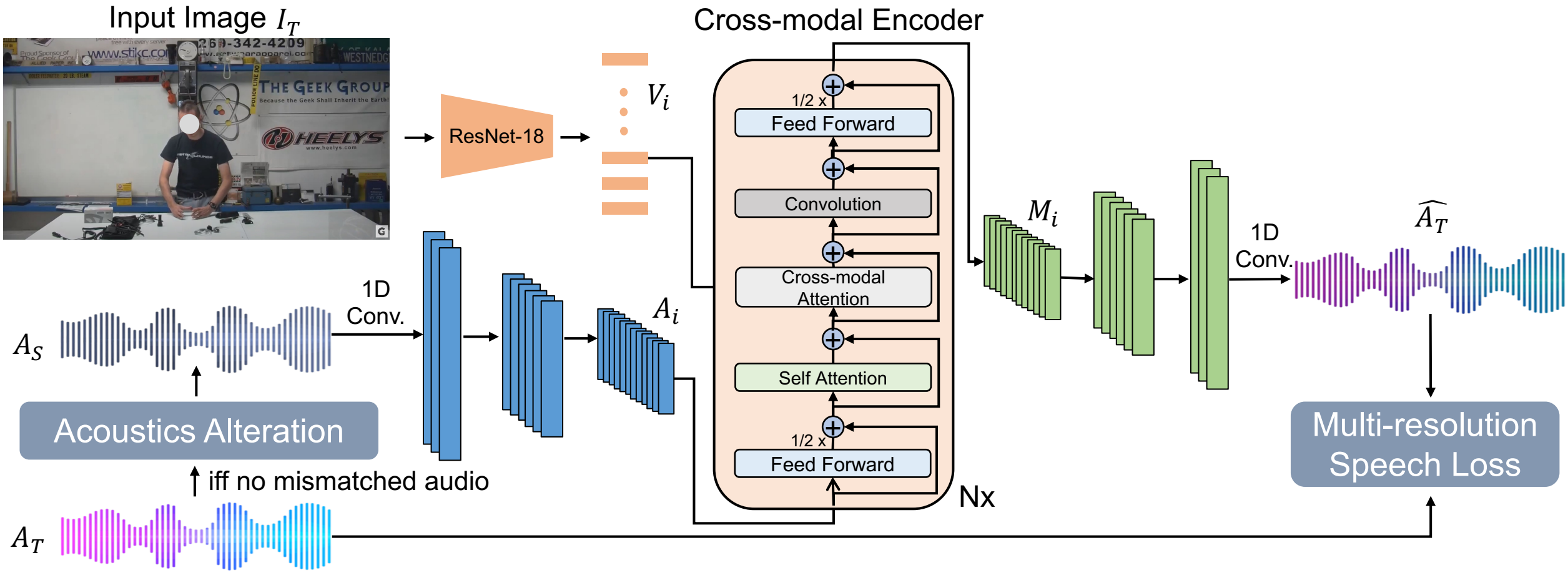
The visual acoustic matching task

We propose to transform the sound recorded in one space to another depicted in the target visual scene.



Chen et al., Visual Acoustic Matching, CVPR 2022

Audio-Visual Transformer for Audio Generation (AViTAR)

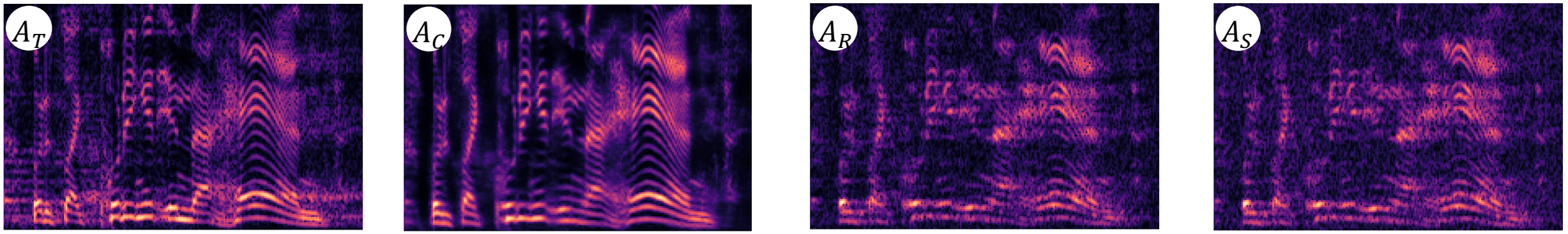


Chen et al., Visual Acoustic Matching, CVPR 2022

Acoustics alteration strategy

Goal: create audio with the same content but different acoustics as self-supervision.

Dereverberation → Acoustic Randomization → Adding Noise



Chen et al., Visual Acoustic Matching, CVPR 2022

Datasets

SoundSpaces-Speech

- Panoramic observation of the environment
- Impulse responses are available
- Serves as a clean test bed

Acoustic AVSpeech

- A web speech video dataset
- Single speaker and no interfering noise
- No impulse responses available
- Use acoustics alteration strategy to obtain inputs

Chen et al., Visual Acoustic Matching, CVPR 2022

Evaluation Metrics

STFT Distance:

- Closeness to the ground truth (applicable only to synthetic dataset)
- Mean squared error between two magnitude spectrograms

RT60 Error (RTE):

- Correctness of the synthesized acoustics
- RT60 is defined as the time of reverberation decaying by 60dB

Mean Opinion Score Error (MOSE):

- The speech quality preserved in the synthesized speech
- Difference between MOS of target speech and synthesized speech

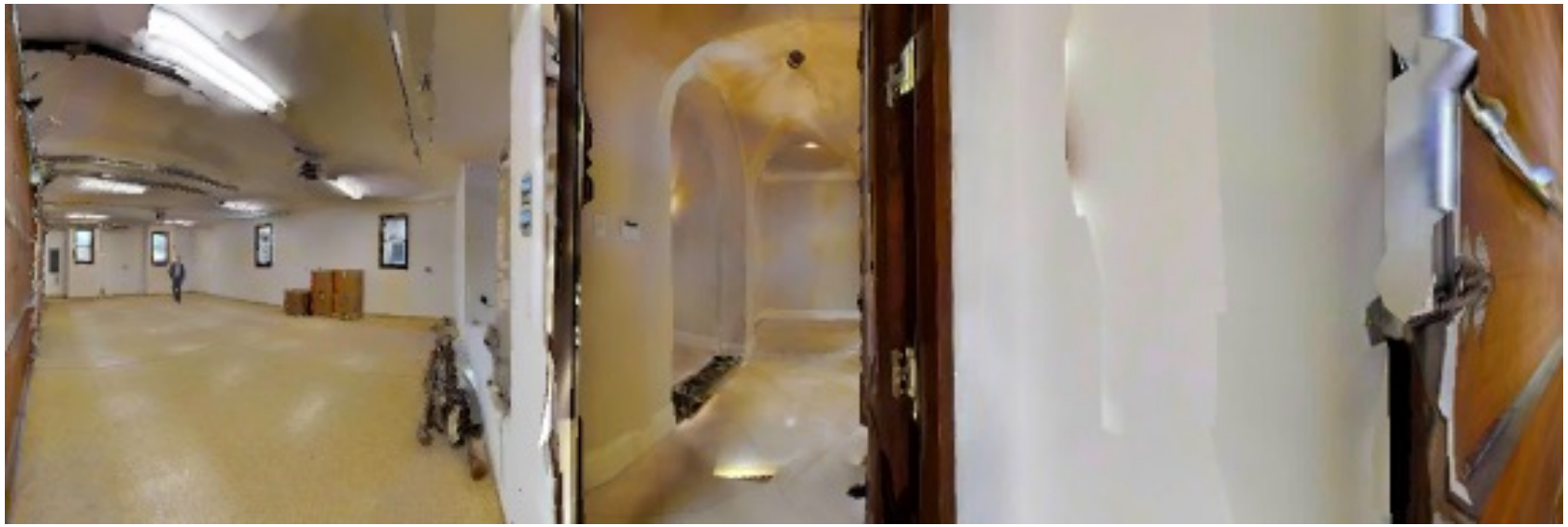
Experiment results

- Strongly outperforms traditional and heavily supervised approaches
- Acoustics is better estimated for seen images

	SoundSpaces-Speech						*Acoustic AVSpeech*			
	Seen			*Unseen*			*Seen*		*Unseen*	
	STFT	RTE (s)	MOSE	STFT	RTE (s)	MOSE	RTE (s)	MOSE	RTE (s)	MOSE
Input audio	1.192	0.331	0.617	1.206	0.356	0.611	0.387	0.658	0.392	0.634
Blind Reverberator [61]	1.338	0.044	0.312	-	-	-	-	-	-	-
Image2Reverb [52]	2.538	0.293	0.508	2.318	0.317	0.518	-	-	-	-
AV U-Net [20]	**0.638**	0.095	0.353	**0.658**	0.118	0.367	0.156	0.570	0.188	0.540
AViTAR w/o visual	0.862	0.140	0.217	0.902	0.186	0.236	0.194	0.504	0.207	0.478
AViTAR	0.665	**0.034**	**0.161**	0.822	**0.062**	**0.195**	**0.144**	**0.481**	**0.183**	**0.453**

Examples on SoundSpaces-Speech

In this example, we show comparison of our model with baselines on SoundSpaces-Speech (unseen).



Anechoic GT Target AViTAR Image2Reverb[1] AV U-Net [2]

[1] Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis, Singh et al., ICCV 2021
[2] 2.5D Visual Sound, Gao et al., CVPR 2019

Matching different environments on AVSpeech

	Office	Garage	Auditorium
Input			
AViTAR			
RT60	0.34s	0.40s	0.58s

Our AViTAR model reasons the image content and learns to inject more reverberation into the speech as the environment gets larger.

Chen et al., Visual Acoustic Matching, CVPR 2022

Augmented reality demo

Here we show one example of virtual phone call in augmented reality where we want a remote participant to sound like they are in the room with us.

Input

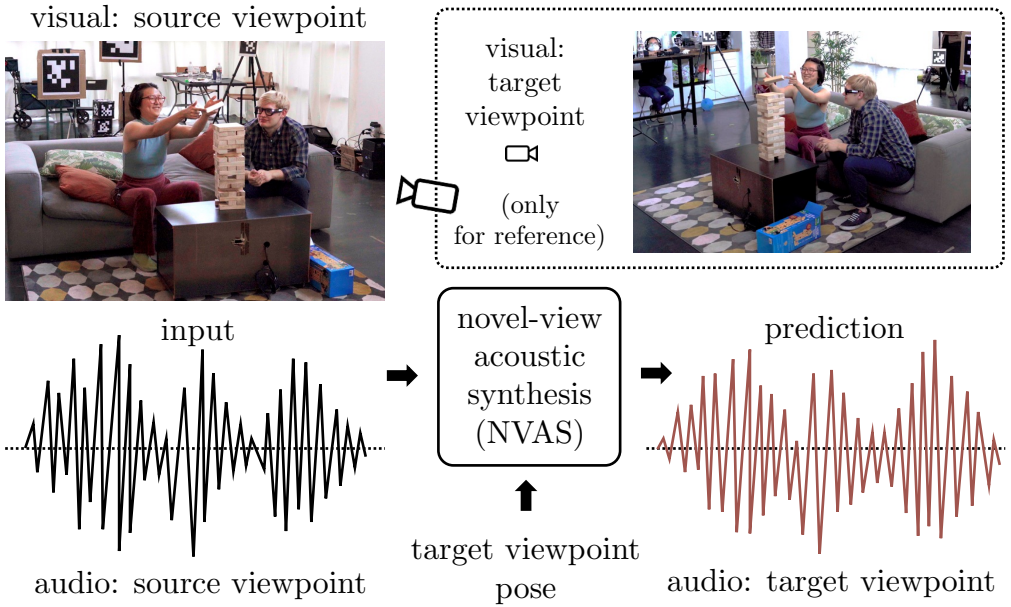
Output

Chen et al., Visual Acoustic Matching, CVPR 2022

Novel-view Acoustic Synthesis

Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Ithapu,
Natalia Neverova, Kristen Grauman, Andrea Vedaldi

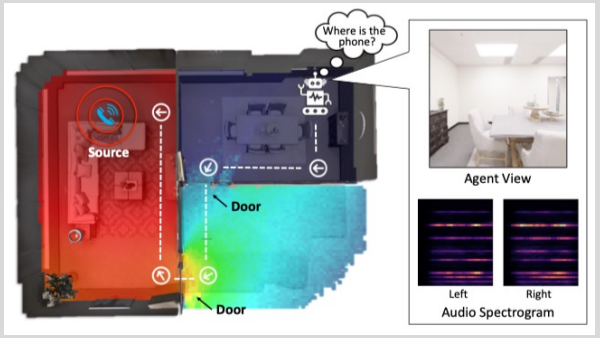
Under Review



Collected hundreds of hours of multi-view AV data

Conclusion

Robot Learning



SoundSpaces: Audio-Visual Navigation in 3D Environments, ECCV 2020 (Spotlight)
Changan Chen, Unna Jain, Carl Schissler, Sebastia V. Amengual Gari, Ziad Al-Halah, Vamsi K. Ithapu, Philip Robinson, Kristen Grauman

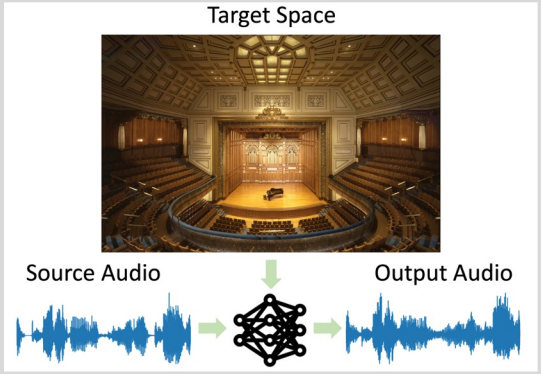
Learning to Set Waypoints for Audio-Visual Navigation, ICLR 2021
Changan Chen, Sagnik Majumder, Ziad Al-Halah, Ruohan Gao, Santhosh K. Ramakrishnan, Kristen Grauman

Semantic Audio-Visual Navigation, CVPR 2021
Changan Chen, Ziad Al-Halah, Kristen Grauman

Sound Adversarial Audio-Visual Navigation, ICLR 2022
Yinfeng Yu, Wenbing Huang, Fuhun Sun, Changan Chen, Yikai Wang, Xiaohong Liu

SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, NeurIPS 2022
Changan Chen*, Carl Schissler*, Sanchit Garg*, Philip Kobernik, Alexander Clegg, Paul Calamia, Dhruv Batra, Philip W Robinson, Kristen Grauman

Acoustic Learning



Visual Acoustic Matching, CVPR 2022 (Oral)
Changan Chen, Ruohan Gao, Paul Calamia, Kristen Grauman

VisualEchoes: Spatial Image Representation Learning through Echolocation, ECCV 2020
Ruohan Gao, Changan Chen, Carl Schissler, Ziad Al-Halah, Kristen Grauman

Few-Shot Audio-Visual Learning of Environment Acoustics, NeurIPS 2022
Sagnik Majumder, Changan Chen*, Ziad Al-Halah*, Kristen Grauman

Learning Audio-Visual Dereverberation, Under Review
Changan Chen, Wei Sun, David Harwath, Kristen Grauman

Novel-view Acoustic Synthesis, Under Review
Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Ithapu, Natalia Neverova, Kristen Grauman, Andrea Vedaldi