Visual-Acoustic Learning

Changan Chen
changan.io
UT Austin
06/10/2023

The University of Texas at Austin

Human perception is a multisensory experience

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

Traditional audio-visual learning

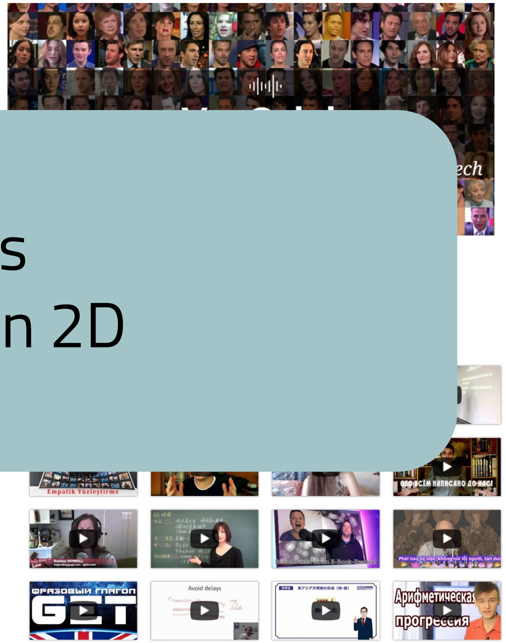
Audio-visual classification

ACAV100M, 2021

Audio-visual
localization/separation

MUSIC-Synthetic, 2020

Speech separation



AVSpeech, 2018

Datasets: single, passively collected videos
Tasks: discriminative and focused on objects in 2D

1 drum kit 5 different spaces

Video source: Shred Shed Studio

Augmented reality / virtual reality

Immersive experience

Enhanced hearing

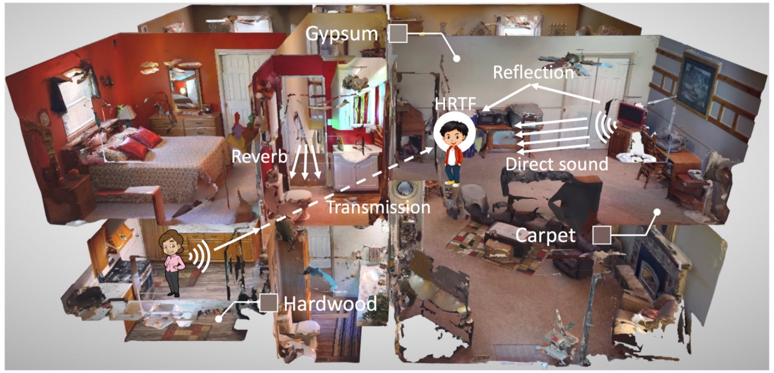
AR/VR systems that create immersive experience for users as well as augment the hearing ability of the device wearer

Visual-acoustic learning

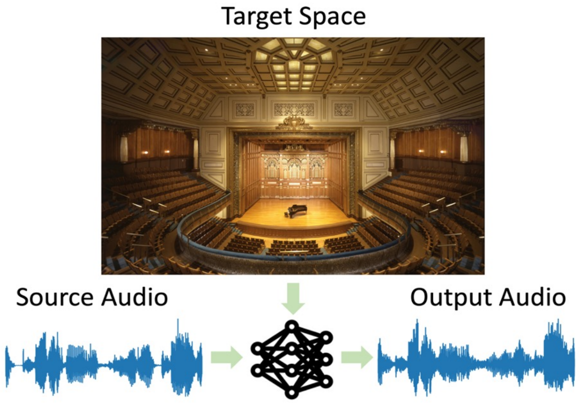
Learning how sounds are situated, produced and transformed physically in spaces based on visual inputs

AV4D: 3 spatial dimensions + 1 temporal dimension

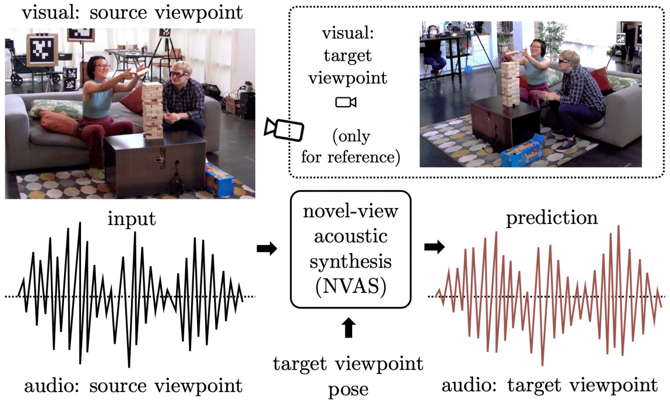
SoundSpaces 2.0, NeurIPS 22



Visual-acoustic Matching, CVPR 22



Novel-view Acoustic Synthesis, CVPR 23



Collecting data is expensive!

- Acoustic data measured with room impulse response
- A recording is only good for one source/receiver location pair
- Expensive to scale up

Image source: OpenAIR dataset

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]

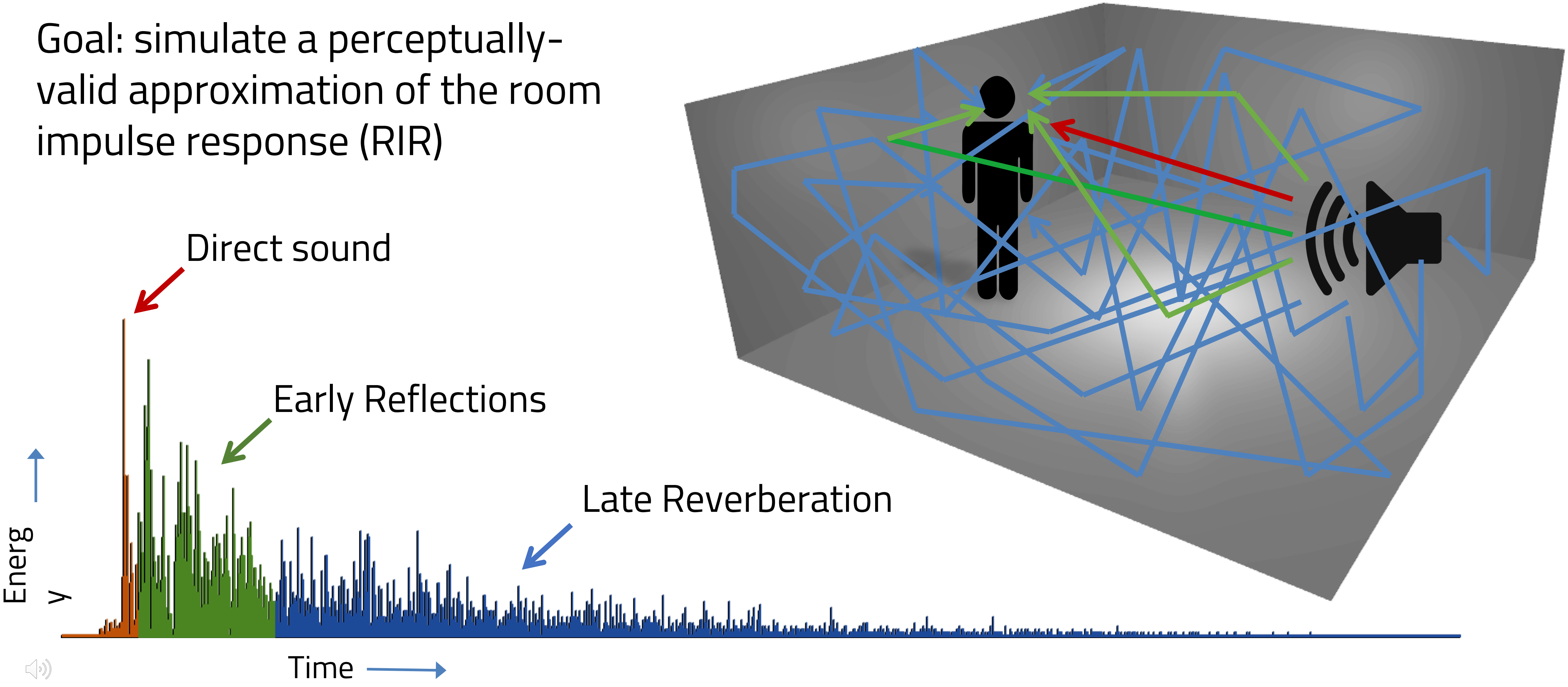
[1]UT Austin, [2]Reality Labs at Meta, [3]Georgia Tech, [4]FAIR

NeurIPS 2022

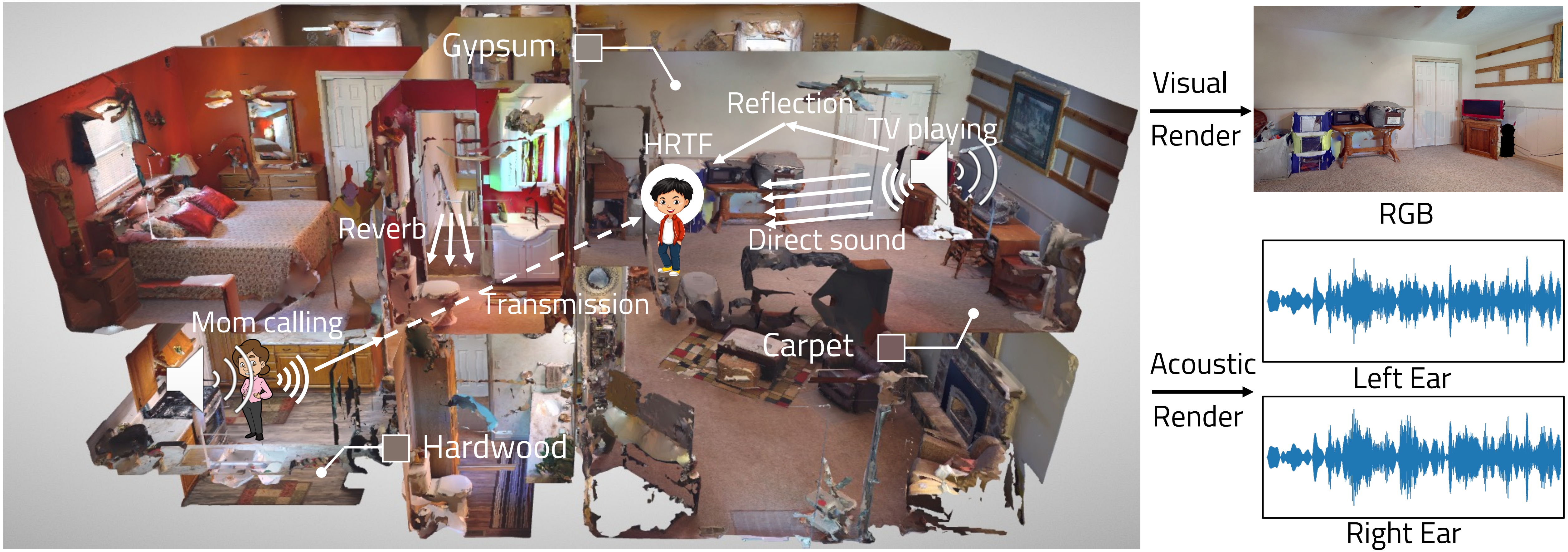
SoundSpaces

Acoustic simulation

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



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



SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, Changan Chen et al., NeurIPS 2022

SoundSpaces demo for navigation

SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, Changan Chen et al., NeurIPS 2022

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
- ...

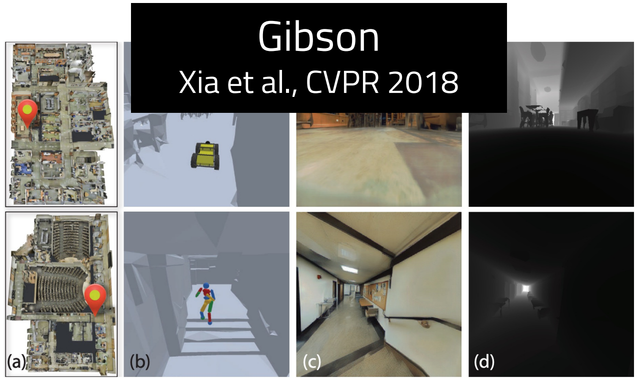
Material properties

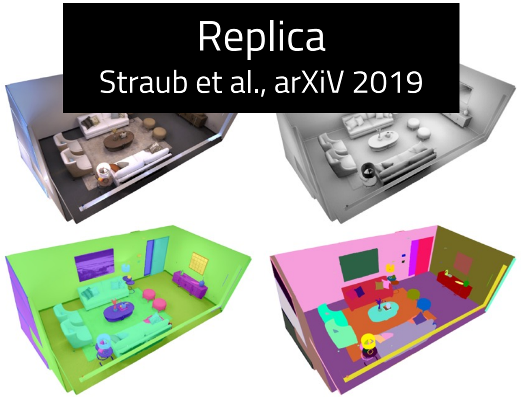
- Absorption coefficients
- Scattering coefficients
- Transmission coefficients
- Damping coefficients
- Frequency band specs
- Instance level config
- ...

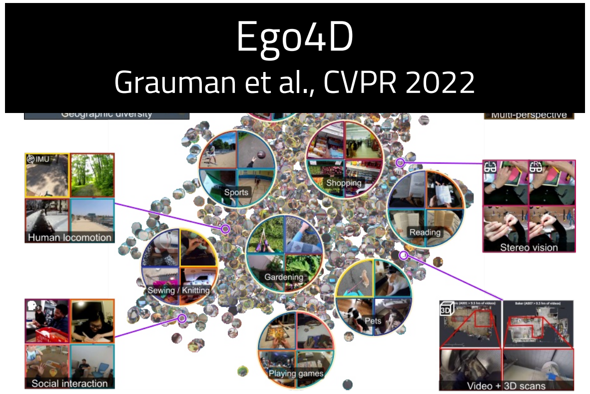
SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, Changan Chen et al., NeurIPS 2022

Generalizable simulation

We support arbitrary scene datasets.





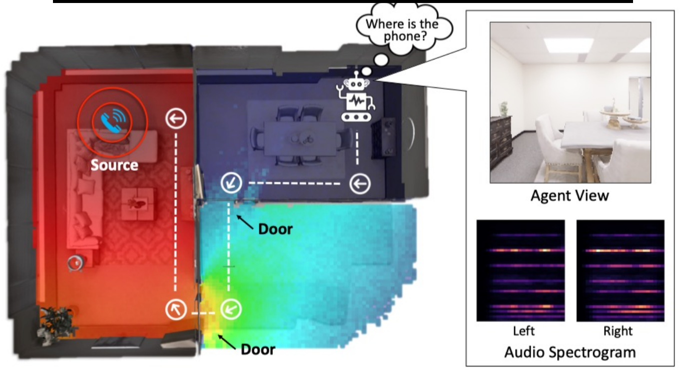


SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, Changan Chen et al., NeurIPS 2022

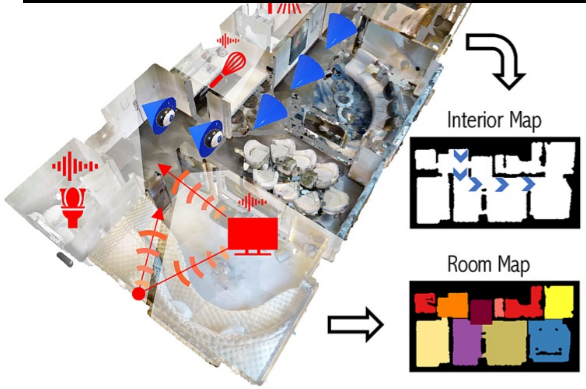
Visual-acoustic learning

We support an array of tasks!

Audio-Visual Navigation
Chen et al., ECCV 2020



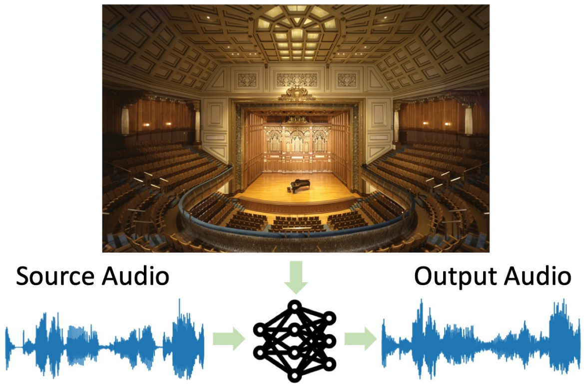
Audio-Visual Mapping
Purushwalkam et al., ICCV 2021



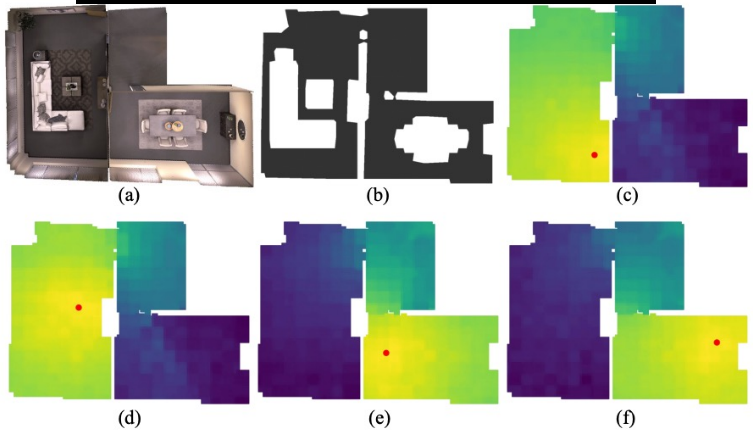
Audio-Visual Separation
Majumder et al., ICCV 2021

Visual Acoustic Matching
Chen et al.,CVPR 2022

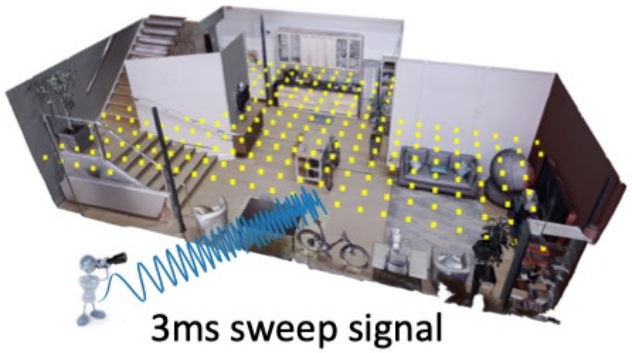
Target Space



Neural Acoustic Rendering
Luo et al., NeurIPS 2022



Echolocation Learning
Gao et al., ECCV 2020



SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, Changan Chen et al., NeurIPS 2022

Learning acoustics from vision

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

Augmented reality

Film dubbing

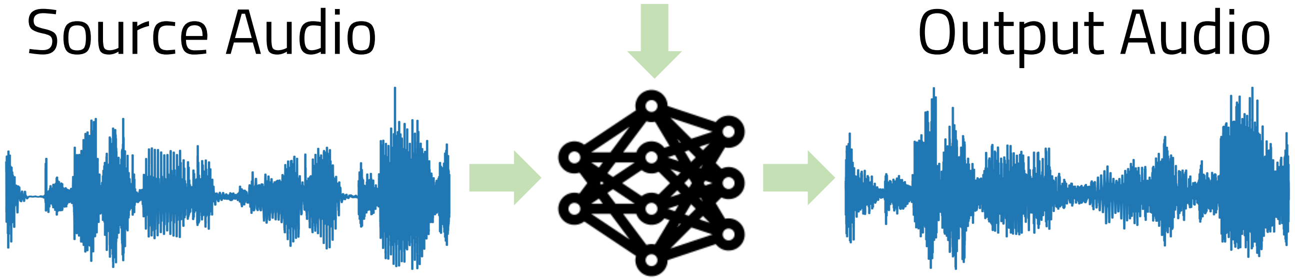
The visual acoustic matching task

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

Target Space

Source Audio

Output Audio

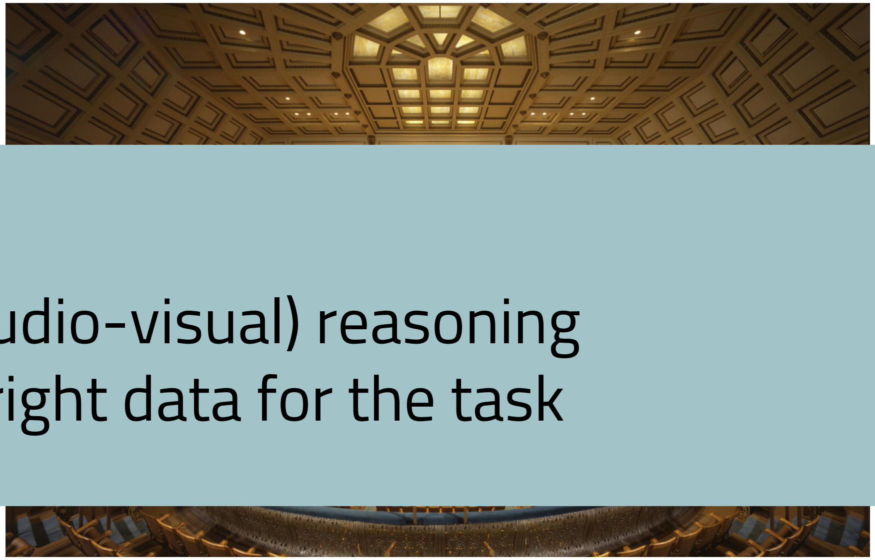


Visual Acoustic Matching, Changan Chen et al., CVPR 2022

The visual acoustic matching task

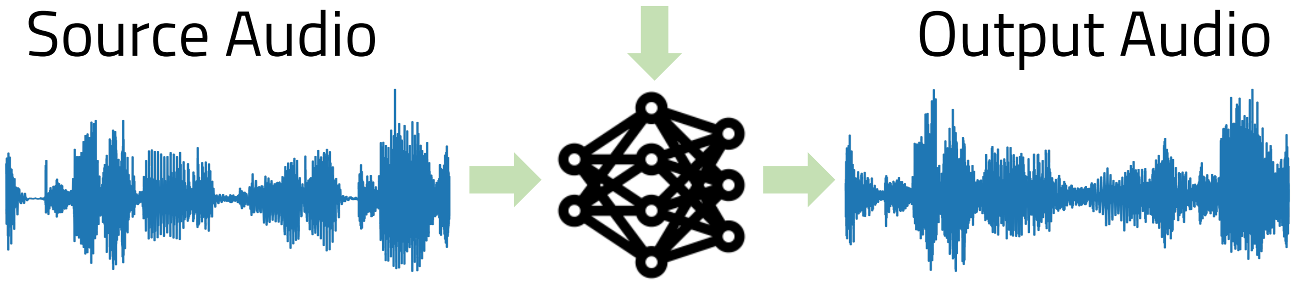
We propose to transform the sound recorded in one space to another depicted in the target visual scene.

Target Space



Main challenges:

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

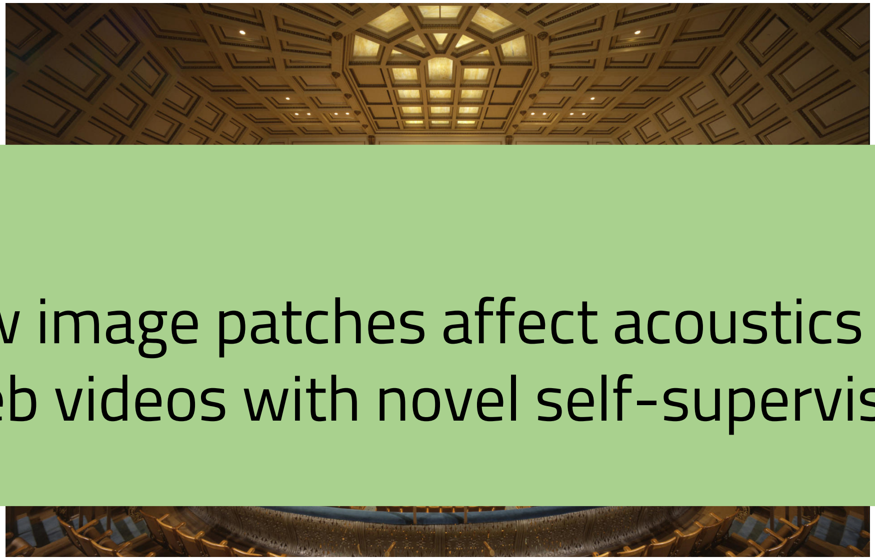


Visual Acoustic Matching, Changan Chen et al., CVPR 2022

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

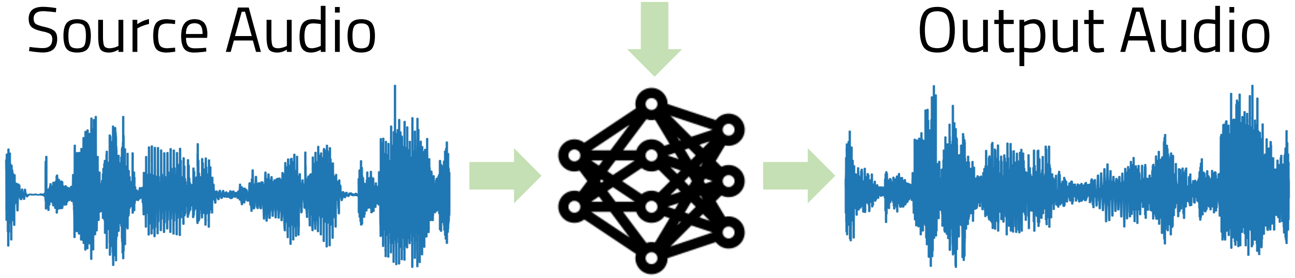


Key ideas:

1. Reasoning how image patches affect acoustics with attention.
2. Leveraging Web videos with novel self-supervision for learning.

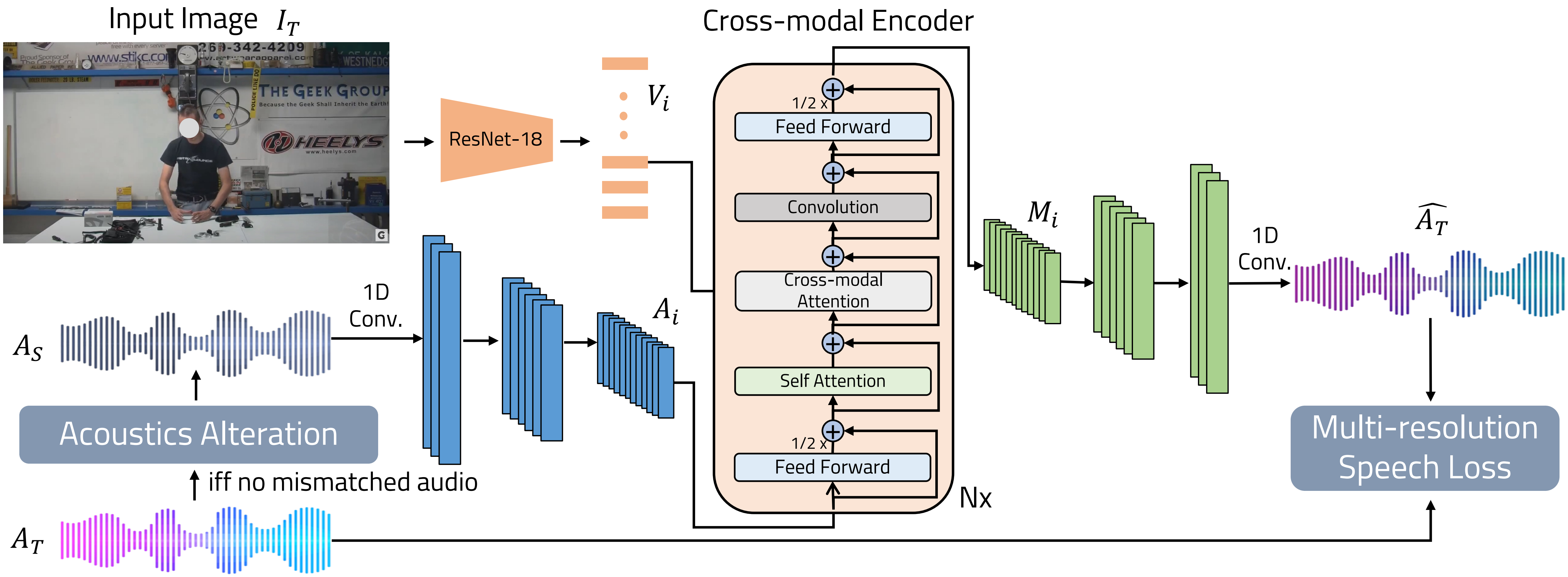
Source Audio

Output Audio



Visual Acoustic Matching, Changan Chen et al., CVPR 2022

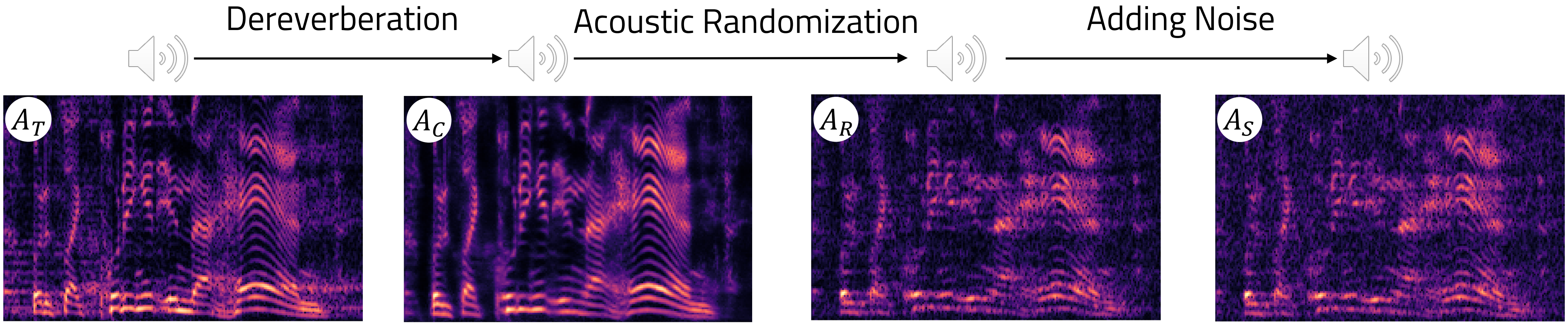
Audio-Visual Transformer for Audio Generation (AViTAR)



Visual Acoustic Matching, Changan Chen et al., CVPR 2022

Acoustics alteration strategy

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



Visual Acoustic Matching, Changan Chen et al., 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

Visual Acoustic Matching, Changan Chen et al., 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

Visual Acoustic Matching, Changan Chen et al., CVPR 2022

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**

Visual Acoustic Matching, Changan Chen et al., CVPR 2022

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.

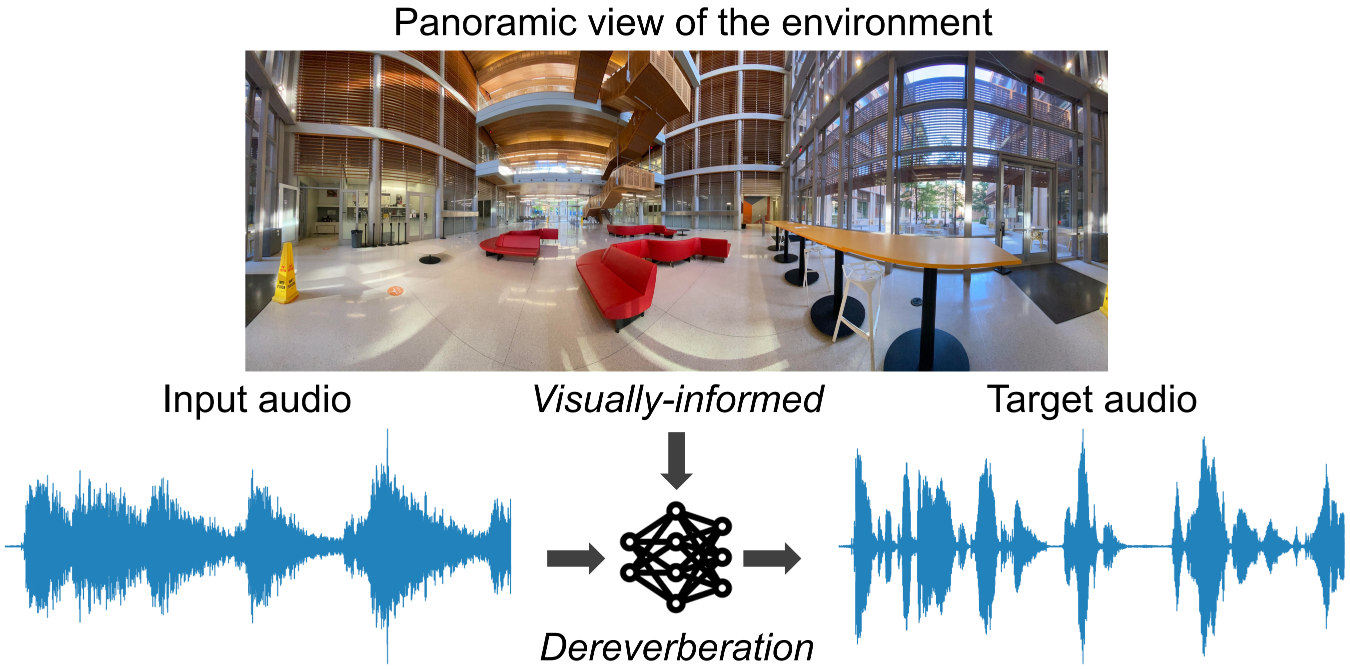
Visual Acoustic Matching, Changan Chen et al., CVPR 2022

Learning Audio-Visual Dereverberation

Changan Chen[1,2], Wei Sun[1], David Harwath[1], Kristen Grauman[1,2]

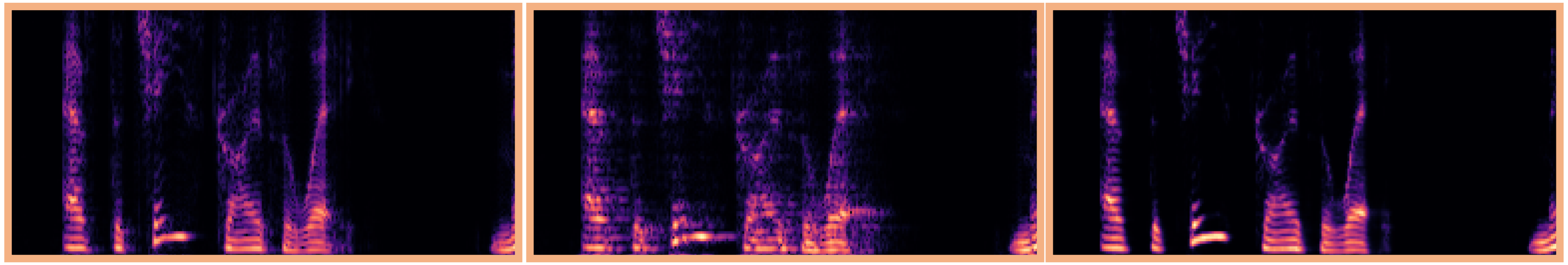
UT Austin[1], Meta AI[2]

ICASSP 2023

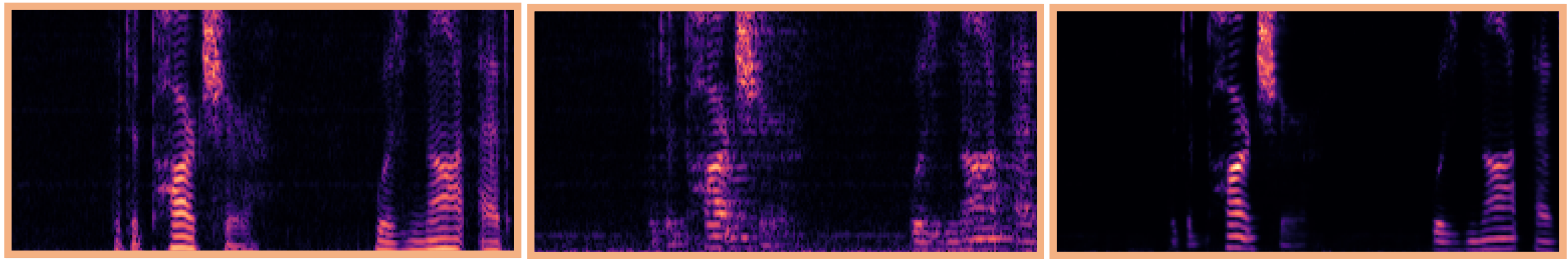


Qualitative examples

Panorama RGB	Clean (GT)	Reverberant	De-reverberated by VIDA



Long corridor, distance speaker



Classroom, close speaker

Novel-view Acoustic Synthesis

Changan Chen[1,3], Alexander Richard[2], Roman Shapovalov[3], Vamsi Krishna Ithapu[2], Natalia Neverova[3], Kristen Grauman[1,3], Andrea Vedaldi[3]

University of Texas at Austin[1], Reality Labs Research at Meta[2], FAIR, Meta AI[3]

CVPR 2023

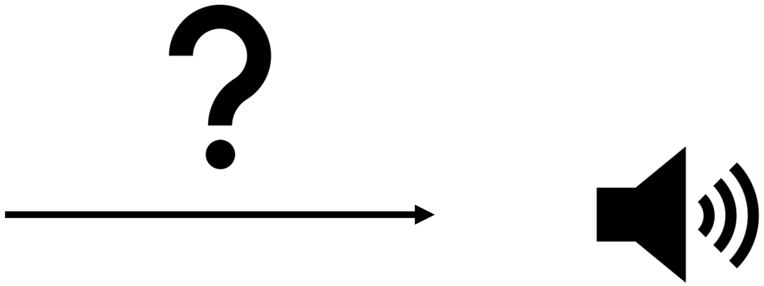
Replaying videos to relive a moment

- Many of our important life moments are recorded in videos
- Videos are however passively collected from one viewpoint
- Recreating the moment in 3D is important for immersive AR/VR applications

Novel-view Acoustic Synthesis, Changan Chen et al., CVPR 2023

Novel-view synthesis

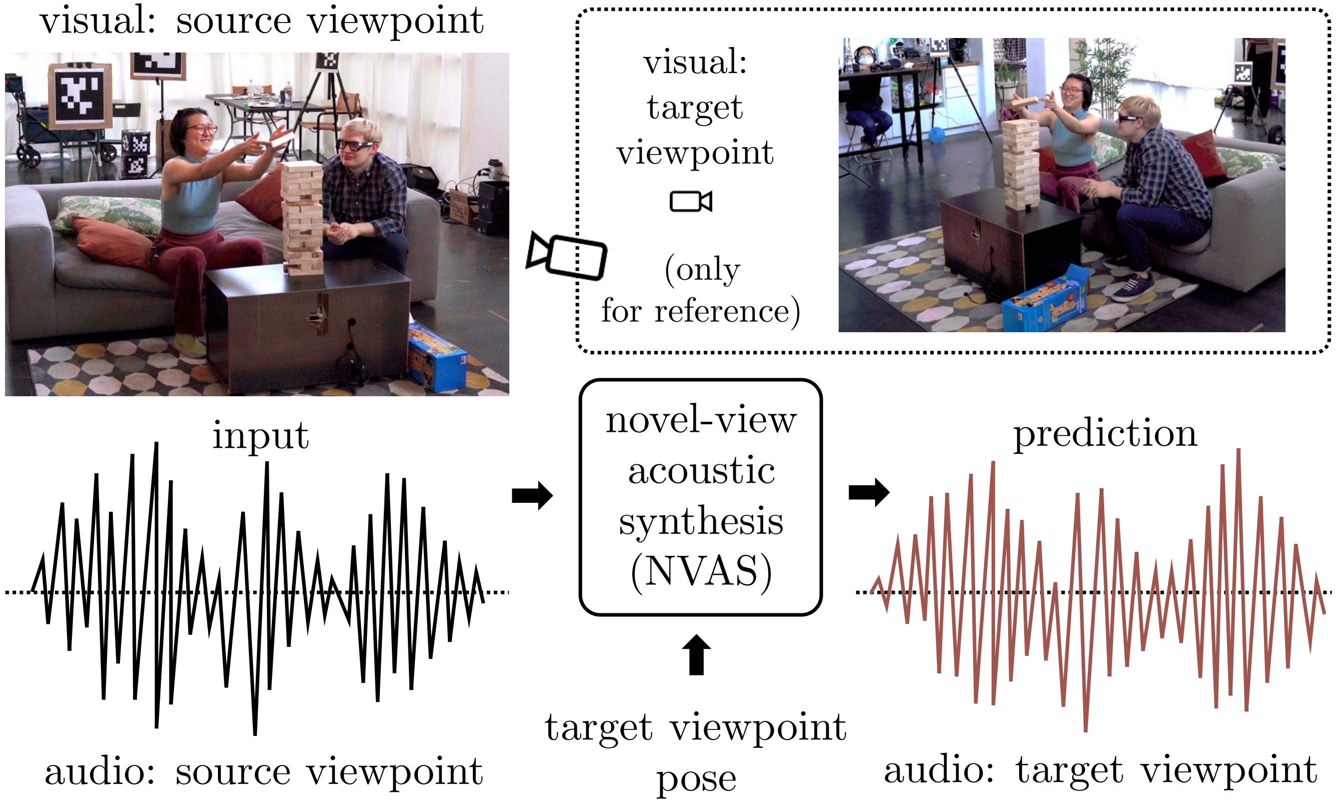
- Novel-view synthesis (NVS) receives lots of attention lately
- However, it is by far vision-only and does not handle sound



Novel-view Acoustic Synthesis, Changan Chen et al., CVPR 2023

Novel-view Acoustic Synthesis

We propose the novel-view acoustic synthesis task:



Novel-view Acoustic Synthesis, Changan Chen et al., CVPR 2023

Difference between NVS & NVAS

NVS:

- 3D scenes change limitedly during th
- Camera directio
- Freque providin triangulation and segmentation

NVAS:

- Sound changes substantially over namic
- e at best
- s a wide
- Sounds are often mixed together

1. Lack of supporting dataset and benchmark

2. Lack of existing model that is capable of NVAS

Novel-view Acoustic Synthesis, Changan Chen et al., CVPR 2023

Replay-NVAS dataset

- 46 scenarios captured from 8 different viewpoints
- Each viewpoint is equipped with a DSLR camera and binaural mic
- 2-4 actors act on a certain topic, e.g., chatting, doing yoga, etc.
- Each actor has a near-range mic to record their voice
- In total 37 hours of video data

Novel-view Acoustic Synthesis, Changan Chen et al., CVPR 2023

Replay-NVAS example

Novel-view Acoustic Synthesis, Changan Chen et al., CVPR 2023

SoundSpaces-NVAS dataset

- Constructed based on SoundSpaces 2.0[1] audio-visual simulator
- Renders acoustic effects such as direct sound, reverberation, transmission, and diffraction
- Use LibriSpeech[2] (audio book) as the source audio
- 1,000 speakers, 120 3D scenes, 200K viewpoints and 1.3K hours of audio-visual data

[1]SoundSpaces 2.0: A Simulation Platform for Visual-Acoustic Learning, Chen et al., NeurIPS 2022
[2]Librispeech: An ASR corpus based on public domain audio books, Chen et al., ICASSP 2015

SoundSpaces-NVAS examples

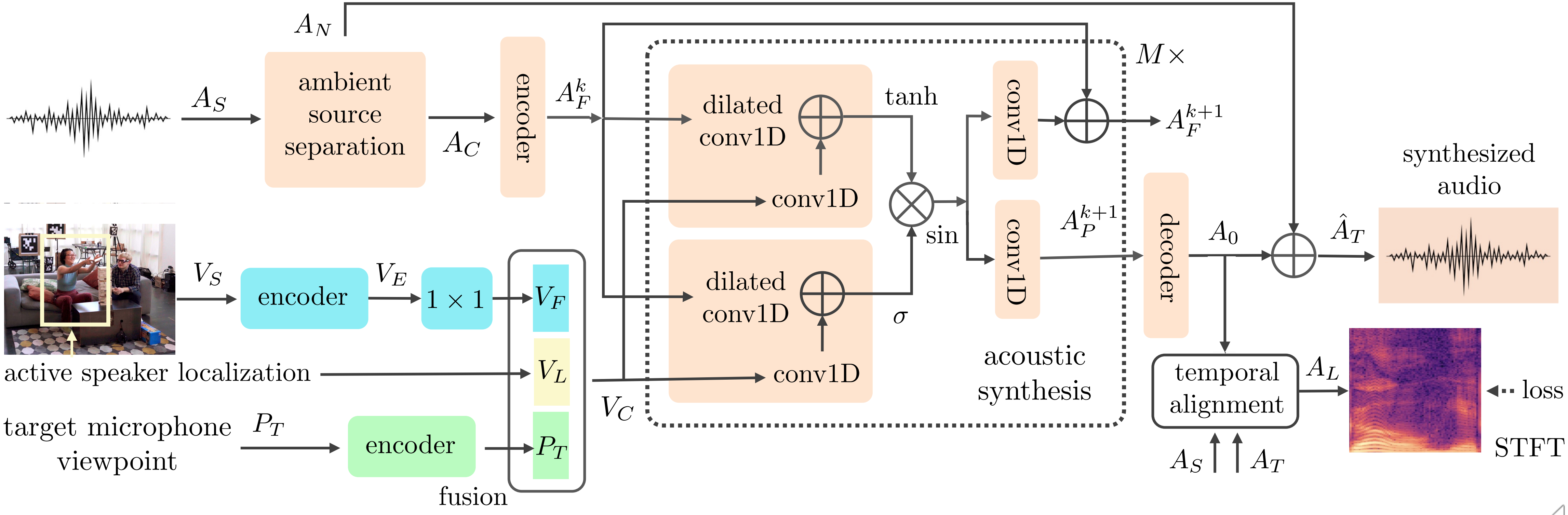
Here we show the near-range audio (clean) of the female speaker and then the audio-visual observations at different viewpoints.

Near-range | Viewpoint 1 | Viewpoint 2 | Viewpoint 3 | Viewpoint 4

Novel-view Acoustic Synthesis, Changan Chen et al., CVPR 2023

Visually Guided Acoustic Synthesis (ViGAS)

Learn an implicit neural transfer function that reasons the sound source location, acoustics of the space and the target pose in 3D to synthesize the target sound.



Novel-view Acoustic Synthesis, Changan Chen et al., CVPR 2023

Evaluation

- Closeness to the ground truth
  - Magnitude spectrogram distance (Mag)
- Correctness of the spatial sound
  - Left-right energy ratio error (LRE)
- Correctness of the acoustics properties
  - Reverberation time decaying by 60dB error (RTE)

Single-environment: train and test on the same environment

Novel-environment: train and test on disjoint environments

Baselines

- Input audio
  - Copies input to the prediction
- TF estimator[1] + nearest neighbor
  - Estimates transfer functions indexed by ground truth location during training and retrieve the nearest nighbor during test
- Digital signal processing (DSP)[2]
  - Estimates the distance, azimuth and elevation of the sound source, then apply an inverse head-related transfer function (HRTF)
- Visual acoustic matching (VAM)[3]
  - A recent audio-visual generative model for matching acoustics with images

[1]Extrapolation, interpolation, and smoothing of stationary time series. Norbert Wiener. Report of the Services 19, 1942
[2]Introduction to head-related transfer functions (hrtfs): representations of hrtfs in time, frequency, and space. Cheng et al., AES 2001
[3]Visual Acoustic Matching, Chen et al., CVPR 2022

Results

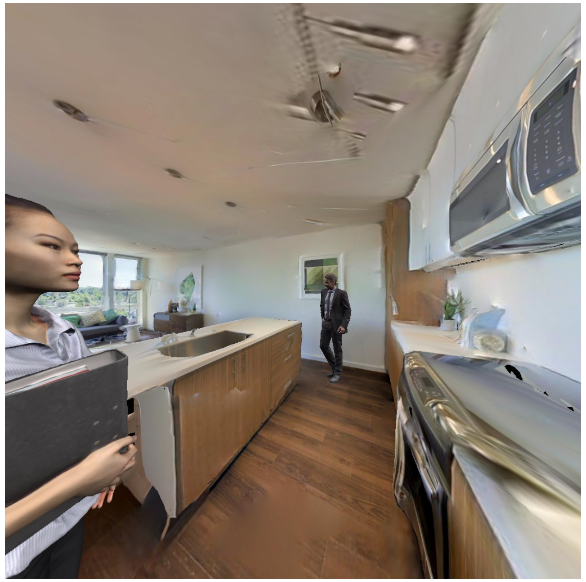
- Our model outperforms all baselines including audio-only ablation
- Generalizing to the acoustics of novel environments is challenging

	SoundSpaces-NVAS						**Replay-NVAS**		
	Single Environment			*Novel Environment*			*Single Environment*		
	Mag	LRE	RTE	Mag	LRE	RTE	Mag	LRE	RTE
Input audio	0.225	1.473	0.032	0.216	1.408	0.039	0.159	1.477	0.046
TF Estimator [1]	0.359	2.596	0.059	0.440	3.261	0.092	0.327	2.861	0.147
DSP [2]	0.302	3.644	0.044	0.300	3.689	0.047	0.463	1.300	0.067
VAM [3]	0.220	1.198	0.041	0.235	1.131	0.051	0.161	0.924	0.070
ViGAS w/o visual	0.173	0.973	0.031	0.181	1.007	0.036	0.146	0.877	**0.046**
ViGAS	**0.159**	**0.782**	**0.029**	**0.175**	**0.971**	**0.034**	**0.142**	**0.716**	0.048

[1] Extrapolation, interpolation, and smoothing of stationary time series. Norbert Wiener. Report of the Services 19, 1942
[2] Introduction to head-related transfer functions (hrtfs): representations of hrtfs in time, frequency, and space. Cheng et al., AES 2001
[3] Visual Acoustic Matching, Chen et al., CVPR 2022

Qualitative examples on SoundSpaces-NVAS

Here we compare ViGAS with three other baseline methods (all audio clips are 2.5 seconds).



Source

Target

ViGAS

DSP

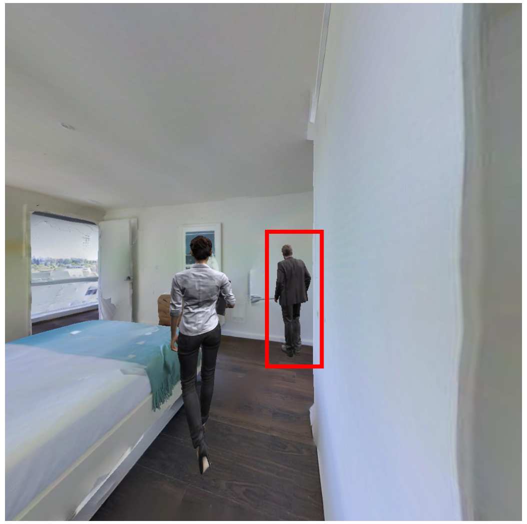
TF Estimator

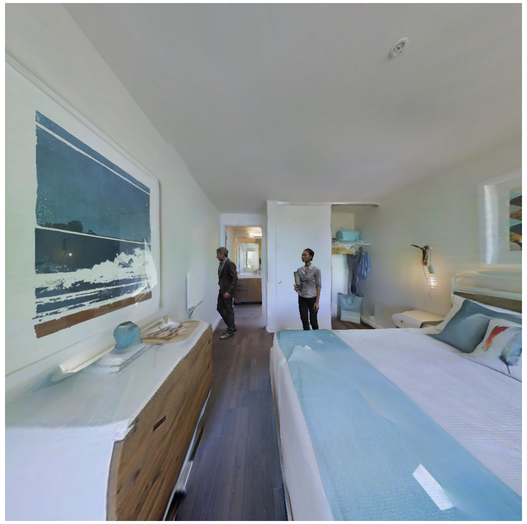
VAM

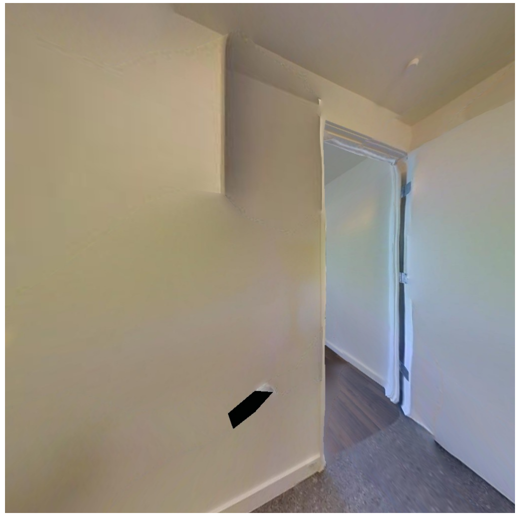
Novel-view Acoustic Synthesis, Changan Chen et al., CVPR 2023

ViGAS examples for SoundSpaces-NVAS

Here we show that for one source viewpoint, our model predicts the audio for four different viewpoints.





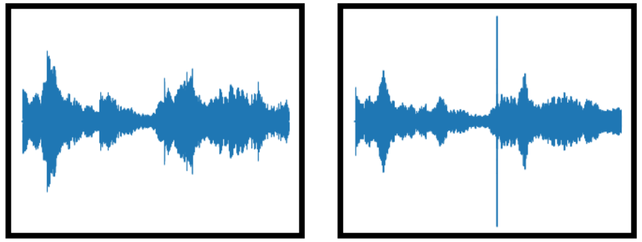


Target

Source

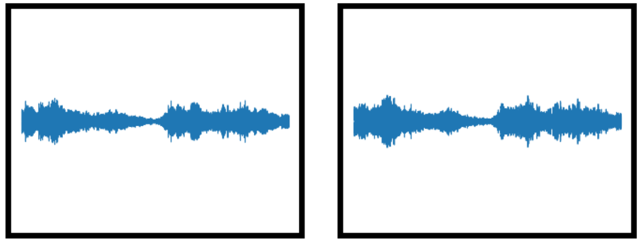
Prediction

Replay-NVAS example 1



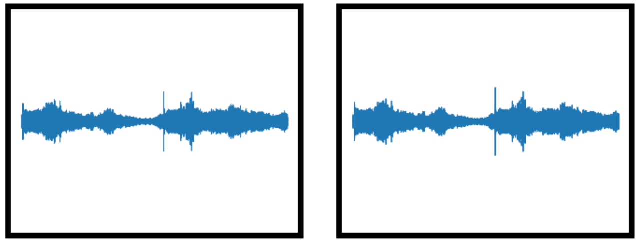
Left channel Right channel

Source



Left channel Right channel

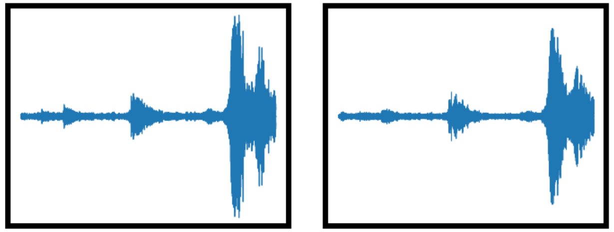
Target



Left channel Right channel

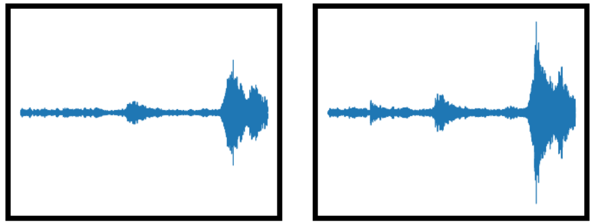
Ours

Replay-NVAS example 2



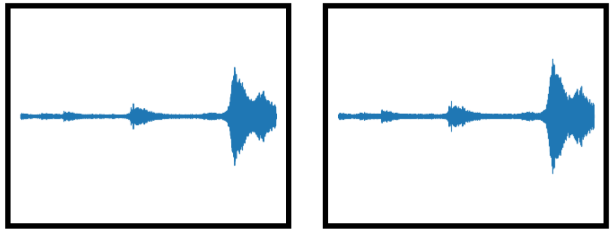
Left channel Right channel

Source



Left channel Right channel

Target



Left channel Right channel

Ours

Summary

- Studying the 4D correspondence between sight and sound
- Data
    - Build the first audio-visual simulation that's configurable and generalizable
    - Devise self-supervised objectives to leverage in-the-wild web data
    - Collect a large-scale multi-view audio-visual dataset
- Tasks and benchmarks
    - Embodied agents, e.g., audio-visual navigation, active separation
    - Matching acoustics with a reference image
    - Inform the dereverberation process with a panoramic snapshot of the env
    - Novel-view acoustic synthesis