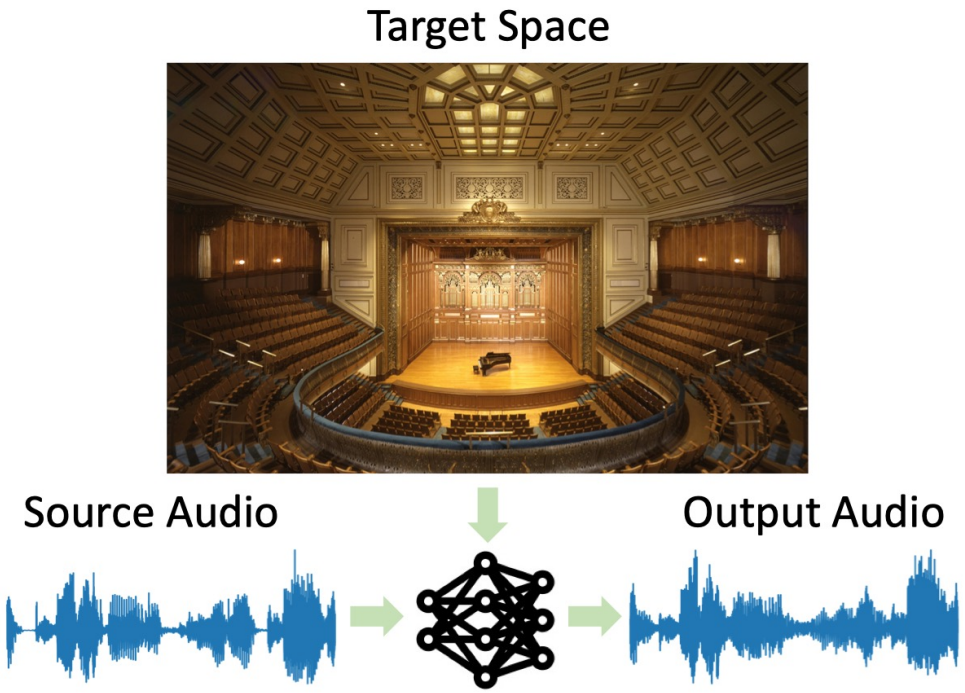
Visual Acoustic Matching

Changan Chen[1,4], Ruohan Gao[2], Paul Calamia[3], Kristen Grauman[1,4]

[1]UT Austin, [2]Stanford, [3]Reality Labs at Meta, [4]Meta AI



The University of Texas at Austin

Meta AI

1 drum kit 5 different spaces

Source: Shred Shed Studio

Acoustics is everywhere

Whenever we hear sound, it is shaped by the space
- Sound wave propagates in space, reflects off, gets absorbed or transmits through surfaces
- Geometry, materials, source & listener locations

Lots of reverb

Less reverb

Chen et al., Visual Acoustic Matching, CVPR 2022

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

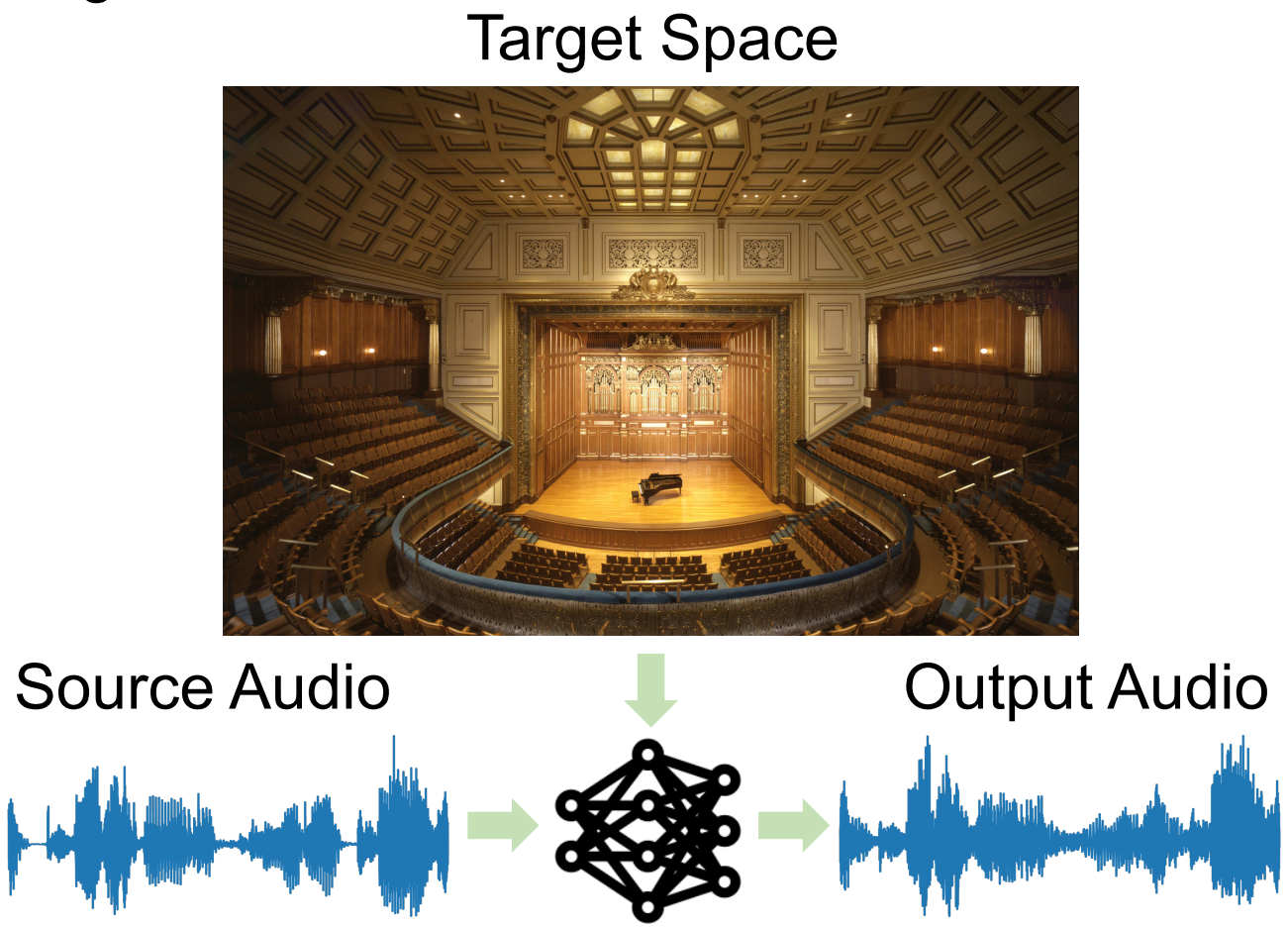
Film dubbing

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.

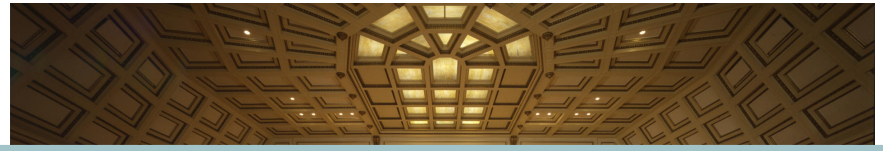


Chen et al., Visual Acoustic Matching, CVPR 2022

The visual acoustic matching task

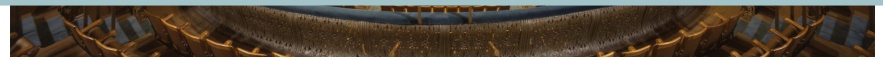
We propose to transform the sound recorded in one space to another depicted in the target visual scene.

Target Space

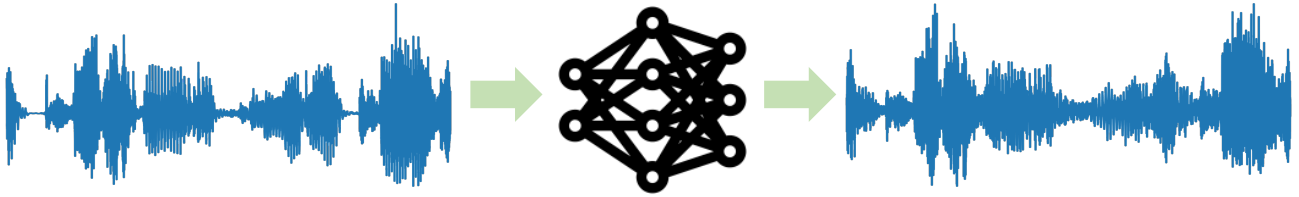


Main challenges:

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



Source Audio



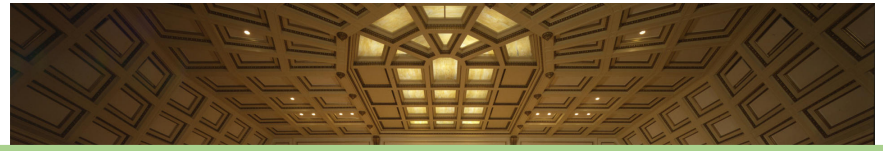
Output Audio

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.

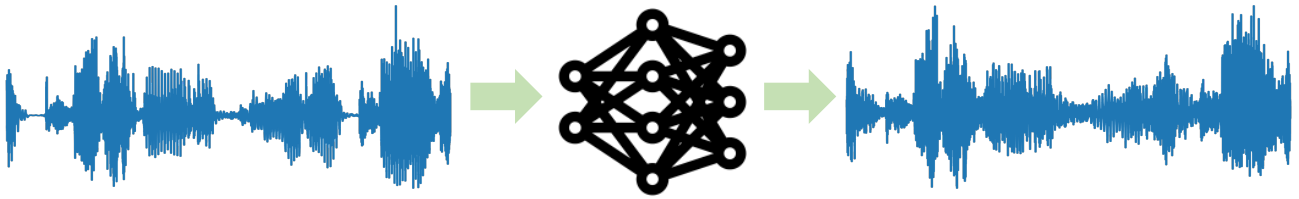
Target Space



Key ideas:

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

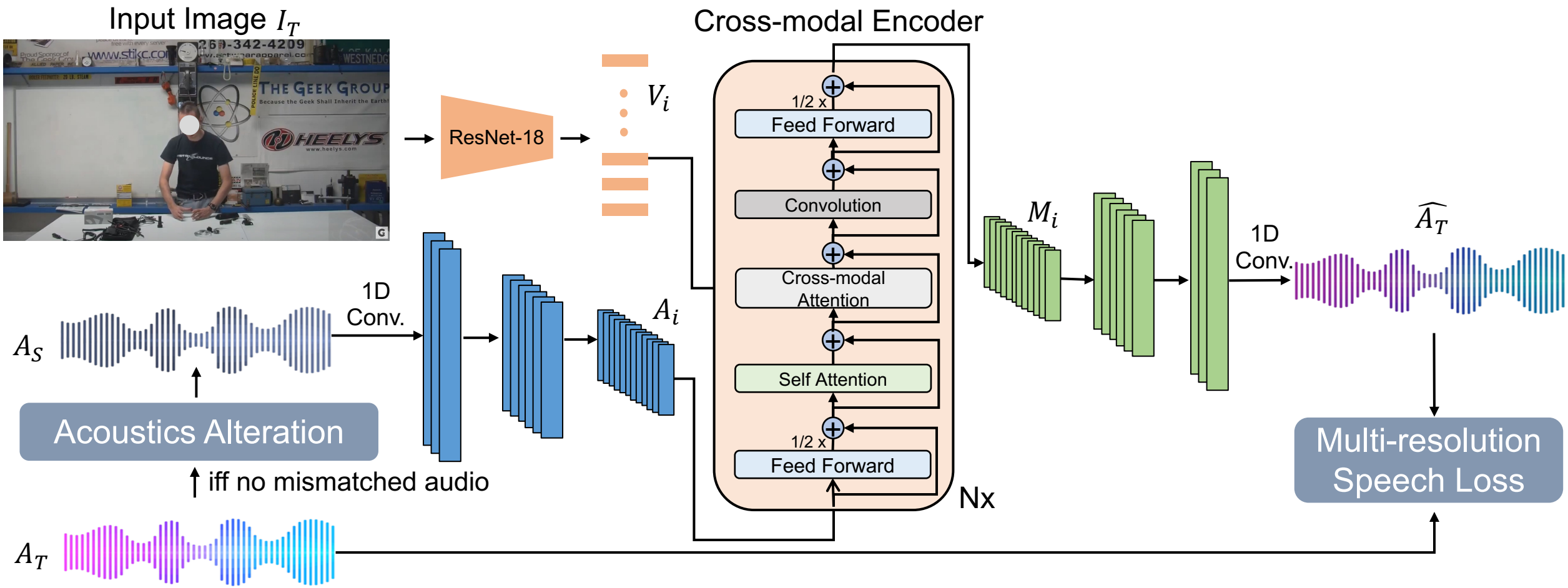
Source Audio



Output Audio

Chen et al., Visual Acoustic Matching, CVPR 2022

Audio-Visual Transformer for Audio Generation (AViTAR)



Acoustics alteration: creates acoustic mismatch by dereverberation and acoustic randomization for self-supervision

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

Experiment results

- Closeness to ground truth, correctness of acoustics and speech quality
- Strongly outperforms traditional and heavily supervised approaches

	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 [1]	1.338	0.044	0.312	-	-	-	-	-	-	-
Image2Reverb [2]	2.538	0.293	0.508	2.318	0.317	0.518	-	-	-	-
AV U-Net [3]	**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**

[1] More than 50 years of artificial reverberation, Valimaki et al., The 60th AES Conference on DREAMS
[2] Image2Reverb: Cross-Modal Reverb Impulse Response Synthesis, Singh et al., ICCV 2021
[3] 2.5D Visual Sound, Gao et al., CVPR 2019

Matching different environments on AVSpeech

	Office	Garage	Auditorium
Input			
AViTAR			
Reverb time	0.34s	0.40s	0.58s

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

Augmented reality demo

Making a remote participant sound like he is speaking in the room with us in a virtual phone call.

Input

Output

Chen et al., Visual Acoustic Matching, CVPR 2022

Visual Acoustic Matching

Changan Chen[1,4], Ruohan Gao[2], Paul Calamia[3], Kristen Grauman[1,4]

[1]UT Austin, [2]Stanford, [3]Reality Labs at Meta, [4]Meta AI

Code and demo available at:
https://vision.cs.utexas.edu/projects/visual-acoustic-matching

TEXAS
The University of Texas at Austin

Meta AI