Vision, Audio and Depth

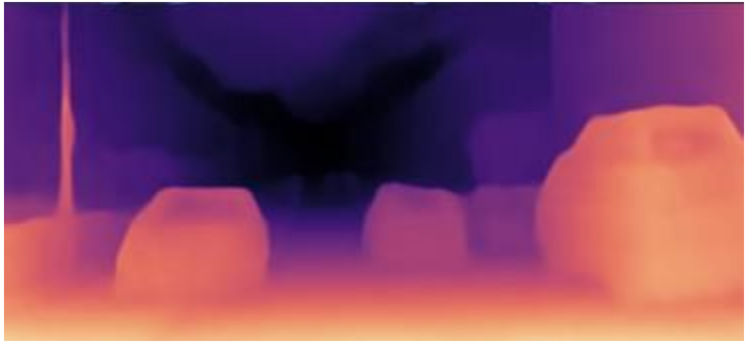
Kranti Kumar Parida

2 Papers

1. Structure from Silence - Estimating depth of the scene from ambient sound

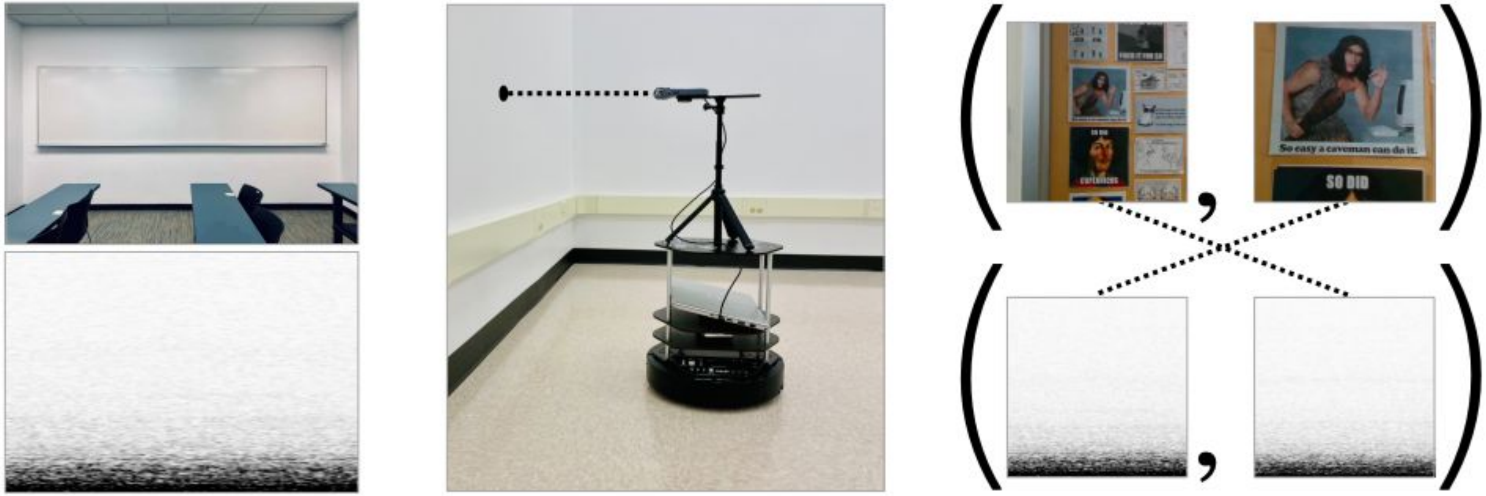
2. Audio-Visual Dereverberation - Enhancing Sound using visual/depth information

Image and Depth



Structure from Silence: Learning Scene Structure from Ambient Sound

Ziyang Chen*, Xixi Hu*, Andrew Owens
University of Michigan
https://ificl.github.io/structure-from-silence



(a) *Quiet Campus* dataset (b) Depth estimation (c) Multimodal self-supervision

CoRL 2021
https://arxiv.org/pdf/2111.05846.pdf, https://ificl.github.io/structure-from-silence/

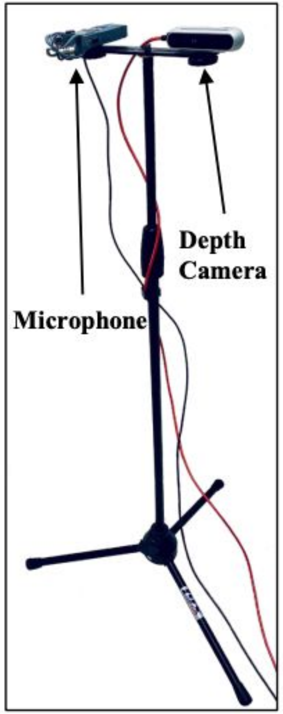
Introduction

- Does ambient sound convey information about 3D structure?
- Humans capable of estimating scene structure from subtle ambient sound cues
- Estimate Depth from Sound
- Not depth but a simplified version

Dataset

- Data Collected using audio and RGB-D camera
- Indoor ambient audio recordings
- No other sound producing objects
- Both Motion and Static
- Camera Pointing to wall/flat surfaces

Figure 2: **The *Quiet Campus* Dataset**. We collected a dataset of paired audio and RGB-D recordings from a variety of quiet indoor scenes. We show selected images from the *static* and *motion* subsets, which contain stationary and moving microphones respectively. Please refer to the project webpage for audio-visual examples.

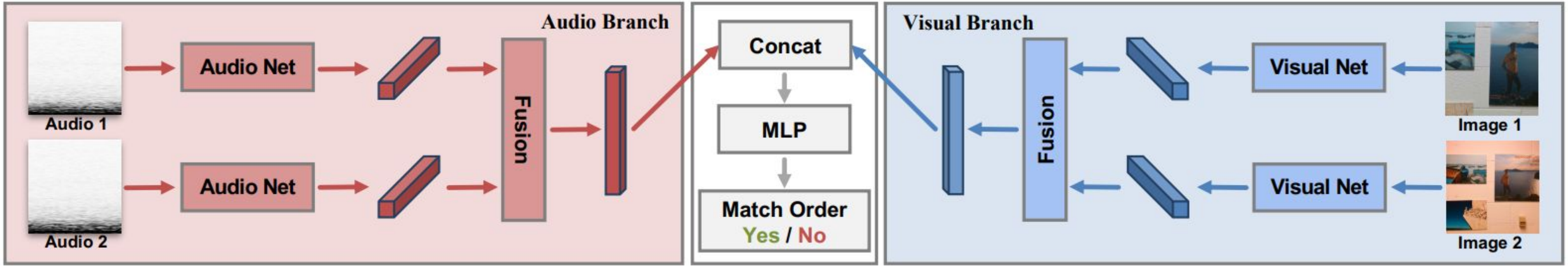


Tasks

1. Depth Estimation
    a. Obstacle Detection : Whether microphone is within a small distance of wall (0.5 m)
    b. Relative Depth Order - Given two audio clip predict which one is closer to wall
    c. Relative Depth Estimation - Given two audio clips, predict the difference of distance between them.
    d. Absolute Depth estimation - Given an audio clip, directly predict the distance to the wall.

- Depth : center crop 320x240 and average the depth values

Self-Supervised Learning



Input Representation and Network

- Audio Input: 0.96s in the form of log-mel spectrogram
- Audio Network: VGGish network, final layer replaced either for classification or regression
- Visual Network: ResNet-18 with 224x224 image

Results

Table 1: **Obstacle detection and relative depth order.** We evaluate our model's ability to determine whether a microphone is within 0.5 meters of a wall and identify which sound has a smaller distance to the wall. *Pre* refers to pretraining.

			Obstacle detection		Relative order	
Model	Pre.	Task	AP(%)	Acc(%)	AP(%)	Acc(%)
Audio		static	68.3 (±1.3)	60.0 (±0.9)	85.5 (±1.0)	77.2 (±0.8)
Image		static	99.2 (±0.2)	95.5 (±0.5)	94.6 (±0.5)	86.4 (±0.7)
Image	✓	static	99.5 (±0.1)	98.4 (±0.4)	97.7 (±0.2)	92.1 (±0.5)
Chance		static	46.4 (±1.4)	50.0 (±1.0)	47.2 (±1.3)	50.0 (±1.0)
Audio		motion	65.6 (±1.4)	64.5 (±0.9)	87.1 (±1.0)	81.3 (±0.8)
Image		motion	73.4 (±1.2)	68.2 (±1.0)	87.9 (±0.9)	81.2 (±0.8)
Image	✓	motion	88.6 (±0.7)	78.5 (±0.8)	97.1 (±0.3)	90.6 (±0.6)
Chance		motion	50.4 (±1.4)	50.0 (±1.0)	50.5 (±1.4)	50.0 (±1.0)

Results

Table 2: **Relative depth ratio.** We evaluate our model's ability of predicting relative depth ratio from two ambient sounds, for the *motion* recordings.

	Regression			Regression-by-Classification		
Model	MAE ↓	Med. ↓	R^2 ↑	Top-1 ↑	Top-5 ↑	Avg. Dist ↓
Audio	0.55 (±.01)	0.44 (±.01)	0.48 (±.02)	22.8 (±0.8)	80.7 (±0.7)	1.66 (±.03)
Image	0.54 (±.01)	0.42 (±.01)	0.49 (±.02)	26.6 (±0.8)	83.6 (±0.7)	1.47 (±.02)
Image (Pre.)	0.39 (±.01)	0.29 (±.01)	0.72 (±.01)	34.2 (±0.9)	90.5 (±0.6)	1.15 (±.02)
Chance	0.89 (±.01)	0.79 (±.01)	0.00 (±.00)	9.45 (±0.6)	52.6 (±1.0)	2.79 (±.04)
No input	0.82 (±.01)	0.75 (±.01)	0.00 (±.00)	10.7 (±0.6)	51.6 (±1.0)	4.50 (±.05)

Table 3: **Absolute depth estimation.** We evaluate our model's ability of predicting absolute distance to the wall for the *motion* recordings.

		Regression			Regression-by-Classification		
	Model	MAE ↓	Med. ↓	R^2 ↑	Top-1 ↑	Top-5 ↑	Avg. Dist ↓
Single	Audio	0.28 (±.00)	0.25 (±.01)	-0.34 (±.03)	30.8 (±0.9)	88.3 (±0.6)	1.11 (±.02)
	Image	0.31 (±.00)	0.27 (±.01)	-0.67 (±.07)	35.6 (±0.9)	95.9 (±0.4)	1.05 (±.02)
	Image (Pre.)	0.26 (±.00)	0.21 (±.01)	-0.24 (±.04)	50.8 (±1.0)	99.2 (±0.2)	0.62 (±.01)
	No input	0.28 (±.00)	0.27 (±.01)	-0.19 (±.02)	24.3 (±0.8)	88.3 (±0.6)	1.07 (±.01)
Conditional	Audio	0.21 (±.00)	0.17 (±.00)	0.19 (±.02)	36.9 (±1.0)	90.0 (±.06)	1.17 (±.02)
	Image	0.22 (±.00)	0.18 (±.00)	0.12 (±.02)	38.2 (±0.9)	95.5 (±0.4)	0.93 (±.02)
	Image (Pre.)	0.18 (±.00)	0.14 (±.00)	0.39 (±.02)	51.7 (±1.0)	99.8 (±0.1)	0.59 (±.01)
	No input	0.25 (±.00)	0.23 (±.00)	0.01 (±.01)	26.4 (±0.8)	95.9 (±0.4)	1.43 (±.03)
	Chance	0.78 (±.01)	0.84 (±.01)	-3.38 (±0.23)	23.3 (±1.2)	56.9 (±0.9)	2.83 (±.02)

Results (Self-Supervised Learning)

- Given audio and image pair, predict if they are matched or mismatched
- Evaluate on depth estimation task

Table 4: **Linear probing experiments.** We evaluate our self-supervised feature set for **obstacle detection** and **relative depth order**, for the *motion* recordings. Here, *Audio* means taking audio only as inputs. *Visual* means taking images only as inputs. *Both* means taking both audio and image as inputs.

			Obstacle detection		Relative order	
	Model	Pre.	AP(%)	Acc(%)	AP(%)	Acc(%)
Audio	Scratch		61.9 (±1.5)	60.3 (±0.9)	78.0 (±1.4)	73.1 (±0.9)
	VGGish [75]		58.2 (±1.3)	56.0 (±1.0)	61.1 (±1.4)	61.2 (±1.0)
	AV-Sync		**69.1** (±1.4)	**64.0** (±0.9)	80.2 (±1.3)	74.1 (±0.8)
	AV-Order		63.4 (±1.4)	61.5 (±0.9)	**84.2** (±1.2)	**79.4** (±0.7)
	VGGish [75]	✓	59.0 (±1.5)	56.7 (±1.0)	67.7 (±1.4)	64.5 (±0.9)
	AV-Sync	✓	**65.3** (±1.4)	**62.8** (±0.9)	82.1 (±1.2)	76.4 (±0.8)
	AV-Order	✓	62.8 (±1.5)	64.5 (±0.9)	**85.5** (±1.1)	**80.7** (±0.8)
Visual	Scratch		70.1 (±1.3)	64.0 (±0.9)	79.7 (±1.1)	71.5 (±0.9)
	AV-Sync		**77.1** (±1.1)	**69.2** (±0.9)	85.3 (±0.9)	76.1 (±0.8)
	AV-Order		76.8 (±1.1)	68.8 (±0.9)	**87.4** (±0.9)	**79.1** (±0.8)
	ImageNet [77, 76]	✓	80.4 (±1.2)	74.5 (±0.8)	94.0 (±0.5)	85.8 (±0.7)
	AV-Sync	✓	**89.0** (±0.8)	75.6 (±0.8)	92.8 (±0.6)	85.4 (±0.7)
	AV-Order	✓	86.5 (±1.1)	**76.3** (±0.8)	**95.8** (±0.4)	**88.9** (±0.6)
Both	AV-Order		77.1 (±1.1)	69.1 (±0.9)	89.0 (±0.8)	80.8 (±0.8)
	AV-Order	✓	88.1 (±0.9)	76.9 (±0.8)	95.8 (±0.4)	88.9 (±0.6)

Table 5: **Linear probing experiments.** We evaluate our learned representation for **relative depth ratio** for the *motion* recordings.

	Model	Pre.	Top-1 (%) ↑	Top-5 (%) ↑	Avg. Dist ↓
Audio	Scratch		19.2 (±0.8)	72.8 (±0.8)	2.33 (±0.04)
	VGGish [75]		14.4 (±0.7)	53.9 (±1.0)	3.78 (±0.06)
	AV-Sync.		19.2 (±0.7)	72.7 (±0.8)	2.07 (±0.03)
	AV-Order		**22.2** (±0.9)	**79.6** (±0.8)	**1.86** (±0.03)
	VGGish [75]	✓	15.6 (±0.7)	54.0 (±1.0)	3.59 (±0.05)
	AV-Sync.	✓	20.7 (±0.8)	75.1 (±0.9)	1.99 (±0.03)
	AV-Order	✓	**23.6** (±0.9)	**80.5** (±0.7)	**1.75** (±0.03)
Visual	Scratch		18.5 (±0.8)	70.8 (±0.8)	2.66 (±0.05)
	AV-Sync.		22.2 (±0.8)	76.8 (±0.8)	1.85 (±0.03)
	AV-Order		**24.7** (±0.8)	**80.2** (±0.8)	**1.71** (±0.03)
	ImageNet [77, 76]	✓	27.4 (±0.9)	87.1 (±0.7)	1.60 (±0.03)
	AV-Sync.	✓	27.5 (±0.8)	85.2 (±0.7)	1.53 (±0.03)
	AV-Order	✓	**28.9** (±0.9)	**88.6** (±0.6)	**1.40** (±0.03)
Both	AV-Order		23.8 (±0.8)	81.5 (±0.7)	1.59 (±0.03)
	AV-Order	✓	30.0 (±0.9)	89.3 (±0.6)	1.31 (±0.03)

Task

3. Audio-Visual Robotic Navigation

- Detect if there is a wall near the left/right and move accordingly

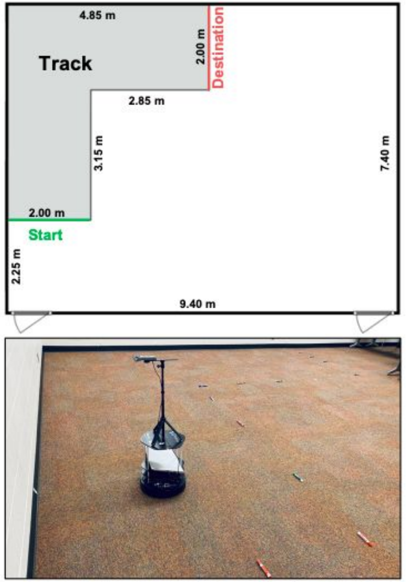


Figure 7: Classroom floor plan and track setting.

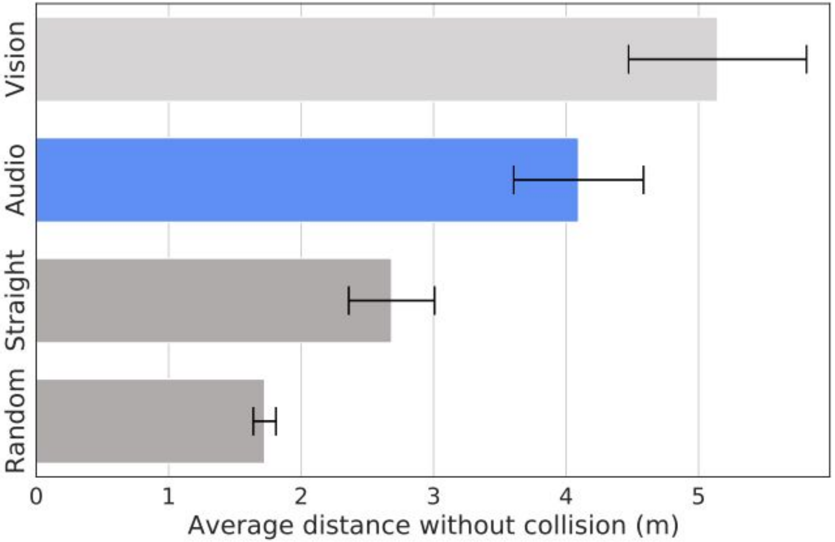


Figure 9: Robot navigation results.

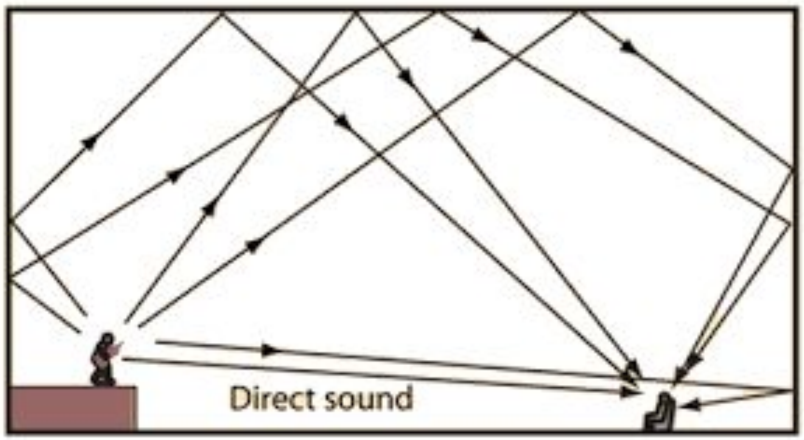
Learning Audio-Visual Dereverberation

Changan Chen[1,2] **Wei Sun**[1] **David Harwath**[1] **Kristen Grauman**[1,2]
[1]UT Austin [2]Facebook AI Research

ArXiv 2021
https://arxiv.org/pdf/2106.07732.pdf, https://vision.cs.utexas.edu/projects/learning-audio-visual-dereverberation/

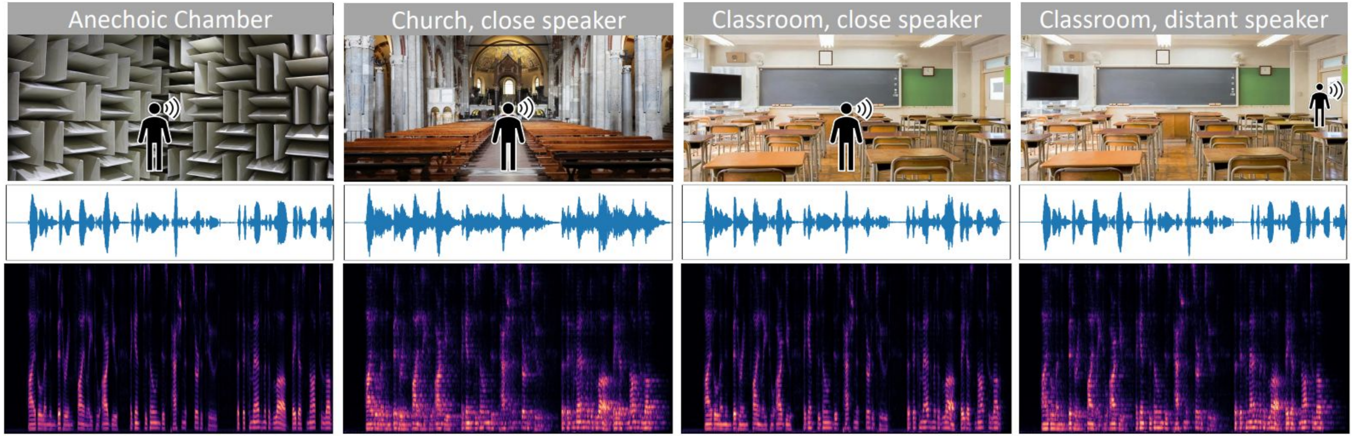
Audio Reverberation

- Multiple reflections from different objects and surfaces
- Alters original signal
- Degrades perceptual experience and ASR systems



Background

- Reverberation explained by Room Impulse Response (RIR)
- Function of room geometry, materials and speaker location



$$A_r(t) = A_s(t) * R(t)$$

Reverb Audio   Source Audio   RIR

Dereverberation Past Approaches

- Signal processing and statistical signals
- Neural Network based approach
- Rely completely on audio

Goal:

Given RGB image, depth Image, received (reverb) audio predict source audio

$$\hat{A}_s(t) = f_p([I_r, I_d, A_r(t)])$$

Dataset

- No existing dataset was available
- Both simulated and real data proposed

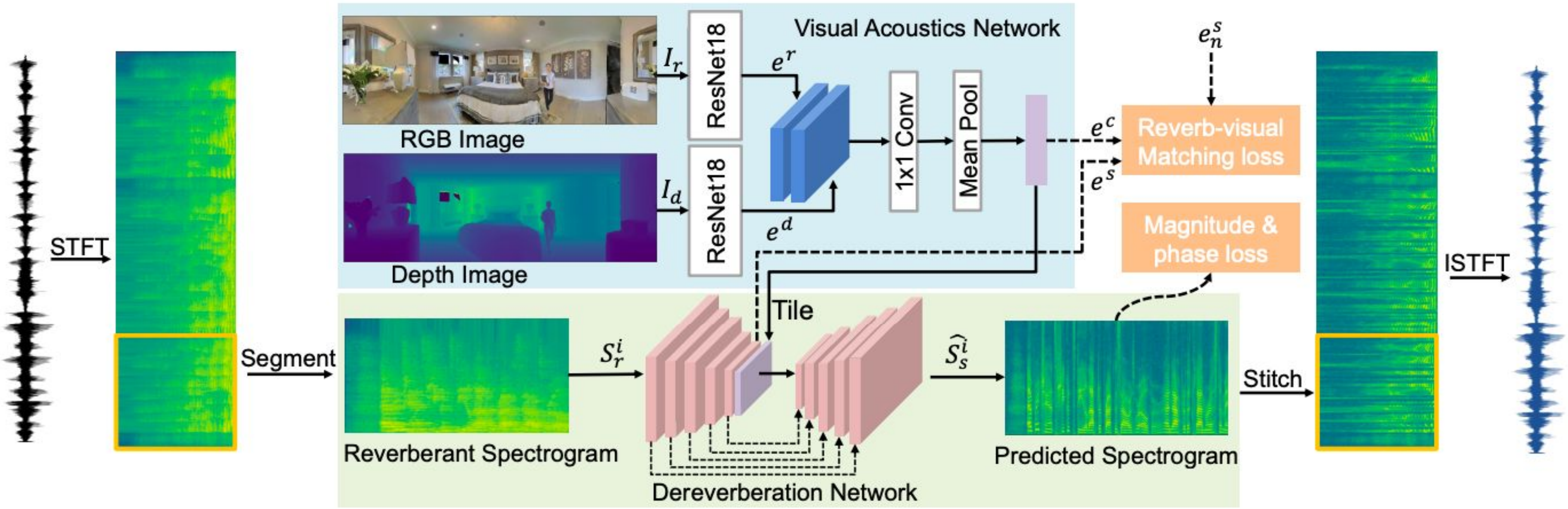
Simulated Dataset:

- Audio-visual simulator SoundSpaces (contains pre-computed RIR)
- Samples from LibriSpeech used as source audio
- Convolve speech waveform with RIR at random location
- Augment 3D humanoid of same gender at speaker location
- Obtain RGB and Depth images, both panorama and normal FOV

Real Dataset:

- Collected data in auditoriums, meeting rooms, atriums, corridors and classrooms
- Source speech obtained from Librispeech and played through a loudspeaker
- Image captured using iPhone11 camera, depth estimated using pre-trained network
- Audio recorded using external microphone ZYLIA ZM-1
- Both microphone and camera placed at same height

Approach



Visual Acoustics Network
RGB Image
Depth Image
I_r
I_d
ResNet18
ResNet18
e^r
e^d
1x1 Conv
Mean Pool
e^c
e^s
e_n^s
Reverb-visual Matching loss
Magnitude & phase loss
STFT
Segment
Tile
S_r^i
$\widehat{S_s^i}$
Reverberant Spectrogram
Dereverberation Network
Predicted Spectrogram
Stitch
ISTFT

Losses

Magnitude Loss

$$L_{magnitude} = ||M_s^i - \hat{M}_s^i||_2$$

Phase Loss

$$L_{phase} = ||\sin(P_s^i) - \sin(\hat{P}_s^i)||_2 + ||\cos(P_s^i) - \cos(\hat{P}_s^i)||_2$$

Reverb-visual Matching Loss

$$L_{matching}(e^c, e^s, e_n^s) = \max\{d(f_n(e^c), f_n(e^s)) - d(f_n(e^c), f_n(e_n^s)) + m, 0\}$$

Evaluation

Evaluated on 3 downstream tasks;

1. Speech Enhancement
2. Automatic Speech Recognition
3. Speaker Verification

Results

Results on simulated data

	Speech Enhancement	*Speech Recognition*		*Speaker Verification*	
	PESQ ↑	WER (%) ↓	WER-FT (%) ↓	EER (%) ↓	EER-FT (%) ↓
Clean (Upper bound)	4.64	2.50	2.50	1.62	1.62
Reverberant	1.54	8.86	4.62	4.69	4.57
MetricGAN+ [16]	2.33 (+51%)	7.49 (+15%)	4.86 (-5%)	4.67 (+0.4%)	2.75 (+39%)
WPE [45]	1.63 (+6%)	8.18 (+8%)	4.30 (+7%)	5.19 (-11%)	4.48 (+2%)
Audio-only dereverb.	2.32 (+51%)	4.92 (+44%)	3.76 (+19%)	4.67 (+0.4%)	2.61 (+43%)
VIDA w/ normal FoV	2.33 (+51%)	4.85 (+45%)	3.73 (+19%)	4.53 (+3%)	2.79 (+39%)
VIDA w/o matching loss	**2.38 (+55%)**	4.59 (+48%)	3.72 (+19%)	4.02 (+14%)	2.62 (+43%)
VIDA w/o human mesh	2.31 (+50%)	4.57 (+48%)	3.72 (+19%)	4.00 (+15%)	2.52 (+45%)
VIDA	2.37 (+54%)	**4.44 (+50%)**	**3.66 (+21%)**	**3.99 (+15%)**	**2.40 (+47%)**

Results

Results on Real data (Sim2Real Transfer)

	Speech Enhancement PESQ ↑	*Speech Recognition* WER (%) ↓	*Speaker Verification* EER (%) ↓
Clean (Upper bound)	4.64	2.52	1.42
Reverberant	1.22	18.39	3.91
MetricGAN+ [16]	**1.62 (+33%)**	21.42 (-16%)	5.70 (-46%)
Audio-only dereverb.	1.41 (+16%)	15.18 (+17%)	4.24 (-8%)
VIDA w/ normal FoV	1.44 (+18%)	14.71 (+20%)	3.79 (+3%)
VIDA	1.49 (+22%)	**13.02 (+29%)**	**3.75 (+4%)**

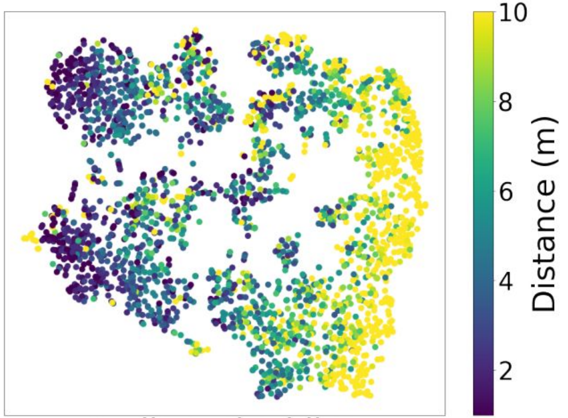
Ablation

Adding Noise

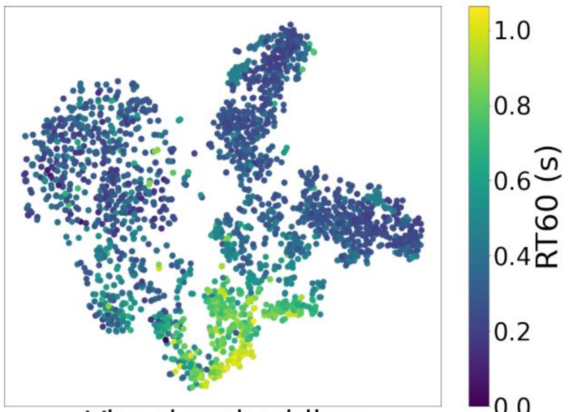
	Speech Enhancement PESQ ↑	*Speech Recognition* WER (%) ↓	WER-FT (%) ↓	*Speaker Verification* EER (%) ↓	EER-FT (%) ↓
Clean (Upper bound)	4.64	2.50	2.50	1.62	1.62
Reverberant	1.36	12.27	6.38	4.69	5.10
MetricGAN+ [16]	**2.12 (+57%)**	9.40 (+23%)	7.09 (-11%)	4.94 (-5%)	3.38 (+34%)
WPE [45]	1.39 (+2%)	11.32 (+8%)	7.00 (-10%)	**4.48 (+4%)**	4.95 (+3%)
Audio-only dereverb.	1.76 (+29%)	7.37 (+40%)	5.52 (+14%)	5.75 (-23%)	3.58 (+30%)
VIDA w/ normal FoV	1.76 (+29%)	7.51 (+39%)	5.51 (+14%)	5.54 (-18%)	3.40 (+33%)
VIDA w/o matching loss	1.81 (+33%)	6.76 (+45%)	5.31 (+17%)	4.95 (-6%)	3.26 (+36%)
VIDA	1.82 (+34%)	**6.53 (+47%)**	**5.29 (+17%)**	4.83 (-3%)	**3.13 (+39%)**

Contribution of each modality

	Speech Enhancement PESQ ↑	*Speech Recognition* WER (%) ↓	*Speaker Verification* EER (%) ↓
Reverberant	1.54	8.86	4.69
Audio-only dereverb.	2.32 (+51%)	4.92 (+44%)	4.67 (+0.4%)
VIDA w/o RGB	**2.38 (+55%)**	4.76 (+46%)	**3.82 (+19%)**
VIDA w/o depth	**2.38 (+55%)**	4.52 (+49%)	3.99 (+15%)
VIDA w/ early fusion	**2.38 (+55%)**	4.56 (+48.5%)	3.94 (+16%)
VIDA	2.37 (+54%)	**4.44 (+50%)**	3.99 (+15%)



Audio embedding



Visual embedding

Qualitative Results