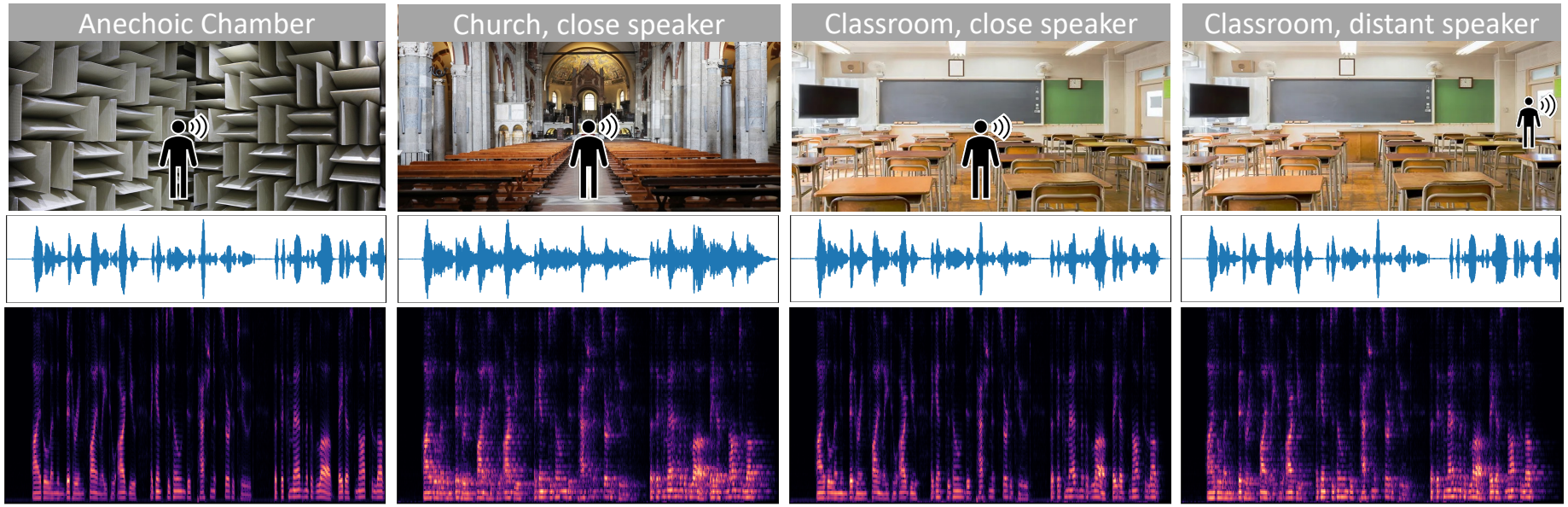
Learning Audio-Visual Dereverberation

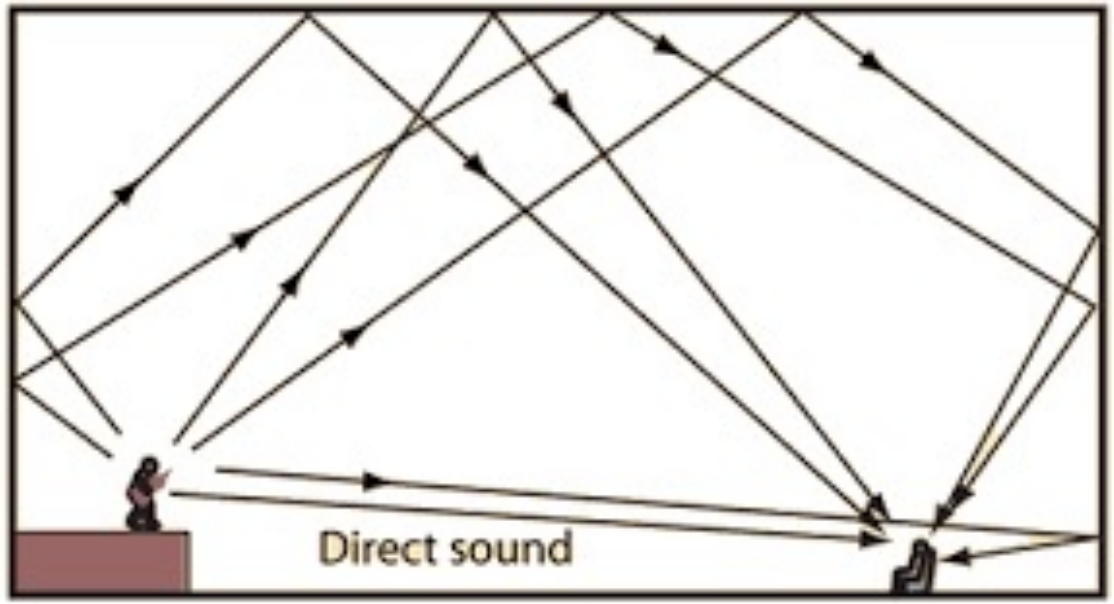
Changan Chen[1,2], Wei Sun[1], David Harwath[1], Kristen Grauman[1,2]

UT Austin[1], FAIR[2]



Reverberation in Our Daily Life

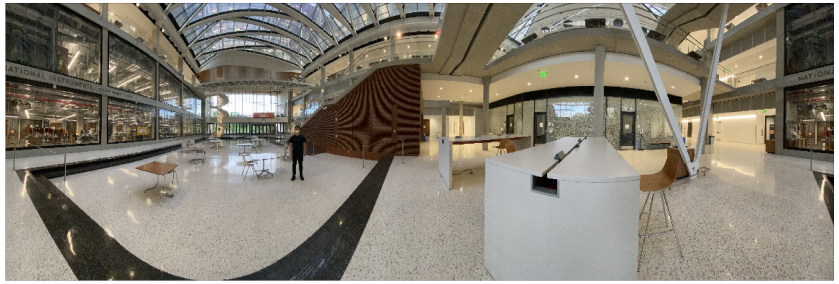
- Reverberation is everywhere in our daily life
- Audio signal reflects off surfaces and objects
- Function of room geometry, materials and speaker location



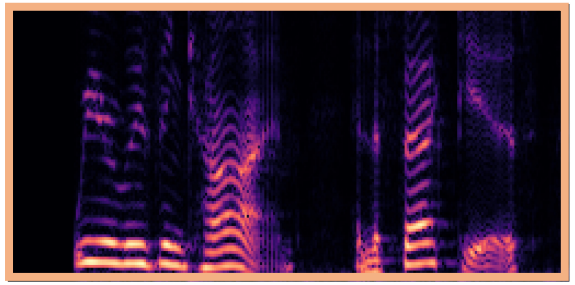
Reverberation Impacts Speech Recognition

- Reverberation damages the performance of automatic speech recognition
- Dereverberation: restore the clean speech
- Applications: robotic speech recognition, video conferencing, AR/VR

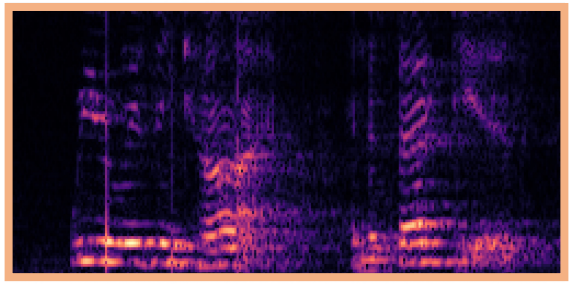
Panorama RGB



Clean (GT)



Reverberant

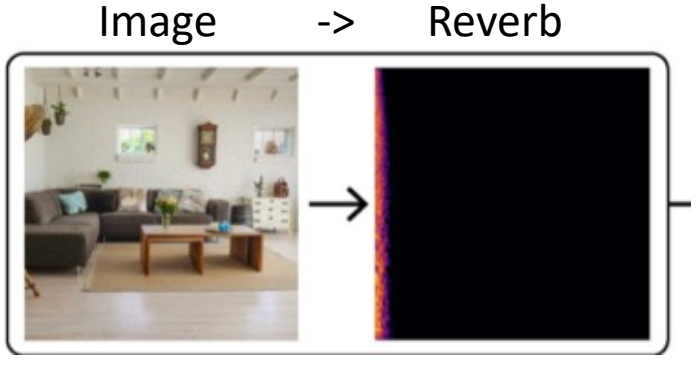


Audio Dereverberation and Speech Enhancement

- Prior work only uses audio for dereverberation
- MetricGAN+
  - State-of-the-art speech enhancement algorithm
  - Optimizes the speech metric (PESQ) directly
- Room-aware dereverberation
  - Room characteristics are estimated from reverberation in the audio

Visual Understanding of Room Acoustics

- Room acoustics:
  - How sound propagates in a closed or seme-closed space
  - Can be measured by room impulse response
  - Macro characteristics: Reverberation time by 60dB(RT60), Direct-to-reverberant ratio (DRR) etc.
- Image2Reverb:
  - Generate RIRs with generative models based on single images
  - Images taken at an unknown location different from the microphone
- Goal: to estimate room acoustics features from visual for dereverberation



The Audio-Visual Dereverberation Task

The reverberant speech A_r can be modeled as:

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

$A_s(t)$ is the anechoic source speech and R is the room impulse response.

Given the RGB I_r, depth Image I_d, received audio A_r, predict A_s

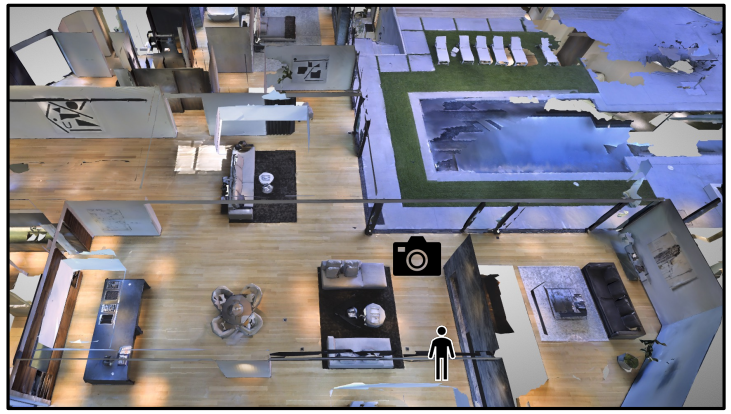
$$\hat{A}_s(t) = f_p([I_r, I_d, A_r(t)])$$

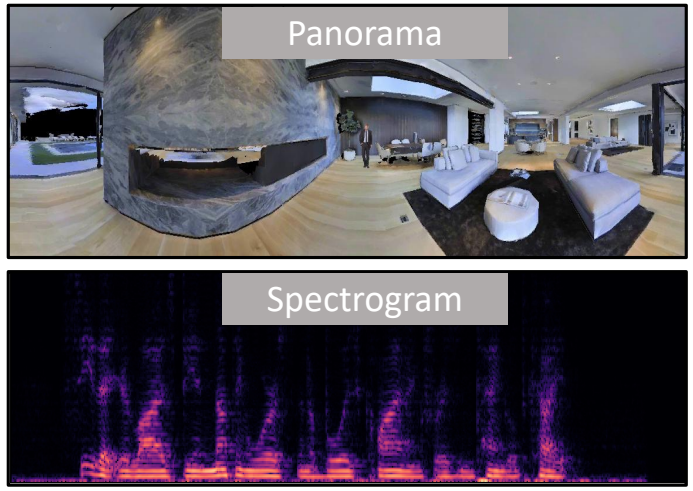
Dataset Curation

- Obtaining the right data is challenging
- Video data does not have clean anechoic signal
- Recorded RIR datasets do not have camera at mic locations or speaker
- Introduce both simulated and real data

Simulated Data

- The ability to control environment settings (positions of the speaker, listener, speech content and room)
- Use the audio-visual simulator SoundSpaces and Matterport3D
- Use LibriSpeech as the source speech corpus
- Insert a 3D humanoid of the same gender at the speaker location
- Panorama: 18 images of FOV 20 (192x756)
- Normal view: 4 images of FOV 20 (384x256)



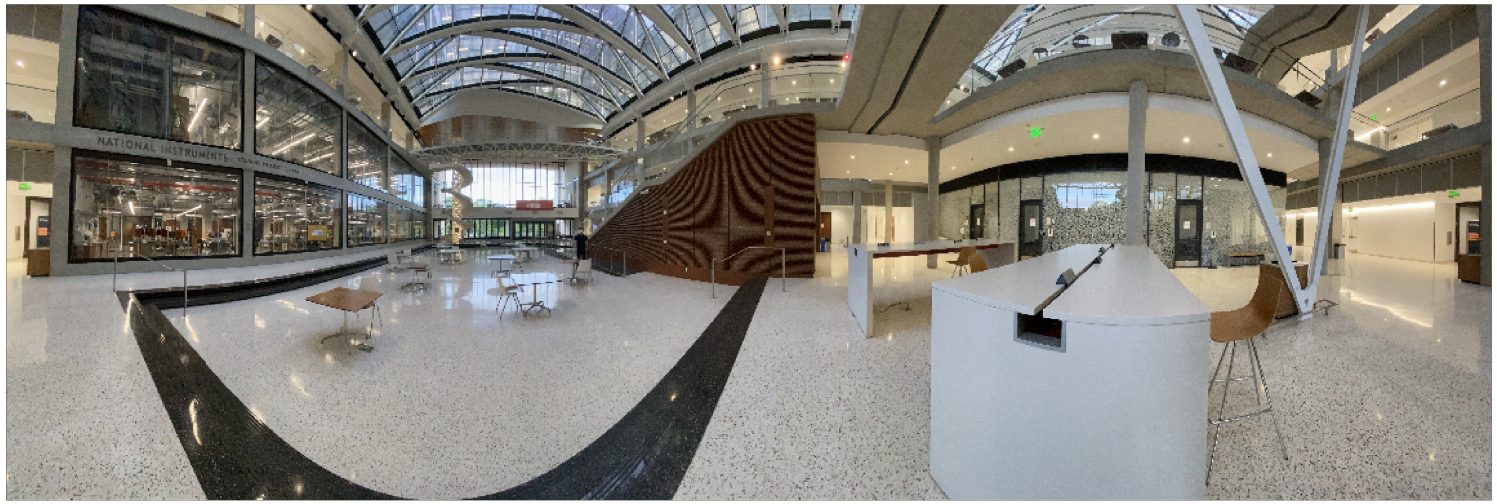


SoundSpaces Demo

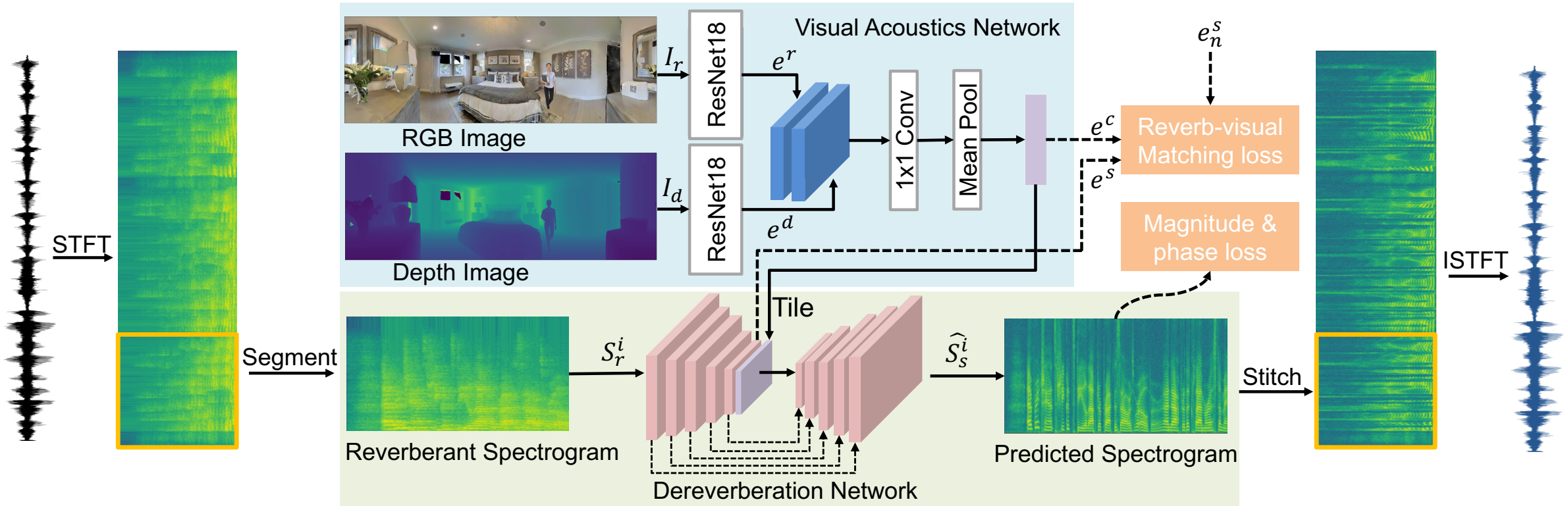
SoundSpaces: Audio-Visual Navigation in 3D Environments, Changan Chen*, Unnat Jain*, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, Kristen Grauman, ECCV 2020

Real Data Collection

- Use iPhone 11 pro camera to capture panoramic RGB image
- Use monocular depth estimation algorithm to generate depth
- Microphone: ZYLIA ZM-1 mic (1 channel)
- Play utterances through a loudspeaker held by a person
- Varying environments: auditoriums, meeting rooms, atriums, corridors and classrooms
- Varying speaker location from near-field to mid-field to far-field
- Record ambient sound during recording



Visually-Informed Dereverberation of Audio



Training

- 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:

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

- Overall objective:

$$L_{total} = L_{magnitude} + \lambda_1 L_{phase} + \lambda_2 L_{matching},$$

Evaluation Tasks and Metrics

- Speech Enhancement (SE):
  - Improve the sonic quality of speech signal
  - Metric: Perceptual Evaluation of Speech Quality (PESQ)
- Automatic Speech Recognition (ASR):
  - Transcribe the sequence of words that spoken in the audio
  - Metric: Word Error Rate (WER)
  - Evaluate with pretrained model and finetuned model
- Speaker Verification (SV):
  - Detect whether two utterances were spoken by the same speaker
  - Metric: Equal Error Rate (EER)
  - Evaluate with pretrained model and finetuned model

Baselines

- MetricsGAN+: state-of-the-art SE model, learning based
- WPE: statistical SE model
- Audio-only dereverberation: an ablation of the proposed VIDA model

Results in Scanned Environments

- VIDA outperforms all baselines
- Panorama input leads to better results compared to normal FOV
- Reverb-visual matching loss help the model learn a better feature representation
- Removing human meshes leads to performance drop

	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 on Real Data

- VIDA generalizes to real data
- It still outperforms baselines on ASR and SV tasks
- MetricGAN+ does better on speech enhancement

	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%)**

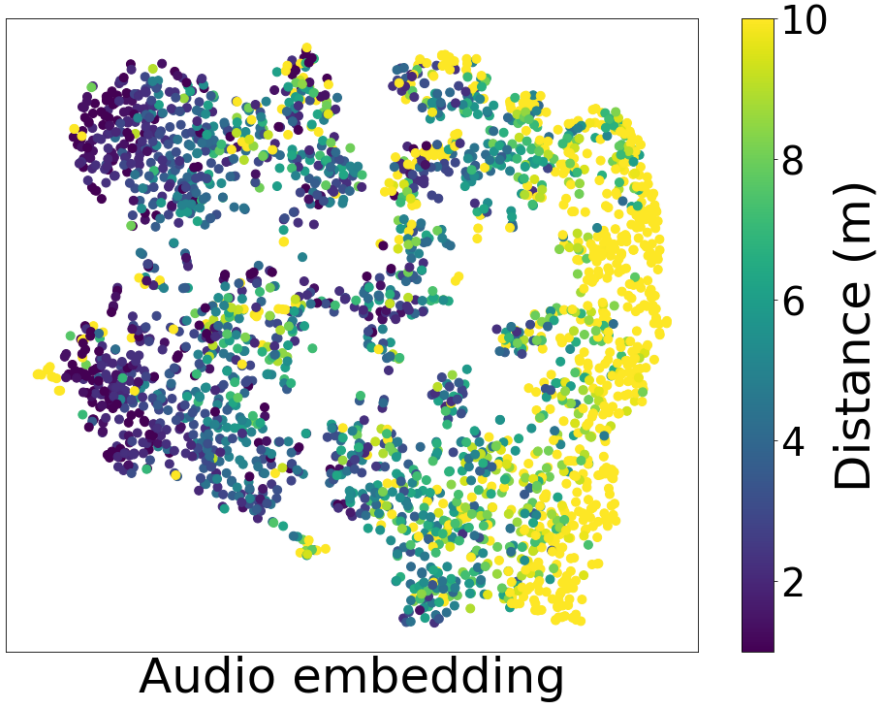
Breakdown of WER

- VIDA outperforms Audio-only dereverb. in most scenarios
- Large environments/distances tend to be more reverberant

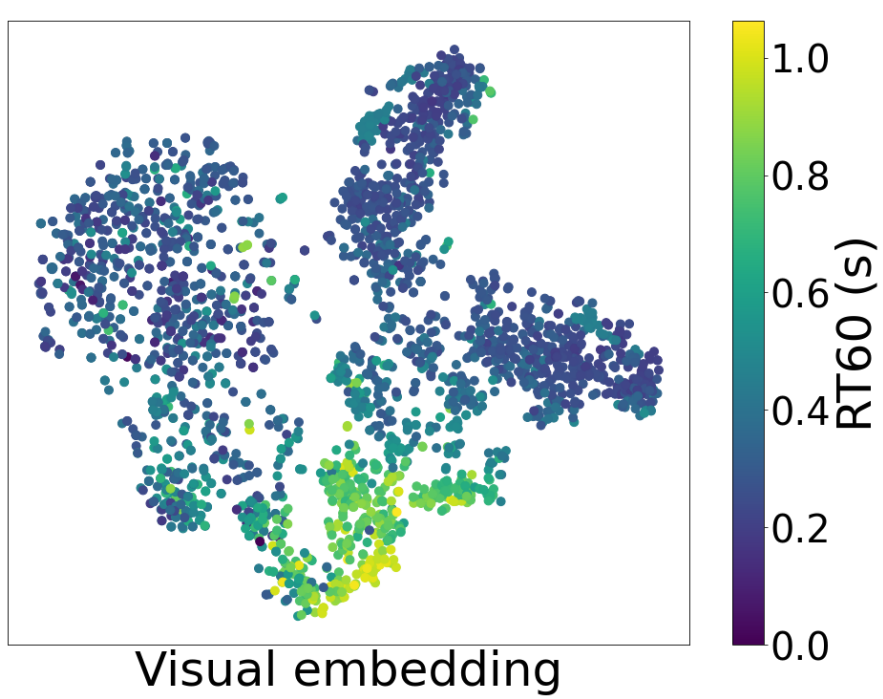
	Atrium	Auditorium	Meeting Room	Classroom	Corridor
Near-field	14.10 / **8.97**	**0.91** / **0.91**	**4.98** / 6.47	6.14 / **5.26**	2.15 / **1.79**
Mid-field	21.78 / **18.94**	**5.06** / 6.32	**7.67** / **7.67**	2.56 / **1.47**	7.27 / **4.36**
Far-field	52.38 / **50.52**	10.44 / **7.46**	21.95 / **6.71**	**5.91** / 6.82	25.23 / **21.10**

TSNE Visualizations

- Color points according to the ground truth distance to speaker
- Color points according to the reverberation time decay by 60 dB (RT60)

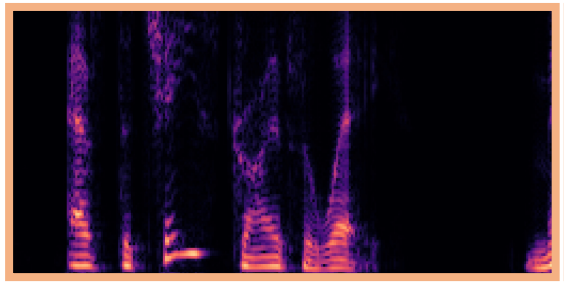
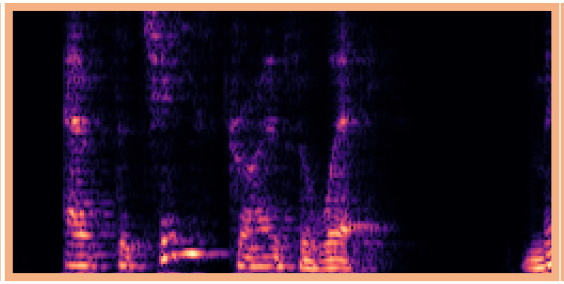
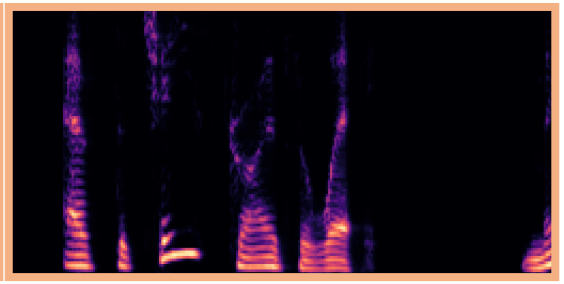
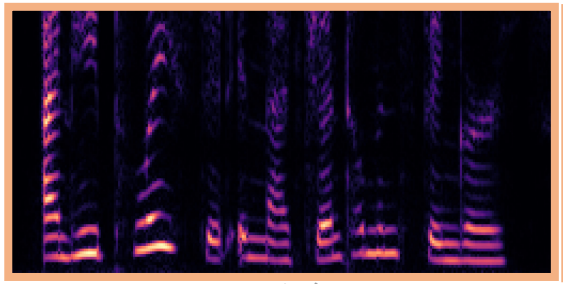
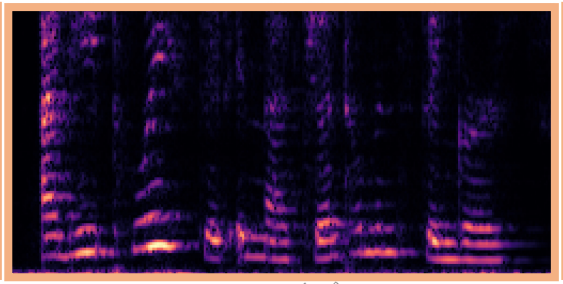
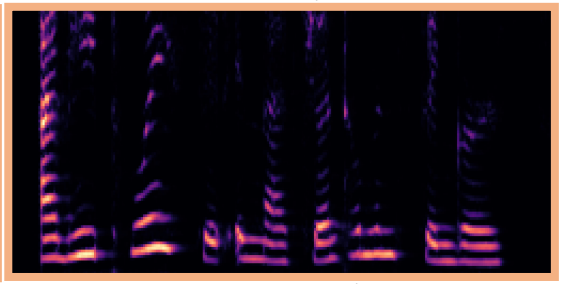
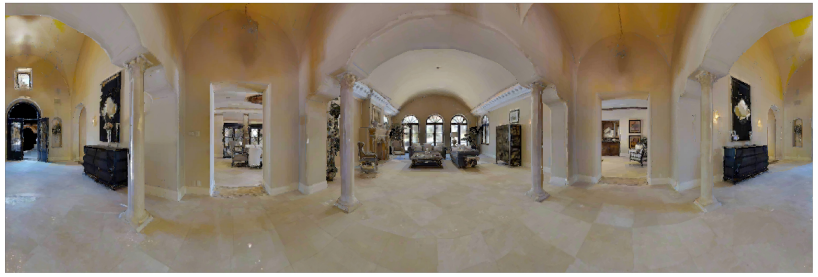
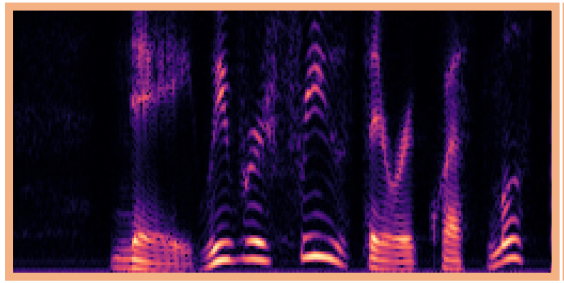
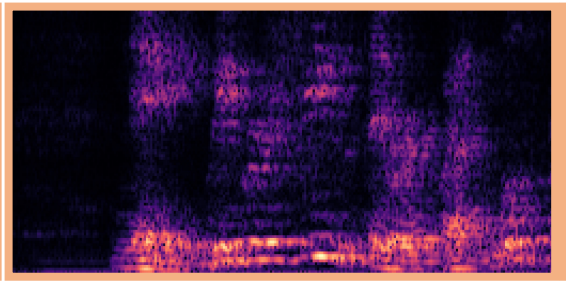
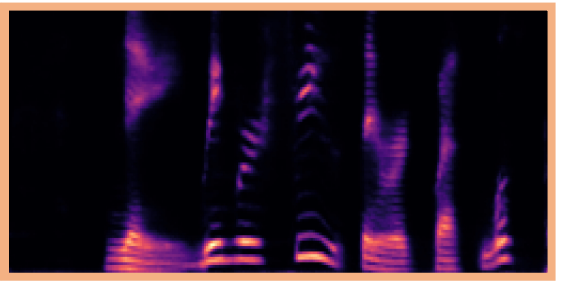


Audio TSNE

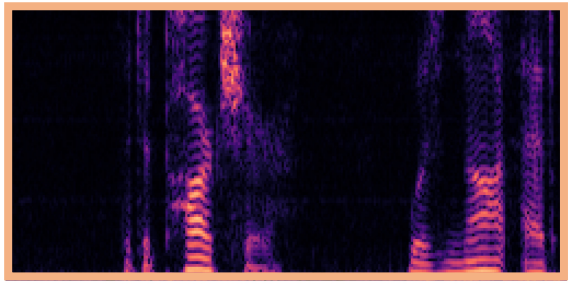
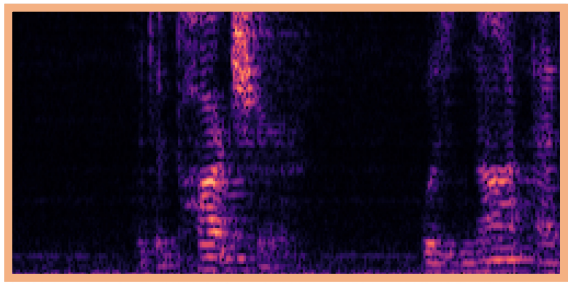
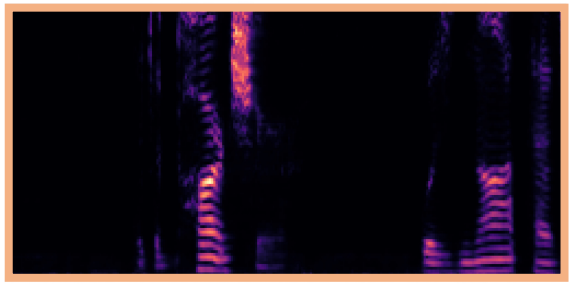
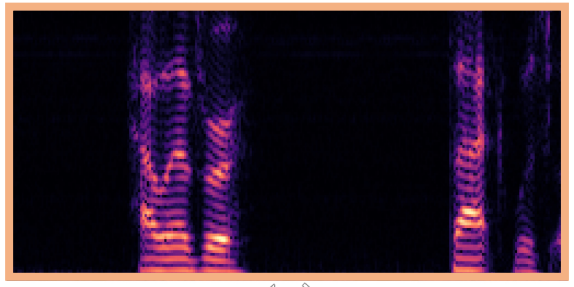
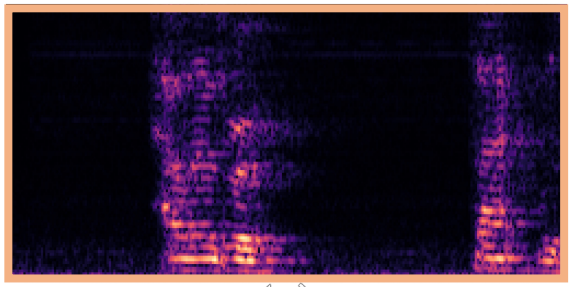
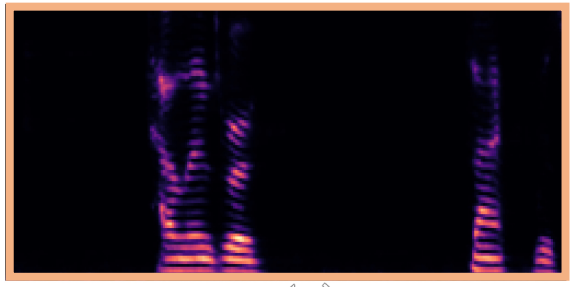
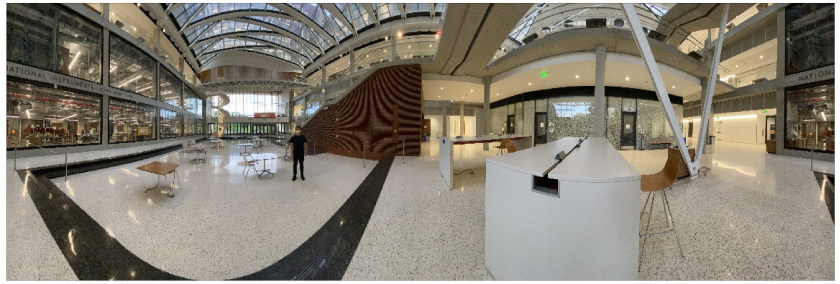
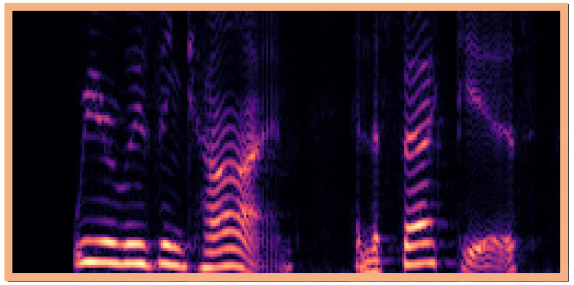
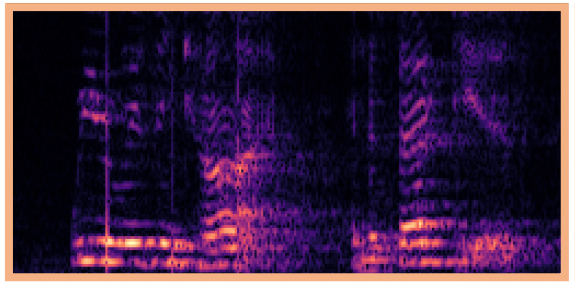
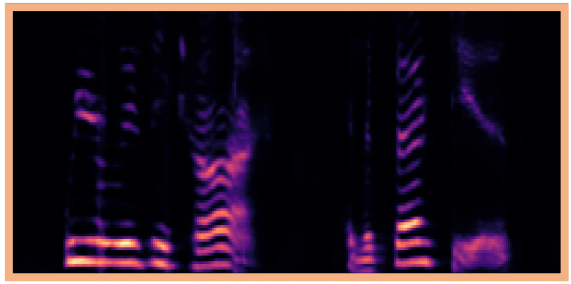


Visual TSNE

Simulated Examples

Panorama RGB	Clean (GT)	Reverberant	De-reverberated by VIDA
			
Long corridor, distant speaker, quite reverberant			
			
Big space, close speaker, quite reverberant			
			
Large space, out of view, very reverberant			

Real Examples

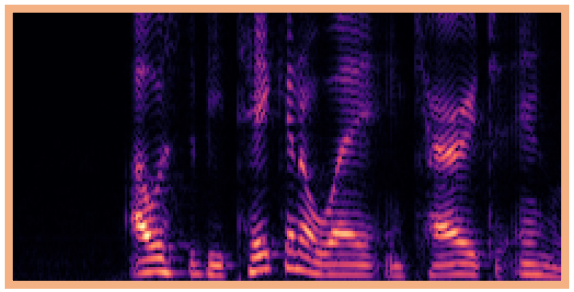
Panorama RGB	Clean (GT)	Reverberant	De-reverberated by VIDA
			
Classroom, close speaker, not very reverberant			
			
Classroom, distant speaker, quite reverberant			
			
Atrium, close speaker, reverberant			

Failure Cases

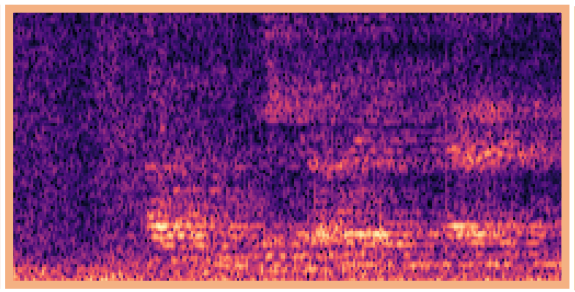
Panorama RGB

Atrium, distant speaker, very reverberant

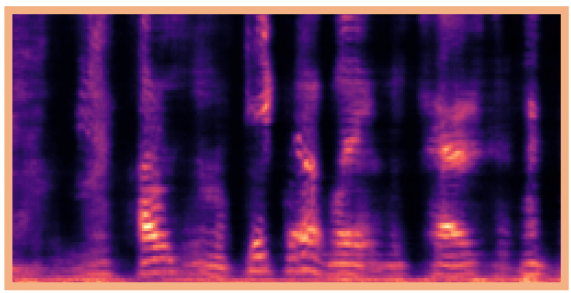
Clean (GT)



Reverberant



De-reverberated by VIDA



When there is a lot of ambient noise and reverberation in the audio, the model almost fails to predict the clean

Thank you!