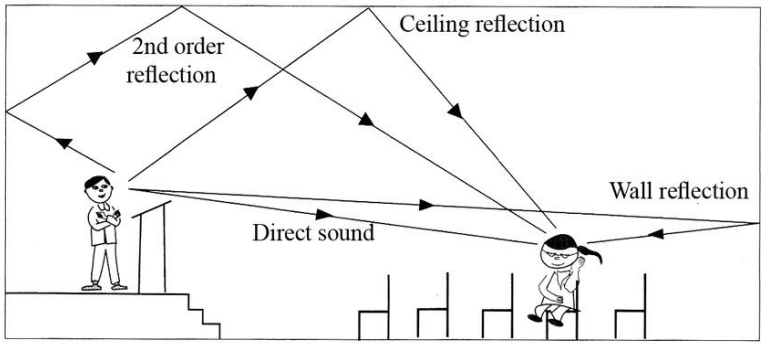
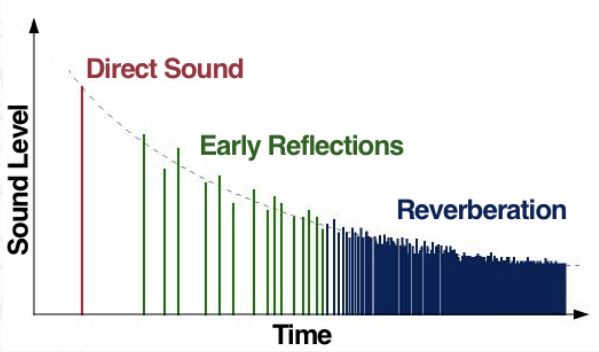
AV-RIR: Audio-Visual Room Impulse Response Estimation

Anton Ratnarajah, Sreyan Ghosh, Sonal Kumar, Purva Chiniya, Dinesh Manocha

Room Impulse Response





Room Impulse Response

Reverberant Speech = Clean Speech * RIR

Motivation

- **Audio-only** RIR estimation techniques are capable of estimating **early components** and are not effective in estimating late components.

- **Visual-only** RIR estimation demonstrates feasibility of predicting **late components** from the RGB image of the environment, however these approaches are not effective in estimating early components.

- Considering the limitation of prior works, we propose AV-RIR, a novel multi-modal multi-task learning approach for RIR estimation.

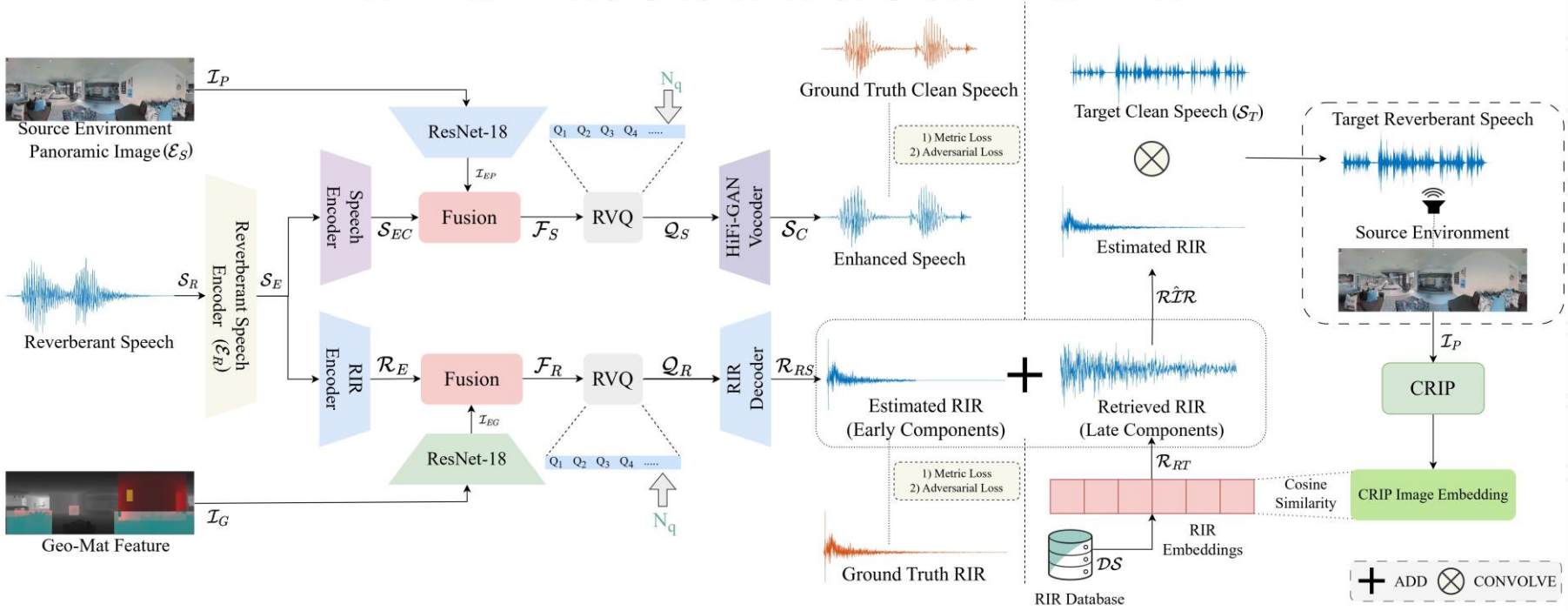
AV-RIR

- AV-RIR employs a **neural codec-based multi-modal architecture** that inputs audio and visual cues and proposed **novel Geo-Mat feature** that captures room geometry and material information.

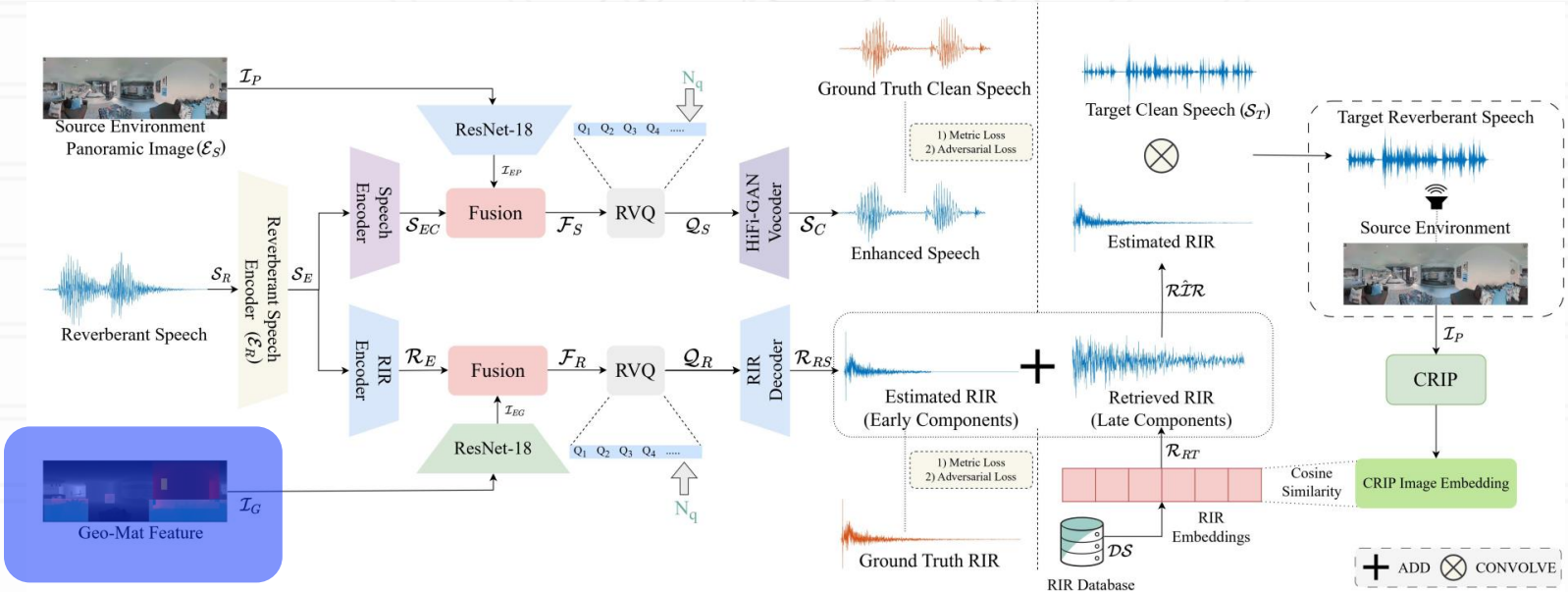
- We also propose **CRIP** to improve the late reverberation of estimated RIR using retrieval and observe that CRIP improves late reverberation by 86%.

- AV-RIR solves an auxiliary speech dereverberation task for learning RIR estimation. Through this, AV-RIR essentially learns to separate anechoic speech and RIR.

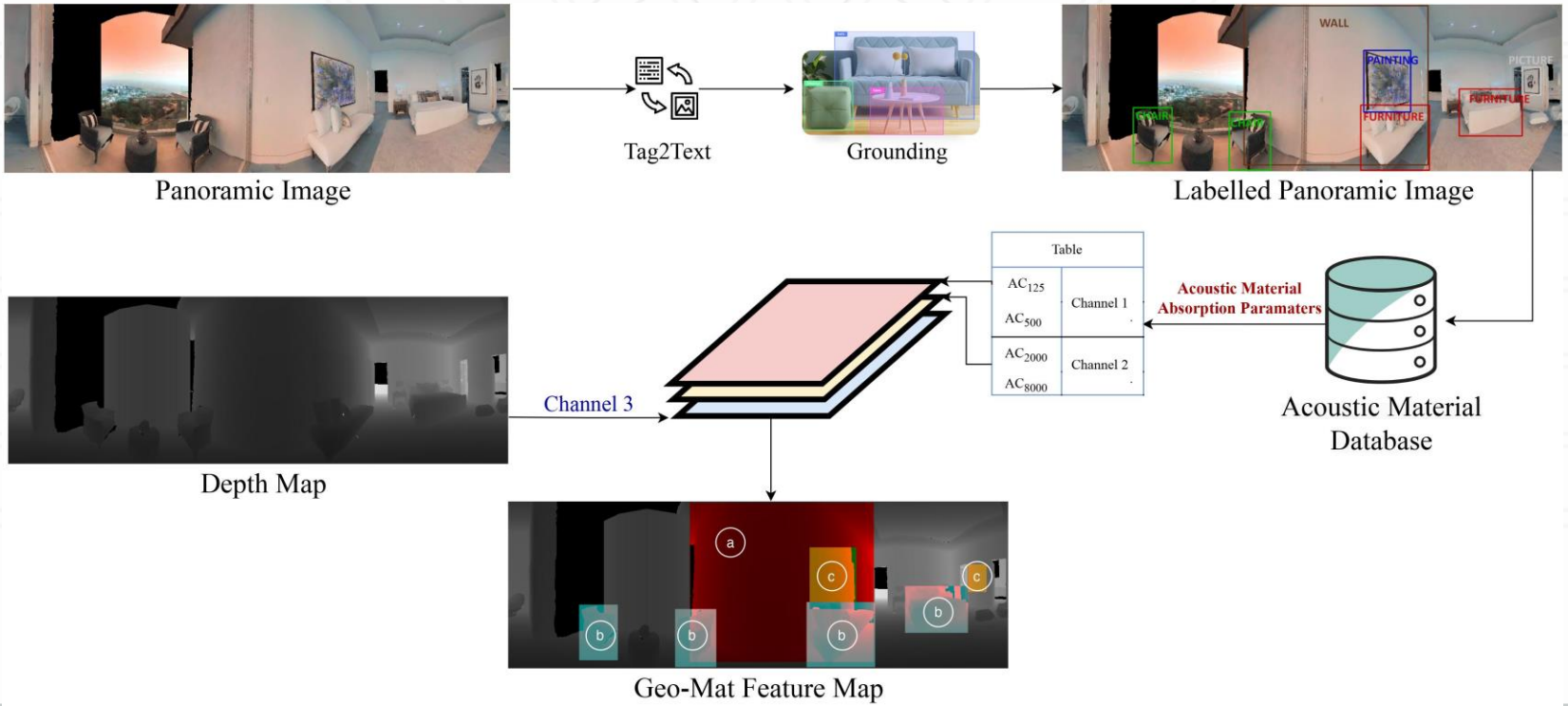
AV-RIR


Source Environment Panoramic Image ($\mathcal{E}_S$)
$\mathcal{I}_P$
ResNet-18
$\mathcal{I}_{EP}$
N_q
Q_1 Q_2 Q_3 Q_4 .....
Speech Encoder
$\mathcal{S}_{EC}$
Fusion
$\mathcal{F}_S$
RVQ
$\mathcal{Q}_S$
HiFi-GAN Vocoder
$\mathcal{S}_C$
Ground Truth Clean Speech
1) Metric Loss
2) Adversarial Loss
Enhanced Speech
Reverberant Speech
$\mathcal{S}_R$
Reverberant Speech Encoder ($\mathcal{E}_R$)
$\mathcal{S}_E$
RIR Encoder
$\mathcal{R}_E$
Fusion
$\mathcal{F}_R$
RVQ
$\mathcal{Q}_R$
RIR Decoder
$\mathcal{R}_{RS}$
$\mathcal{I}_{EG}$
ResNet-18
Geo-Mat Feature
$\mathcal{I}_G$
Estimated RIR (Early Components)
Retrieved RIR (Late Components)
$\mathcal{R}\hat{I}\mathcal{R}$
Estimated RIR
Target Clean Speech ($\mathcal{S}_T$)
Target Reverberant Speech
Source Environment
$\mathcal{I}_P$
CRIP
CRIP Image Embedding
$\mathcal{R}_{RT}$
Cosine Similarity
RIR Embeddings
$\mathcal{DS}$
RIR Database
Ground Truth RIR
ADD
CONVOLVE

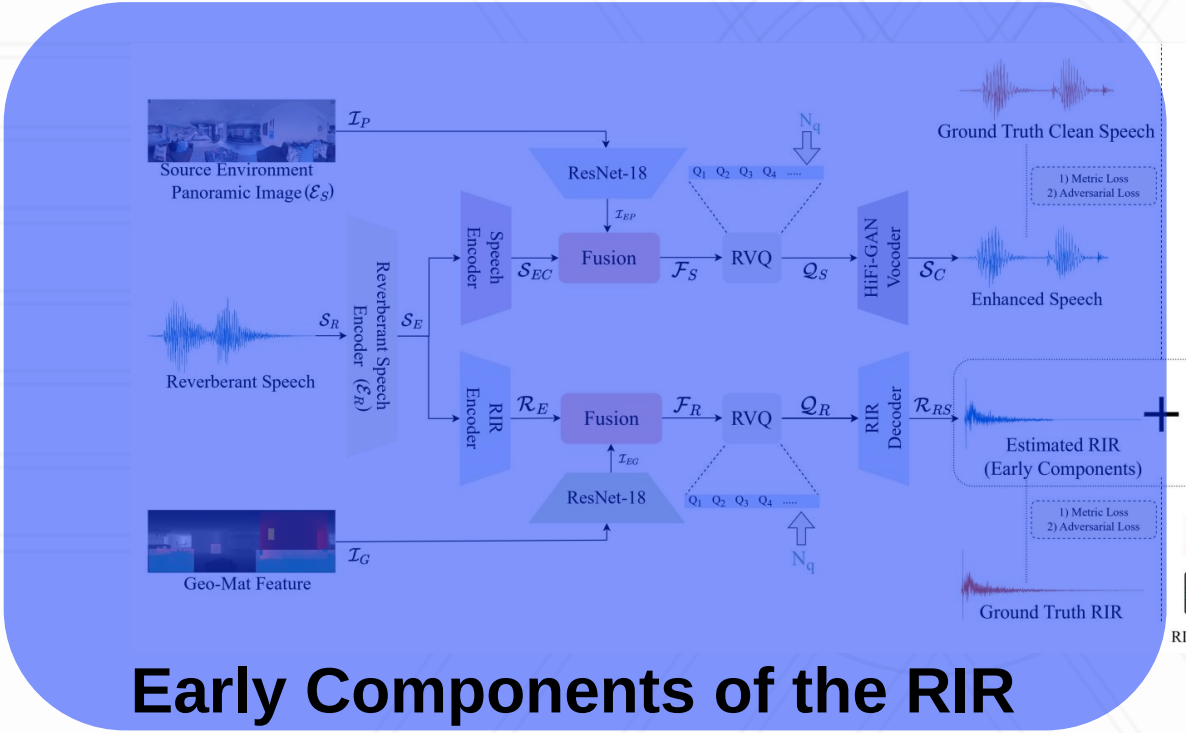
AV-RIR : Geo-Mat Feature



Geo-Mat Feature

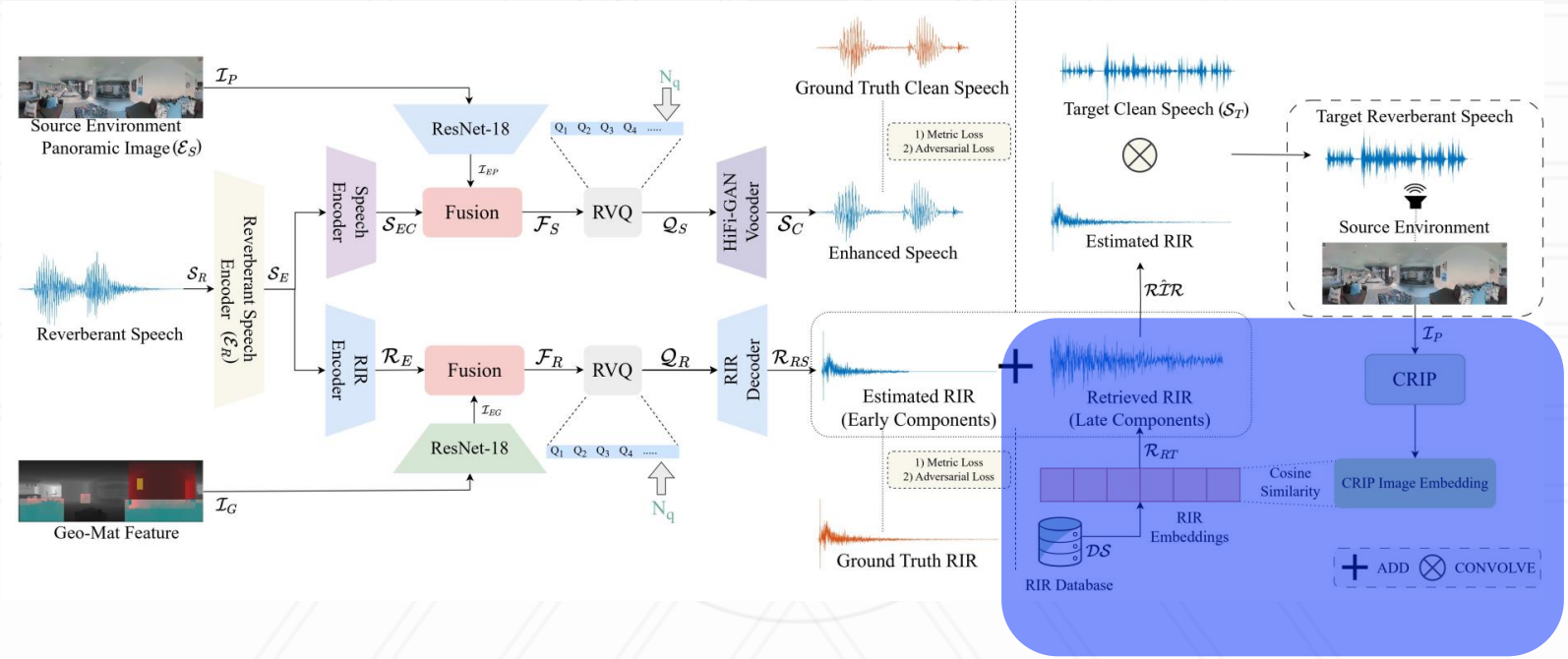


AV-RIR : Training

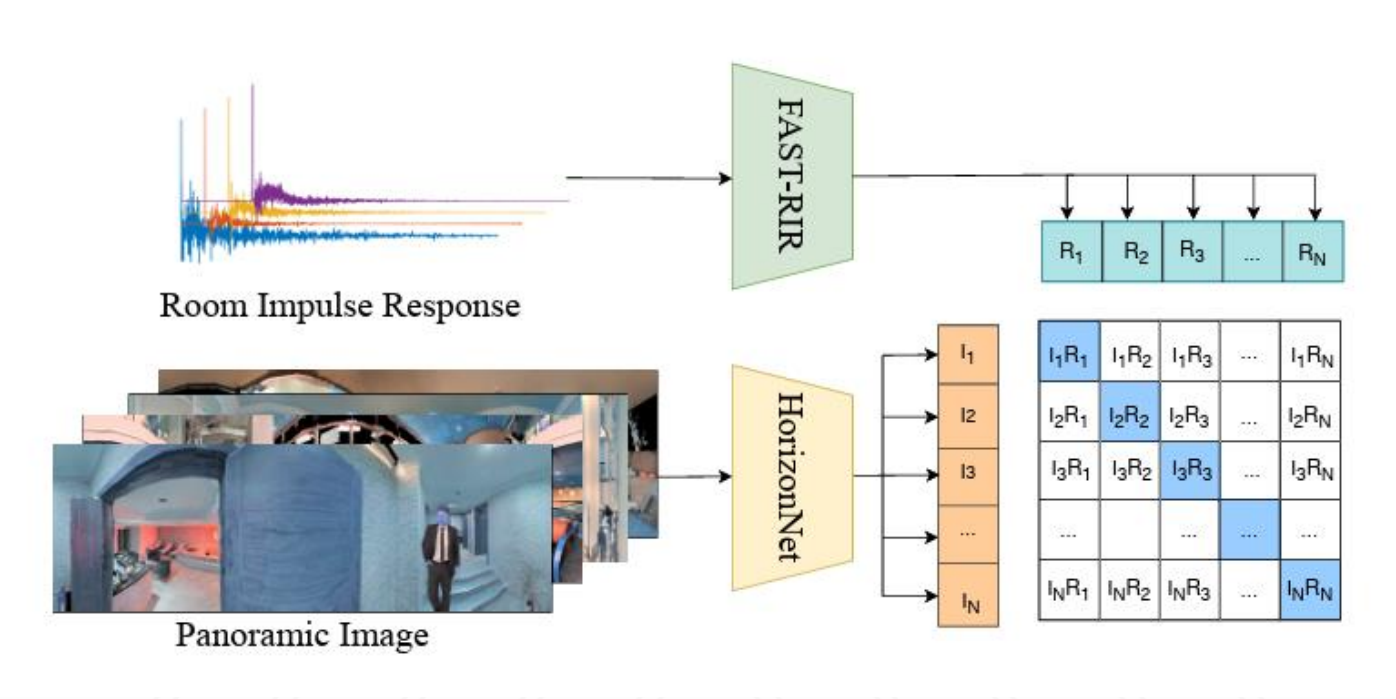


Early Components of the RIR

AV-RIR : Late Component



Contrastive RIR-Image Pretraining (CRIP)



Qualitative Results

- **Novel View Acoustic Synthesis :** In the novel view acoustic synthesis task, given the audio-visual input from the source and target viewpoint, we modify the reverberant speech from the source viewpoint to sound as if it is recorded from the target viewpoint.

- **Visual-Acoustic Matching :** In the visual-acoustic matching task, we resynthesize the speech from the source environment to match the target environment.

Novel View Acoustic Synthesis

- We estimate the Enhanced Speech from source viewpoint Audio-Visual input.
- The RIR of the target view is estimated from target viewpoint Audio-Visual Input.
- The target RIR is convolved with source clean speech to synthesize speech for target viewpoint.

Novel View Acoustic Synthesis - Source View

Source View

Source Reverberant Speech

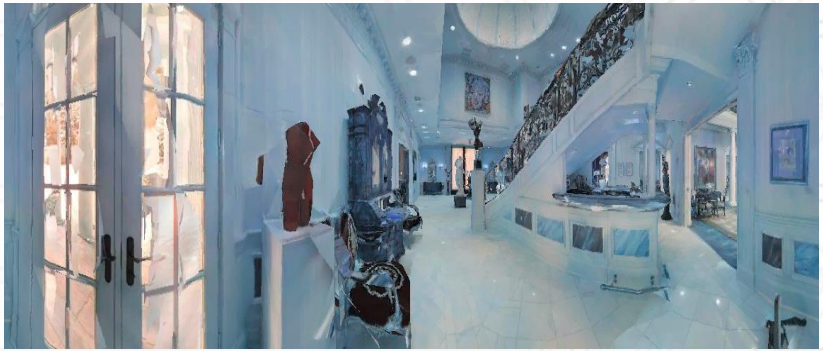
Novel View Acoustic Synthesis - Source View

Source View

Source Reverberant Speech

Enhanced Speech from our AV-RIR

Novel View Acoustic Synthesis - Target View



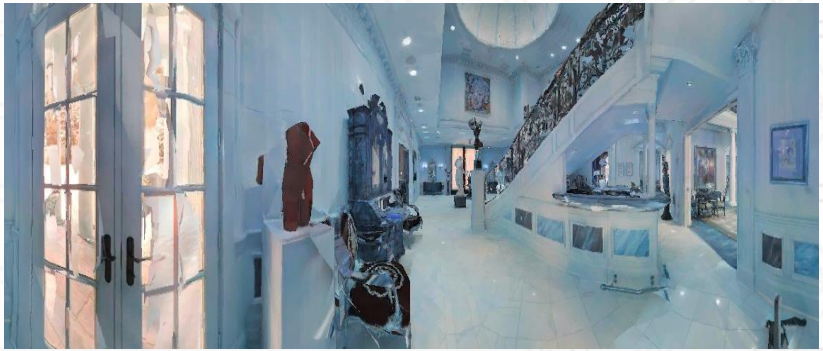
Source View

Target View

Ground Truth Speech

Synthesized Speech from our AV-RIR

Novel View Acoustic Synthesis - Target View



Source View

Target View

Ground Truth Speech

Synthesized Speech from our AV-RIR

Visual Acoustic Matching

- In Visual Acoustic Matching Tasks, the reverberation effects of input clip from source environments is matched to the target environment conditions.
- From our AV-RIR, we estimate the Enhanced Speech from source environment.
- We estimate the RIR of target environment from Audio-Visual Input.
- The target environment RIR is convolved with source environment clean speech to synthesize speech for target environment.

Visual Acoustic Matching - Source Environment

Source Environment

Source Reverberant Speech

Visual Acoustic Matching - Source Environment

Source Environment

Source Reverberant Speech

Enhanced Speech from our AV-RIR

Visual Acoustic Matching - Target Environment

Source Environment

Target Environment

Ground Truth Speech

Synthesized Speech from our AV-RIR

Visual Acoustic Matching - Target Environment

Source Environment

Target Environment

Ground Truth Speech

Synthesized Speech from our AV-RIR

UNIVERSITY OF
MARYLAND