Learning Audio-Visual Dynamics Using Scene Graphs for Audio Source Separation

Moitreya Chatterjee[1,2] Narendra Ahuja[2] Anoop Cherian[1]

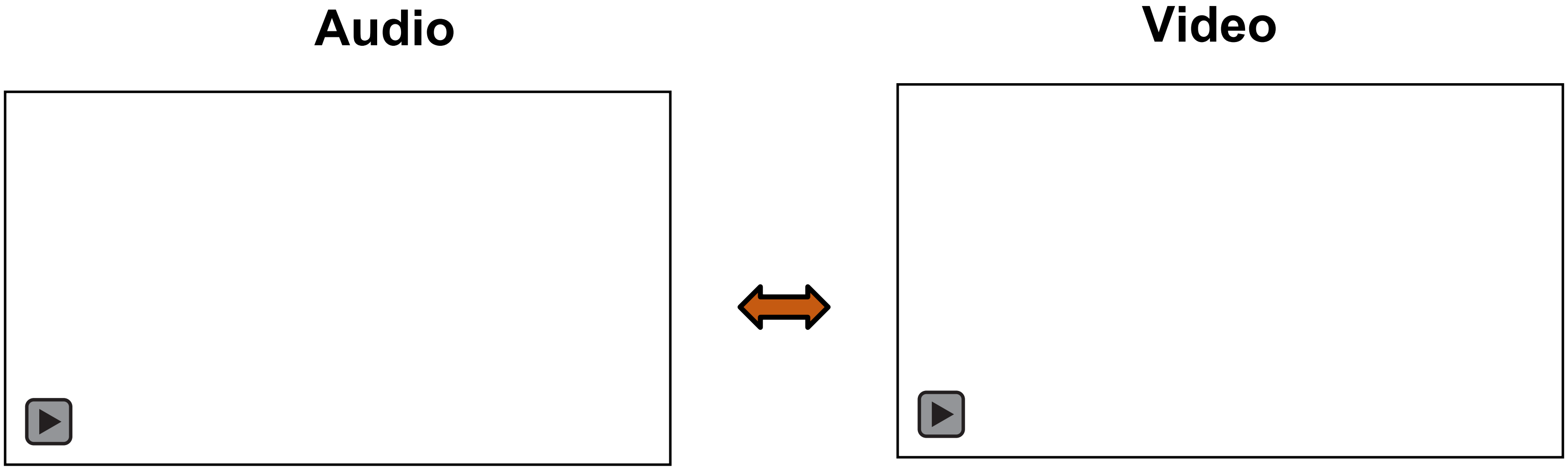
[1] Mitsubishi Electric Research Labs (MERL)

[2] University of Illinois at Urbana – Champaign (UIUC)

NEURAL INFORMATION PROCESSING SYSTEMS

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Motivation



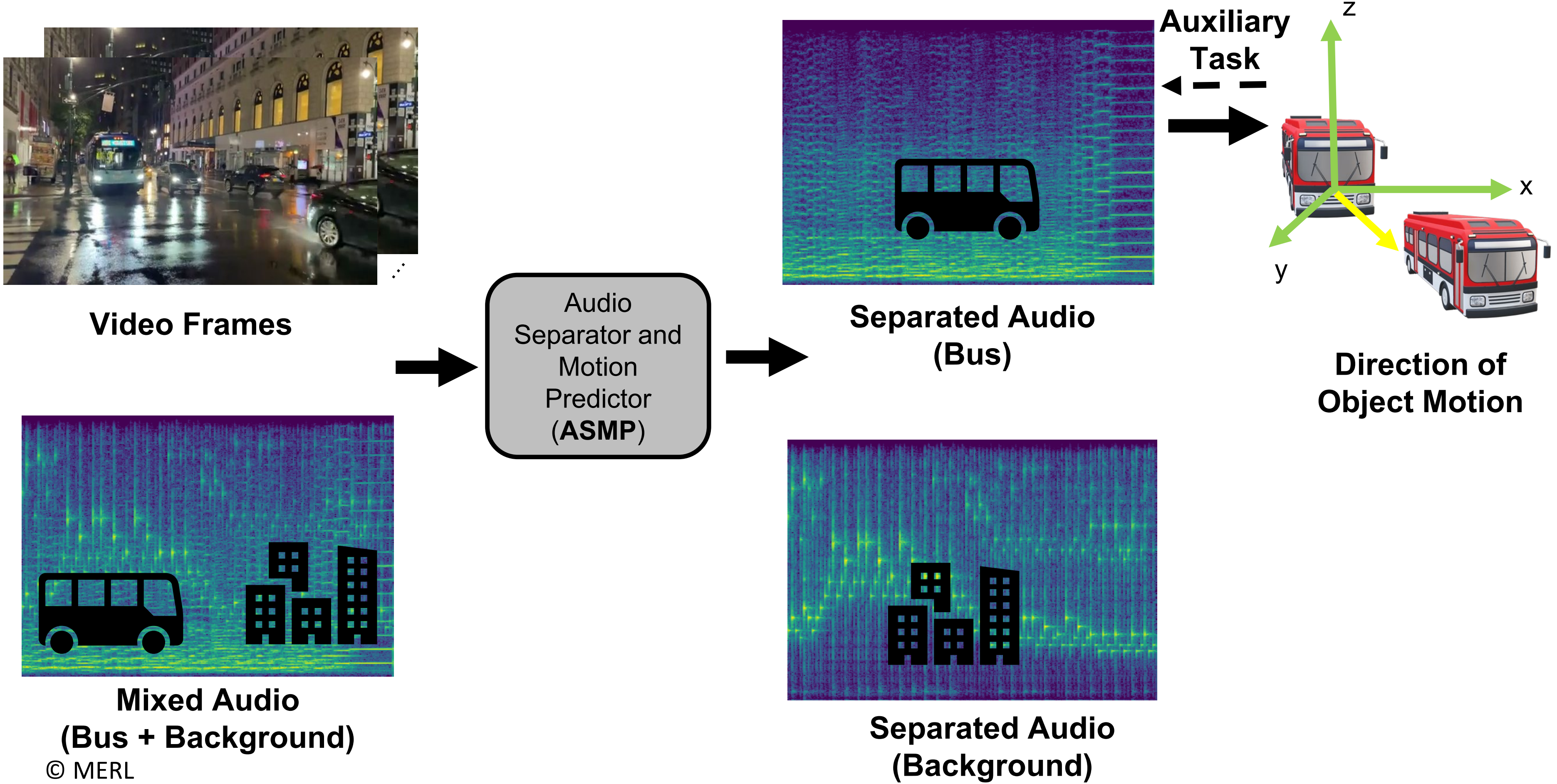
Audio

Video

<u>Motivation:</u>

Intelligent systems need to draw meaningful deductions about objects in the scene by associating their visual appearance and motion with their audio signatures.

Problem Setup



Video Frames

Audio Separator and Motion Predictor (**ASMP**)

Separated Audio (Bus)

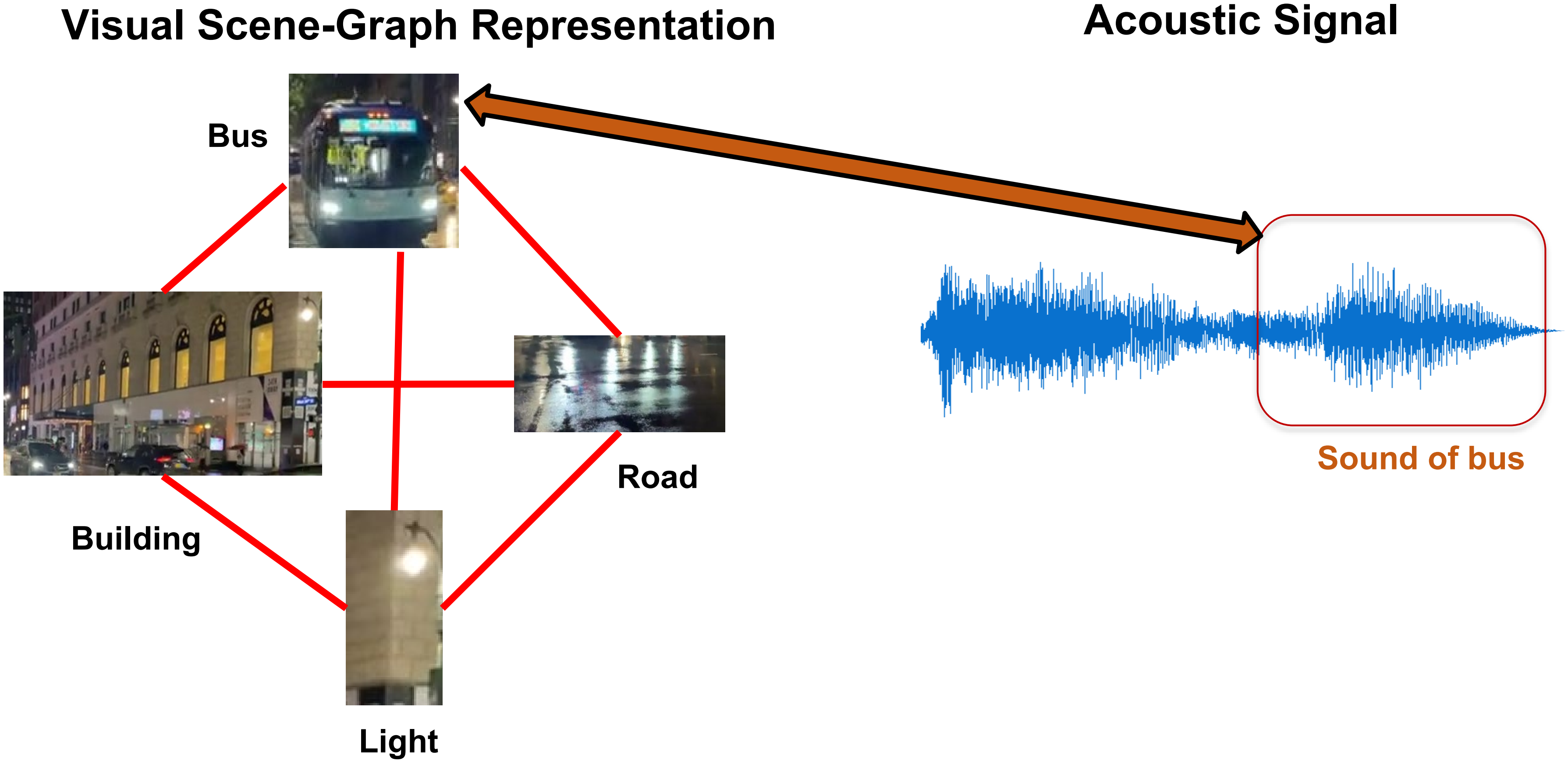
Auxiliary Task

Direction of Object Motion

Mixed Audio (Bus + Background)

Separated Audio (Background)

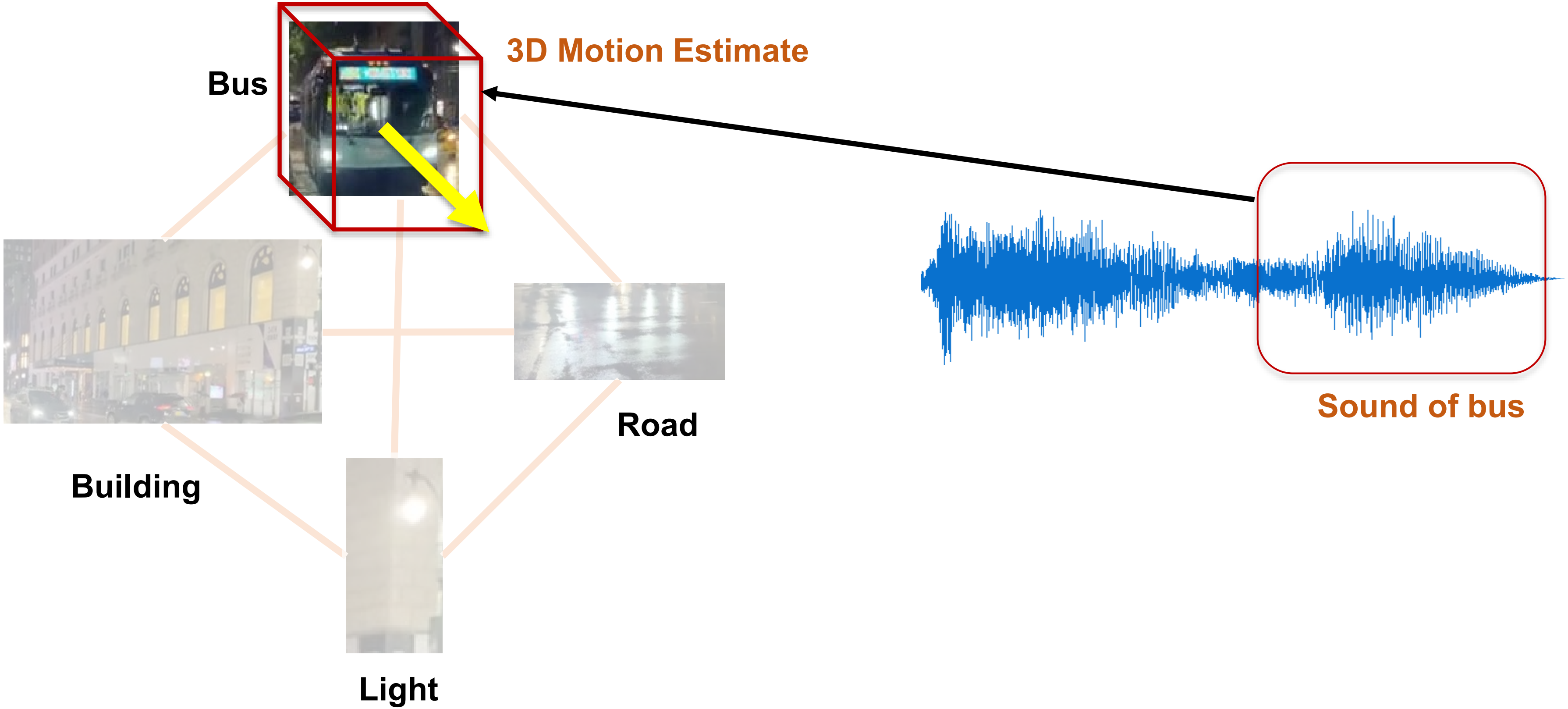
Visual Scene Graphs for Cross-modal Association



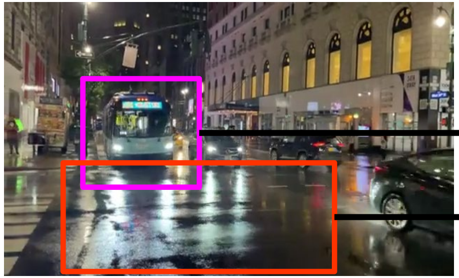
Visual Scene Graphs for Cross-modal Association

Visual Scene-Graph Representation

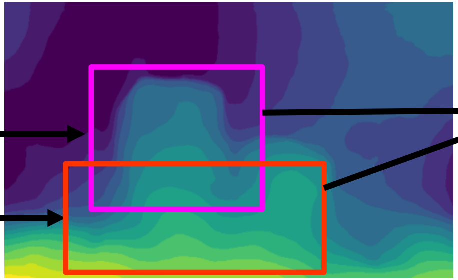
Acoustic Signal



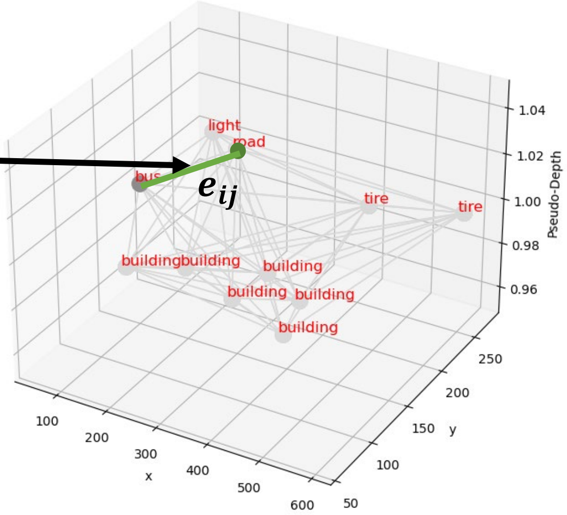
Graph Construction



Registered Reference Frame



Registered Pseudo-Depth Image



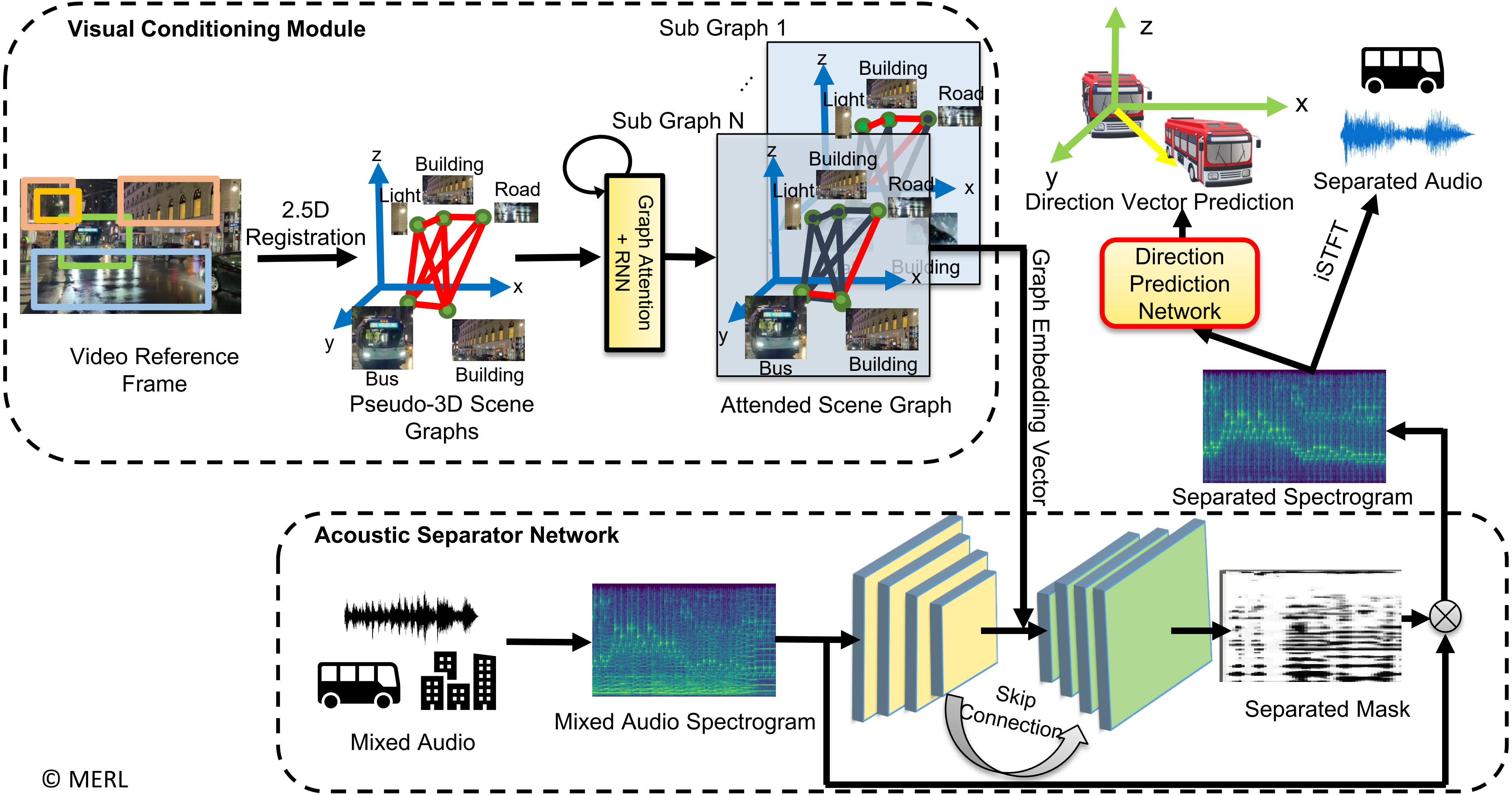
Pseudo-3D Scene Graph (Discard edges with low weight)

Edge Attribute:

Chamfer Distance

$$e_{ij} := \exp\left(\frac{-\boxed{D_{ij}}}{\sigma^2}\right)$$

Model Architecture - Overview



Visual Conditioning Module

Video Reference Frame

2.5D Registration

Pseudo-3D Scene Graphs

Graph Attention + RNN

Sub Graph 1

Sub Graph N

Attended Scene Graph

Graph Embedding Vector

Direction Prediction Network

Direction Vector Prediction

Separated Audio

iSTFT

Separated Spectrogram

Acoustic Separator Network

Mixed Audio

Mixed Audio Spectrogram

Skip Connection

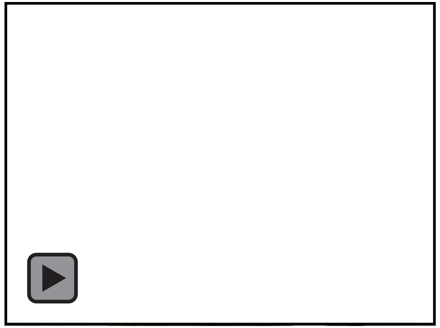
Separated Mask

© MERL

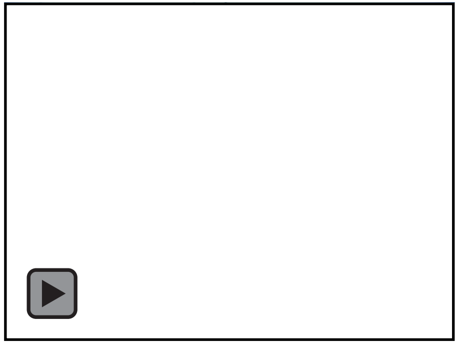
Experiments - Datasets

We conduct experiments on two audio-visual video datasets.

- ❑ **ASIW Dataset:** A novel dataset of 11k+ “in the wild” videos, 10s long, adapted from the AudioCaps dataset, consisting of 14 auditory object categories [1].



- ❑ **The AVE Dataset:** A dataset of 2.5k+ videos, 10s long, collected from YouTube [2]. Consists of 18 stationary as well as moving sound source classes.



[1] Chatterjee, M., Le Roux, J., Ahuja, N., & Cherian, A. (2021). Visual scene graphs for audio source separation. In *Proc. IEEE/CVF International Conference on Computer Vision* (pp. 1204-1213).

[2] Tian, Y., et al. (2018). Audio-visual event localization in unconstrained videos. In *Proc. of ECCV* (pp. 247-263).

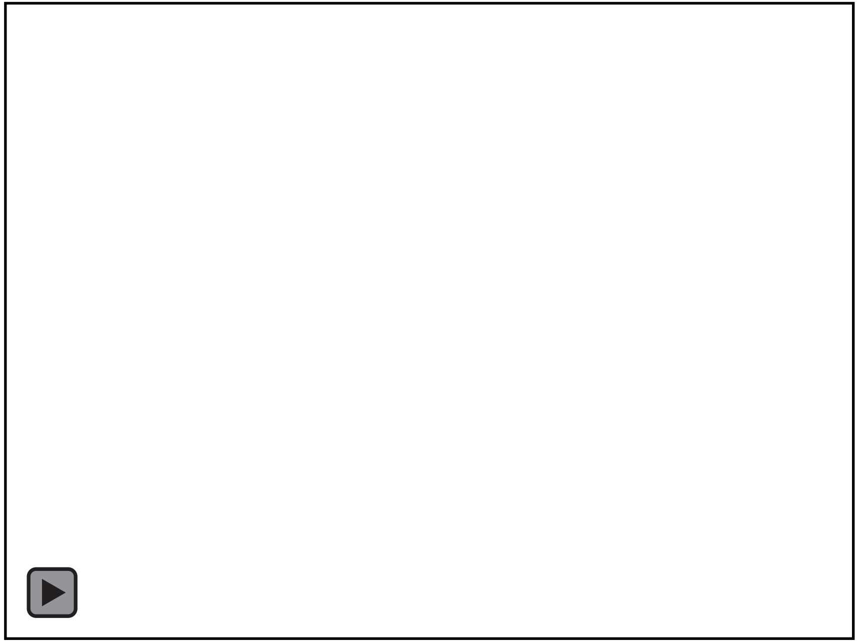
Performance: Mixture of Single Source Audios

Audio Separation	ASIW			AVE		
	SDR	SIR	SAR	SDR	SIR	SAR
Co-Separation [ICCV'19]	6.6	12.9	12.6	3.9	9.3	7.8
AVSGS [ICCV'21]	8.8	14.1	13.0	5.8	10.4	8.2
Ours (Only Graph)	**9.0**	**14.3**	**13.7**	**6.5**	**12.4**	**8.9**
Ours (Graph + Motion)	**9.6**	**14.5**	**14.1**	**7.2**	**13.3**	**9.4**

Direction Prediction	ASIW		AVE	
	10-class	28-class	10-class	28-class
Majority Vote	27.3	25.4	29.2	24.3
Ours (Graph + Motion)	**42.5**	**41.3**	**38.5**	**36.8**

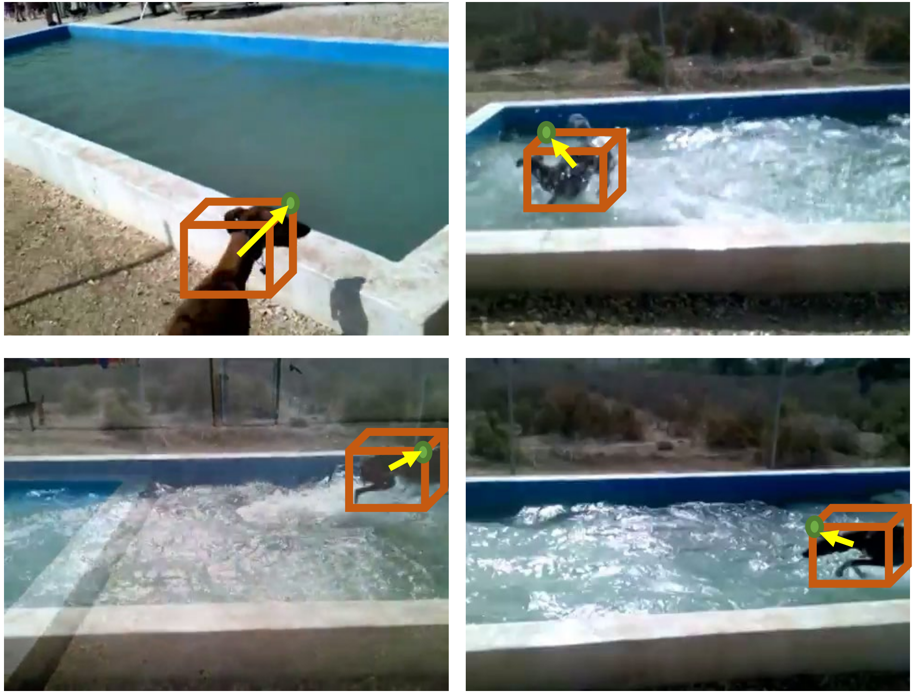
The results show that our method achieves state-of-the-art performance across both datasets, for audio separation as well as for direction prediction.

ASIW Dataset (Duet): Qualitative Result



Input Video (+ Mixed Audio)

Separated Source 1
(dog)

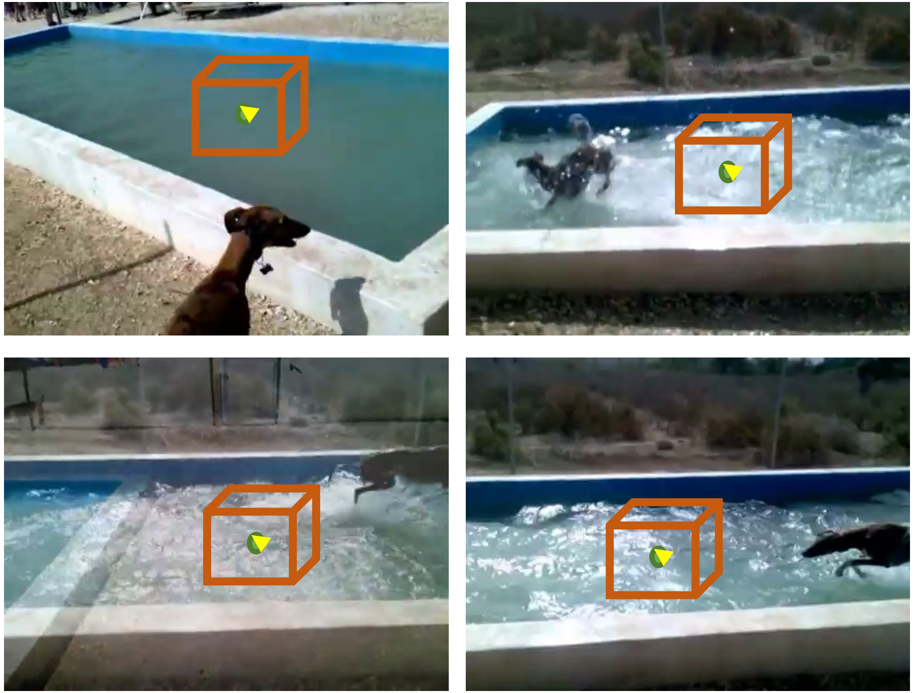


Direction of 3D motion predicted from separated audio

Green dot is the ground truth motion direction

Yellow arrow is the predicted direction

Separated Source 2
(water splash)



Direction of 3D motion predicted from separated audio

Green dot is the ground truth motion direction

Yellow arrow is the predicted direction

Thank you!

Project Page:

https://sites.google.com/site/metrosmiles/research/research-projects/asmp